CRISPRIdentify: Identification of CRISPR arrays using machine learning approach

Alexander Mitrofanov [1,*], Omer S. Alkhnbashi [1,*], Sergey A. Shmakov [2], Kira S. Makarova [2], Eugene V. Koonin [2] and Rolf Backofen [1,3,†]

[1]*Chair of Bioinformatics, University of Freiburg, Freiburg, Germany,* [2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, 20894, USA and* [3]*Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Germany*

ABSTRACT

CRISPR-Cas are adaptive immune systems that degrade foreign genetic elements in archaea and bacteria. In carrying out their immune functions, CRISPR-Cas systems heavily rely on RNA components. These CRISPR (cr) RNAs are repeat-spacer units that are produced by processing of pre-crRNA, the transcript of CRISPR arrays, and guide Cas protein(s) to the cognate invading nucleic acids, enabling their destruction. Several bioinformatics tools have been developed to detect CRISPR arrays based solely on DNA sequences, but all these tools employ the same strategy of looking for repetitive patterns, which might correspond to CRISPR array repeats. The identified patterns are evaluated using a fixed, built-in scoring function, and arrays exceeding a cut-off value are reported. Here, we instead introduce a data-driven approach that uses machine learning to detect and differentiate true CRISPR arrays from false ones based on several features. Our CRISPR detection tool, CRISPRIdentify, performs three steps: detection, feature extraction and classification based on manually curated sets of positive and

---

[*] equal contribution

[†] To whom correspondence should be addressed. Tel: +49 761 2037460; Fax: +49 761 2037462; Email: backofen@informatik.uni-freiburg.de

negative examples of CRISPR arrays. The identified CRISPR arrays are then reported to the user accompanied by detailed annotation. We demonstrate that our approach identifies not only previously detected CRISPR arrays, but also CRISPR array candidates not detected by other tools. Compared to other methods, our tool has a drastically reduced false positive rate. In contrast to the existing tools, our approach not only provides the user with the basic statistics on the identified CRISPR arrays but also produces a certainty score as a practical measure of the likelihood that a given genomic region is a CRISPR array.

INTRODUCTION

CRISPR-Cas is an adaptive immune system that is present in 90% of archaeal and about 40% of bacterial genomes [1, 2, 3]. These systems consist of arrays of direct repeats (CRISPR) and varying suits of CRISPR-associated (*cas*) genes that encode proteins involved in the immune functions.

The CRISPR-Cas immune response consists of three main stages:

(i) Adaptation that consists of selection and excision of short target segments (protospacers) from foreign DNA followed by incorporation of these segments into the CRISPR array, to form spacers via a cut-and-paste mechanism [4, 5, 6]. In most CRISPR-Cas systems, recognition of the protospacers requires the PAM (protospacer adjacent motif), a short sequence motif in the target DNA that flanks the protospacer and is crucial for avoiding self-cleavage [7, 8, 9, 10, 11].

(ii) Biogenesis of CRISPR (cr) RNAs, including expression of the pre-crRNA (a leader followed by an array of repeat-spacer units) and subsequent processing of the transcript into mature crRNA [12, 13].

(iii) Interference that involves invader DNA or RNA degradation at the cognate protospacer, guided by the crRNA and catalyzed by a specific complex of Cas proteins or a single, multidomain protein [1, 6, 14]. Typical of defense systems, CRISPR-Cas loci evolve fast and are often substantially different among otherwise closely related strains of bacteria or archaea. The CRISPR arrays are associated with *cas* genes that belong to more than 50 families and form combinations specific to each type and subtype of CRISPR-Cas systems [2, 3, 15, 16, 17, 18]. Some CRISPR-Cas systems are associated with additional genes whose roles in defense or other functions of CRISPR-Cas remain to be established [19, 20]. The Cas proteins are generally encoded in single operons adjacent to the respective CRISPR array although there are many exceptions to this pattern.

CRISPR-Cas systems are currently classified into two classes, six types (I-VI) and more than 30 subtypes. This classification is based on several criteria, primarily, the identity, co-occurrence and arrangement of the *cas* genes in CRISPR-cas loci [2, 3, 18]. The organization of the adaptation module is largely uniform across the diversity of the CRISPR-Cas systems. Proteins

involved in adaptation are the integrase (endonuclease) Cas1, the structural subunit Cas2 and, in some systems, another endonuclease, Cas4, which facilitates the integration of PAM-compatible spacers [21]. By contrast, the protein machineries involved in the crRNA biogenesis and interference are highly diverse. The two classes of CRISPR-Cas systems differ with regard to the complexity of the effector ribonucleoprotein complexes. Class 1 effectors consist of several structurally and functionally different subunits, some of which possess the pre-crRNA cleavage nuclease activity, and accommodate the crRNA to recognize the protospacer in the target DNA molecule. The target cleavage activity resides either in one of the subunits of the effector complex itself or in a separate nuclease that is recruited by the effector complex when bound to the target DNA [13, 14, 22, 23]. By contrast, Class 2 systems utilize a single, large, multidomain protein, such as Cas9, Cas12 or Cas13 that is responsible for the target recognition and cleavage and, in many case, also for the pre-crRNA processing [2, 13, 18]. The CRISPR-Cas types and subtypes further differ in the repertoires of proteins and domains involved in pre-crRNA processing and interference.

The sequence and structure of the CRISPR array itself is usually disregarded in the classification of CRISPR-Cas systems because, although CRISPR repeat organization can affect the evolution and diversity of CRISPR arrays, the relationship between repeat families and associated CRISPR-Cas types and subtypes is complicated [2, 3, 18, 24, 25, 26]. In particular, nearly identical CRISPR-repeats in the same species can be associated with different subtypes [27, 28]. Nevertheless, CRISPR arrays or, more precisely, the constituent repeats can also be classified into families, capturing the diversity of sequence motifs and RNA secondary structures [24, 25, 26, 29, 30, 31]. The existing approaches for CRISPR array identification, such as CRT [32], CRISPRCasFinder [33], PILER-CR [34], minced (https://github.com/ctSkennerton/minced) and CRISPRDetect [35], rely mainly on the repetitive structure of arrays. CRISPRCasFinder, for example, uses Vmatch (http://www.vmatch.de/) [36] to find pairs of maximal repeats, which are then joined into a consensus repeat sequence. The array defined by this consensus repeat sequence is then scored with a built-in scoring function that takes into account the lengths of the repeats and spacers, similarity between the repeats and other features. The other approaches, while differing in details, rely on a similar overall detection approach.

Here, we present a machine-learning (ML) approach for the processing and evaluation of repetitive elements in order to detect CRISPR arrays. For the repeat scoring, we do not rely on a manually curated evaluation function but rather use a ML method to learn this evaluation function from data. For this purpose, we constructed an extensive set of training and test data including both bona fide CRISPR arrays and negative controls. This allows us to substantially increase the sensitivity and to drastically improve the specificity of CRISPR array detection, while providing the user with a reliable quality measurement. The problem of false positive predictions was not thoroughly investigated in the existing CRISPR detection tools. We show that our ML-based approach has a much lower false positive rate than the other CRISPR detection methods.

MATERIALS and METHODS

The task of CRISPR array detection consists of the search for repetitive elements in a genome, which might form the repeats of a CRISPR array, followed by the evaluation of the identified putative array. The evaluation criteria can include the length of the repeat consensus sequence, the similarity between the repeats within the array, and features of spacers and nearby *cas* genes. The existing approaches use manually curated evaluation functions that combine these features into a score. CRISPRDetect, for example, calculates the score for an array by adding up positive and negative score contributions, such as +3 if the repeat, which is required to be longer than 23 bases, contains a known motif at the end, and a metric for identity among the spacers (-3 to +1), or -1.5 for dissimilarity between repeats. Arrays with scores above 4.0 are then classified as positive. Although such manually curated scoring functions can work well for the detection of positive examples, they have several disadvantages. First updating the scoring system to a growing database is a complicated task. Second, there is no built-in mechanism to balance sensitivity and specificity. Our analysis described here shows that, whereas the sensitivity of the existing tools is usually high, these tools do not adequately control for the false positive rate and thus have a relatively low specificity.

In order to control for both sensitivity and specificity, we set up the CRISPRidentify pipeline, which replaces the manually curated scoring function with an ML approach. The CRISPRidentify pipeline consists of two major steps, namely, a new approach to detect and generate array candidates, aiming at increased sensitivity, and a versatile, machine-learning based evaluation of the detected candidates to increase the specificity. With regard to the first task, any approach to detect CRISPR arrays faces two main subproblems: 1) accurate identification of the start and end positions for each repeat candidate in the array, and 2) accurate identification of the boundaries of the whole array. Mismatches in the repeat sequences or similarity of some spacers around the start and end positions can lead to incorrect identification of the repeat sequence itself. Incorrect identification of the array boundaries could be caused, for example, by degradation of the repeat sequence at the end of the array which makes it difficult to determine the array end point.

The second major task is the evaluation of the candidates. CRISPRidentify learns a classifier for CRISPR arrays from known data to control for the false positive rate, in contrast to the manually curated scoring functions used by the existing tools. In this setting, it is natural to rely on: 1) features that are extracted from the array for estimating the similarity between bona fide CRISPR arrays and array candidates, 2) a benchmark set of positive and negative array examples, and 3) machine learning to learn the evaluation function from this benchmark. We define new training and test sets consisting of carefully selected positive and negative examples. Using these training sets, we learn a classifier for evaluating CRISPR candidates based on 13 array-derived features (see Table S2 in Supplementary), such as the similarity between the repeats, AT-content or stability of the repeat hairpin. The principal advantage of this approach is that we not only increase the sensitivity of CRISPR array detection, but also drastically reduce the false positive rate, i.e. increase the specificity. The problem of false positive predictions has been not properly addressed by the previous approaches. Finally, the pipeline annotates the array with additional information such as its orientation, leader sequence and *cas* genes.

Benchmark datasets

Overall, we prepared six benchmark datasets to test different aspects related to the detection of CRISPR arrays. In particular, we compiled the archaeal and bacterial arrays [29, 30, 37]. The first dataset contained 400 archaeal arrays among which 10 were experimentally validated and the rest were manually verified (Table 1). The second dataset containing 600 bacterial arrays was constructed from published datasets [29, 30, 37], with the constraint that the corresponding repeat sequences should be at least 85% identical to published experimental CRISPR-repeats. For simplicity, we will refer to these datasets as **archaeal** and **bacterial** datasets, respectively. As an independent, high-quality, non-overlapping test set, we additionally selected 550 CRISPR arrays from archaeal and bacterial genomes using the following criteria: 1) the minimal number of repeats in each CRISPR array is three, and 2) the maximum distance between the CRISPR array and the *cas* gene operon is 500 nucleotides. We will call this the **Cas-associated** dataset (See distribution of Cas types in Supplementary materials table S7).

The first three datasets were partially used by other tools, but the fourth one was designed specifically for this work and is required to control the false positive rate. It contains 300 (100 training and 200 test) false CRISPR arrays and was constructed based on the tandem repeat dataset and cases with identical 'spacers' (see Table S2, Table S3 and Table S4 in Supplementary for more details). For simplicity, we will now refer to this set as **false arrays**. To obtain more negative examples, we also created additional distorted CRISPR arrays from the "archaeal dataset" and the "bacterial dataset" by independently shuffling the nucleotides in each repeat and each spacer sequences. We will refer to these as **archaeal shuffled arrays** and **bacterial shuffled arrays**, respectively. All generated datasets where split into training and test sets.

In addition to the datasets we generated, we used two datasets from literature. The first, most comprehensive one consists of all arrays annotated from the most recent CRISPR classifications [2, 3, 18, 19], and will be called **annotated arrays dataset**. It consists of 39,371 arrays. The second dataset was downloaded from CRISPRCasFinder [33]. It contains 3263 arrays and will be denoted as **CRISPRCasFinder** dataset.

Construction of Array candidates

We use Vmatch version 2.3.0 to scan a genomic region for the occurrences of repeat candidates. Vmatch returns a list of putative repeat pairs. By default, in CRISPRIdentify, we set the substring length for Vmatch to be in the range of 21 to 55 nucleotides, and the spacing between the matching pairs has to be in the range between 18 and 78 nucleotides. Both intervals can be changed by the user. To avoid duplicates, we filter out all the repeat candidates except for the unique ones.

The next step is to cluster the repeat matches found by Vmatch, and to extract the bona fide repeat sequence. As a proxy, the repeat sequence is the consensus sequence of all repeat matches belonging to a repeat cluster. However, using a consensus sequence to determine possible repeat candidates poses two problems. First, the length of the consensus sequence is uncertain because a threshold has to be defined for conserved columns, and second, we lose control over the number of mismatches to existing repeat candidate sequences in the genome. Thus, instead of using a single consensus sequence, as the first stage, we generate a set F over each Vmatch cluster containing all possible repeat candidates and their

extensions. The set F is a subset of all subsequences of a maximal consensus sequence partially ordered by the subsequence relation (see Figure 1, box 'Extension', for an example).

To generate F, we use the alignment for each cluster to construct the maximum element of F as a string that is composed, column-wise, of the nucleotides most frequently found in the alignment. The minimum element of the set F is defined as the most consistent substring, and is a subsequence of the maximal element by definition. After the maximum and minimum elements are calculated, we generate all sequences which fulfil the requirements that they 1) are subsequences of the maximum element and 2) contain the minimal element as a subsequence. These sequences then complete the set F. Mathematically, F is a filter (the dual concept of an ideal) in the partial order defined by the subsequence relation. As the last step, the set of repeat candidates is formed via unification of sets F and initial Vmatch results and filtering of duplicates (see Supplementary Materials, Section 1 "E*nhancement of Vmatch results"*). Lastly, we extend the set of repeat candidates with modifications of each repeat candidate in which up to 3 nucleotides are omitted from both ends, resulting in 15 variations of each repeat. After that, duplicates are filtered out for the last time and the remaining candidates are added to the pre-existing clusters. These clustered repeat candidates are then used to form CRISPR candidates.

After the set of potential repeat candidates is extracted, we need to form the bona fide CRISPR array candidates. Unlike the other existing algorithms, where a CRISPR array is formed from a repeat consensus by simply splitting the considered interval into repeats and spacers, our approach utilizes approximate string search to find the locations of the repeat sequences using the python regex library available on PyPI regex. This approach also has an important advantage over simple partitioning of the DNA interval into repeats and spacers because this technique allows incorporating insertions and deletions as editing operations, finding the locations that minimize the number of editing operations. This feature allows our approach to effectively identify potential CRISPR candidates including cases with abnormalities, such as truncated repeat [38, 39, 40] sequences or complete spacer deletion (see section 7 Spacer deletion and section 8 Degenerated repeat in the Supplementary materials).

Feature Vector Associated with an Array Candidate.

Having obtained suitable CRISPR candidates, we generate a feature vector for each candidate to use at the evaluation stage of our pipeline. Users may specify the set of desired features by choosing one or several of the three classifier models we developed. Each of these models is defined by feature selection and encompasses a different combination of the 13 features we devised for classifying CRISPR arrays. When designing these features, we had three conceptual categories in mind, depending on whether we expect their values to be positively or negatively correlated with the likelihood of a true CRISPR array, or whether they provide supporting information. These are, however, only conceptual categories that are not further used during the machine learning approach. Among the factors, which we expect to augment the evidence for a true CRISPR array with their growing value, are **number of repeats**, **length of the repeat sequence** and **repeat similarity** because the more repeats are found, and the longer and more similar they are, the stronger the evidence that the result did not occur by chance.

On top of these features, we compute the RNA minimum free energy (MFE) of the consensus repeat sequence. The negative MFE is called the **MFE score**, and lower MFE scores are positively correlated with the likelihood of a true CRISPR array [22, 27, 28, 41, 42, 43] in some systems. The reason is that, in many cases, for instance, in the vast majority of type I cases, the crRNA tends to form stable stem-loop structures, due to the palindromic organization of the repeat sequences. Therefore, a low MFE score can be used to identify stem-loop structures in these systems, thus buttressing the evidence for true crRNA and, by extension, CRISPR array. The MFE score was computed using RNAfold [44]. Another feature positively correlated with the likelihood of a true CRISPR array is high **similarity between the CRISPR candidate repeat sequence and a verified repeat sequence in the database**. The similarity between a given sequence and the sequences in the database is computed using BLAST [45], which is provided as executable with our pipeline. In our approach, two BLAST similarity scores for each CRISPR candidate were computed. The first score reflects the highest similarity between a candidate and all experimentally validated repeat sequences. The second score measures the similarity to a modified version of the same dataset to which 1 or

2 mismatches were added, thus extending the range of similarity considered by the first BLAST score. These two scores were treated as two separate features.

In contrast to the first group of features, the values of the following features are negatively correlated with the likelihood that a candidate is a true CRISPR array. The first such feature is the **number of mismatches** because repeats are usually highly conserved. Furthermore, spacers of the CRISPR array candidate were assessed in terms of length and similarity. Because in bona fide CRISPR arrays, the spacer length is typically uniform, a high variance in spacer length negatively correlates with the likelihood of a CRISPR array. With respect to similarity, spacers differ from repeats in that they are generally not significantly similar to each other given their origin from different virus or plasmid genomes. Thus, high level of **similarity between spacers** (see Figure S2 in Supplementary) is negatively correlated with the likelihood of a candidate being a CRISPR array. Another feature that is included in our pipeline is whether a candidate region contains a protein-coding sequence. Because CRISPR arrays do not code for proteins, the presence of coding sequences is negatively correlated with the likelihood of an interval being a CRISPR array. Our pipeline identifies Open Reading Frames (ORFs) using Prodigal [46], again provided as executable with our pipeline. If ORFs were found, the result with the highest confidence value was taken as the feature value. To account for protein coding-regions even more thoroughly, our tool also looks for tandem repeat proteins. To this end, the Prodigal output is fed into a Hidden Markov Model [47], which searches for similarity with the PFAM database of tandem repeat proteins. The resulting **tandem protein similarity** score serves as a feature negatively correlated with the likelihood of the interval being a CRISPR array.

In the third category, our approach computes two features whose growing value is not directly correlated with the likelihood of the interval being a CRISPR array but could potentially guide the likelihood in different ways. The first such feature is the AT-content of the identified repeat sequences within a given CRISPR array candidate. Because bacteria and archaea differ in **AT-content**, this feature can help to define implicit subclasses that allow the classifier to model different properties for bacterial and archaeal arrays. The other additional feature is **average spacer length** (see Figure S1 in the Supplementary Materials). This feature is important because, in contrast to repeats, spacers in the spacer array can have different

lengths. Overly short or long average spacer length, however, might signal a false array, which is why such outliers should be considered in the scoring of the candidates. After all the features determined by the classification model were calculated for each CRISPR array candidate, the resulting up to 13-dimensional feature vector was subject to machine learning classification as described below.

Feature Selection

After the first successful results using all thirteen features described above, we went on to evaluate which of the features are actually necessary. This can be determined by feature subset selection (FSS), which generates a subset of features best suited for classification. For this purpose, we applied a wrapper approach [48], which works as follows. We generated all $\sum_{i=1}^{13}\binom{13}{i} = \sum_{i=1}^{13}\frac{13!}{i!(13-i!)} = 8192$ subsets of the full sets of 13 features. We then trained for each of the 8192 subsets a model using only the features from the respective subset, and assessed the prediction quality. The results of these experiments showed that a subset of the 13-dimensional feature vector is sufficient to reach the 99.8% accuracy that is achieved with the complete feature set. The same accuracy levels could be achieved with three of these models. These models included 8, 9 and 10 features, respectively, with each feature appearing in at least one of the models. Although the feature combinations overlapped, they were not subsets of one another (see Table S1 in Supplementary for the models and their feature subsets).

Classification of CRISPR arrays using machine learning

The last step of our CRISPR identification pipeline consists of the evaluation of the obtained CRISPR candidates. For this purpose, we integrated the implementation of the Extra trees classifier from the python Scikit-learn package [49] into our pipeline. We provide the user with an opportunity to pick either of the three pretrained classifiers that we describe in the Feature Selection and Importance Analysis subsection or use all of them simultaneously. It is important to notice that we do not use the class prediction directly but take the label pseudo-probability and treat it as a certainty score. In case of using many classifiers, the score is computed as the average of scores. Based on the distribution of the output CRISPR array scores, we introduced two decisive score values around which we structured the labels: 0.75

and 0.4. CRISPR arrays with scores greater than or equal to 0.75 were labeled to “candidate”, meaning that one of them likely corresponds to a true CRISPR array. CRISPR arrays with scores between 0.75 and 0.4 were assigned to the group “potential candidate”, meaning that there is a smaller probability for one of them to be a true CRISPR array. The CRISPR candidates with scores lower than 0.4 were assigned to the group “low score candidates”, indicating that they are probably not valid CRISPR arrays. After performing the grouping, the algorithm picks the “bona fide candidate” as the candidate with the highest score among the candidates in the “possible candidate” group. Then all the candidates are reported in files corresponding to their groups. Hence, the user may not only inspect the bona fide candidates but also all the alternatives with their respective statistics.

Repeat and Spacer databases

Our tool can also be used to build a user-specific CRISPR array databases. In order to illustrate the potential of this feature, we used data from archaea and bacteria genomes. The archaeal dataset consisted of 251 complete and 736 incomplete genomes, while the bacterial dataset included 1693 complete genomes and 25335 incomplete ones. Having identified CRISPR arrays in each dataset with our tool, we extracted the complete set of spacers as well as the consensus repeat sequence for each CRISPR array. Using these results, we built two databases, one for the identified repeats and the other one for spacers. The output of the repeat database lists, for each consensus repeat sequence, all genomic locations, the organisms, as well as the number of locations and organisms where an array with this consensus repeat can be found (see Table S6 and Figure S6 in Supplementary).

RESULTS AND DISCUSSION

Existing CRISPR Datasets Contain Many Suspicious Arrays

The main drawback of the current tools for CRISPR array identification is the lack of scoring of the obtained results, i.e. absence of a thorough evaluation of the likelihood of each reported array to be a true CRISPR array. We suspected that lacking a stringent quality evaluation might enrich existing datasets with false-positives. A high prevalence of questionable results produced by the commonly utilized CRISPR identification tools could pose a problem not only to experimental researchers, which might invest effort in the investigation of false predictions, but also to tool developer using these datasets as standards [50].

One type of spurious arrays that we found to be common in public databases are arrays that contain damaged repeats or spacers with unusually high similarity to one another. In these cases, a possible explanation is a duplication or repetitive structure stemming from a source other than CRISPR. Consider the example in Figure 2, which is an array that was predicted as positive by the CRT tool. As shown in Figure 2a, the array contains a damaged repeat as well as two nearly identical spacers that differ by only one nucleotide. Further examination showed that the array lies within a predicted protein-coding sequence on the opposite strand, suggesting that this is not a functional array. Moreover, the CRISPR locus is in a region that is predicted to be intrinsically disordered in the protein encoded on the opposite strand, and the repeat (and part of the spacers) are predicted by the Anchor tool [51] to be linear interacting peptides (Figure 2b). Thus, in this case, the repetitive structure at the nucleotide level, most likely, originates from a repetitive structure at the protein level, and accordingly, is not a functional CRISPR array. Other suspicious cases showing, for example, low repeat similarity or a highly non-uniform distribution of spacer lengths can be found in Section 9 "*Spurious Arrays in Existing Tools*" and Figure S3, Figure S4 and Figure S5 of the Supplementary material.

To assess the scale of the problem, we investigated the 3263 array candidates in the CRISPRIdentify benchmark set and systematically searched for suspicious arrays that are predicted as positive by CRT, CRISPRDetect and CRISPRCasFinder. We considered an array to

be suspicious if it had at least one of the following properties: 1) a high similarity between spacers (sequence identity > 0.6), 2) mutations in repeats (identity between repeats < 0.8), 3) similarity between repeat and spacer (identity between repeat and spacer > 0.5), or 4) an uneven spacer length distribution. The spacer length was considered to be distributed unevenly if at least one spacer was 70% shorter than the average. To exclude any bias in the tool parameter setting, we considered only arrays that could be handled by all tools. Thus, we filtered out the cases with only two repeats, or arrays with repeats shorter than 23 or longer than 55 nucleotides as this setting is covered by all tools. Note that each tool predicted slightly different locations overlapping these arrays. Altogether, we obtained 1358 arrays that passed this filter and were predicted as positive by CRT, 1395 by CRISPRCasFinder and 346 by CRISPRDetect. Among these predictions, 602 arrays predicted by CRT, 628 arrays predicted by CRISPRCasFinder and 110 arrays predicted by CRISPRDetect were deemed suspicious under the above criteria (see Table 3). This corresponds to potential false positive rates of 44.3% (CRT), 45.0% (CRISPRCasFinder) and 31.8% (CRISPRDetect) (see Table S5 in Supplementary). Although these numbers are upper bounds because it cannot be ruled out that a suspicious array is actually functional, they clearly show that false positive predictions are a substantial problem for all available methods for CRISPR detection.

The CRISPRIdentify Pipeline using a machine-learning based scoring system

To overcome the problem of false positives, we set up the CRISPRIdentify pipeline (see Figure 1). It consists of two steps, namely, a sensitive approach to detect array candidates, and rigorous evaluation of the detected candidates using machine learning.

In the first step, we used the Vmatch tool to detect pairwise matches of a repeat, from which we reconstruct array candidates by combining compatible pairwise repeat matches. Vmatch builds an efficient computational index structure named suffix arrays that allows for a fast search for matching substrings in the DNA sequence. Pairwise matches that fulfill predetermined length and spacing requirements based on reliable observations are then processed further (see Methods for details). The principal difficulty at this step is that assignment of the leading and trailing nucleotides of the spacers can be ambiguous. Thus, mismatches in the repeat sequences or similarity of some spacers around the start and end

positions can lead to incorrect identification of the repeat sequence itself. Consider the following repeat composition:

$$\ldots \mathbf{C}^{\underline{\text{repeat}}}\mathbf{A}^{\underline{\text{spacer}}}\mathbf{G}^{\underline{\text{repeat}}}\mathbf{T}^{\underline{\text{spacer}}}\mathbf{G}^{\underline{\text{repeat}}}\mathbf{A}\ldots$$

It remains unclear whether the G- and A-nucleotides coloured in green and red, respectively, belong to the repeats, or to the spacers that happen to contain the same first (or, respectively, last) nucleotides in two cases. Repeat finding tools, such as Vmatch, find pairwise matches, which are likely to be copies of the same repeat. Thus, in our example, Vmatch would return the following pairwise matches that could all be instances of the repeat: $^{\underline{\text{repeat}}}$, $\mathbf{G}^{\underline{\text{repeat}}}$ and $^{\underline{\text{repeat}}}\mathbf{A}$. This situation results in three distinct potential array variants. To identify these cases, we first cluster the pairwise matches representing repeat candidate sequences based on the distance between them. However, the correct repeat candidate might not be included in the Vmatch results due to potential similarities in the beginnings or ends of spacers, or mismatches between repeat sequences. In the simplified example above, the candidate $\mathbf{G}^{\underline{\text{repeat}}}\mathbf{A}$ would not have been found as an exact pair match because there is no second example of this form in the genome. To overcome this problem, we extend each cluster in an enhancement procedure by additional repeat candidates. First, we generate a set F over each Vmatch cluster containing all possible repeat candidates and their extensions. To generate F, we use the alignment for each cluster to construct the maximum element of F as a string that is composed, column-wise, of the nucleotides most frequently found in the alignment. The minimum element of the set F is defined as the most consistent substring, and is a subsequence of the maximal element by definition. The set F is a subset of all subsequences of a maximal consensus sequence partially ordered by the subsequence relation (see Figure 1, box candidate extension). Mathematically, F is a filter in the partial order defined by the subsequence relation. It contains all subsequences that contain the minimal repeat element and are itself subsequences of the maximal element (see Methods and Figure 1, box titled "Extension", for details).

As a result, the first step of our pipeline produced multiple CRISPR array candidates for the same genomic region. Thus, an essential second step is to evaluate all these candidates under stringent criteria. On the one hand, this evaluation guarantees that the best candidate is selected for each genomic region. On the other hand, the task of the evaluation is to exclude

repetitive patterns that might be detected as CRISPR arrays by other tools but show otherwise little similarity with bona fide CRISPR arrays, resulting in a better control over the false positive rate. For this purpose, it appears natural to rely on 1) a benchmark set of positive and negative array examples, 2) features that are extracted from the array to estimate the similarity between bona fide CRISPR arrays and array candidates, and 3) machine learning to learn the evaluation function from this benchmark.

The positive examples for the benchmark set are easily generated by selecting an appropriate subset of annotated CRISPR arrays (see Methods). The reconstruction of negative examples, however, is more challenging as there is no curated benchmark set for this purpose. Here, we constructed a negative data set based on the tandem repeat dataset and cases with identical 'spacers' (see Methods for more details).

For the purpose of ML-based evaluation, our pipeline associates to each array candidate a feature vector, which is a set of numerical values (features) that describe different properties of the array. In our pipeline, the feature vector consists of a maximum of 13 features, including repeat-associated features, namely, **number of repeats**, **length of the repeat sequence**, **repeat similarity, number of mismatches, AT-content**, and the **RNA minimum free energy (MFE)** of the repeat, which is indicative of a stem-like structure present in the vast majority of type I arrays. We added also two **BLAST scores** that determine the similarity of the identified repeats to known repeat sequences. Spacer-related features are **similarity between spacers, average spacer length** and **spacer uniformity,** i.e. a measure of the variability of the spacer length. Additionally, we included two features that are associated with the whole array, namely, an **ORF score,** which measures whether the array overlaps with a protein-coding region, and a **tandem protein similarity score,** which is used to detect tandem repeat proteins (see Methods and Table S2 in Supplementary for details). Once the feature vector was associated with a repeat region, a trained classifier was used to decide whether an array candidate is likely a CRISPR array, and the score for this likelihood was calculated.

CRISPRIdentify not only defines and scores CRISPR arrays, but provides the user with more detailed information on the respective CRISPR-Cas systems for further investigation. Thus, we use CRISPRstrand [30] to predict the orientation of the array, and CRISPRleader [37] to predict

the start of the leader sequence. In addition, the results of the CRISPR array analysis are complemented with the information on the detected adjacent *cas* genes. We used Prodigal [49] for the protein prediction, with subsequent classification using Cas-specific Hidden-Markov-Models (HMM). An important aspect of our annotation is the analysis of degenerated repeat elements, as well as detection of insertion sequence (IS) elements, a type of transposable elements that can influence the prediction of CRISPR arrays. A degenerated repeat element can be missed during the CRISPR array identification due to a high number of mismatches exceeding the detection threshold. To overcome this problem, we apply an additional search for degenerated repeat sequences using the approximate search method from the CRISPR identification part.

The IS elements are widespread in archaeal and bacterial genomes [52] and thus can disrupt CRISPR arrays. Integration of an IS element into a CRISPR array would mislead existing CRISPR detection tools to predict two or more distinct CRISPR arrays, despite the relatively close positioning and identity of the repeat sequences. To detect IS-elements, the known IS sequences were clustered using the Markov Cluster Algorithm (MCL) [53] and an HMM model for each cluster using hmmbuild [47]. When two neighboring CRISPR arrays with identical repeat sequences were detected, Prodigal was used to extract protein sequences from the gap region, and the resulting sequences were searched for the presence of an IS-element using the IS-HMM models. If an IS-element is was found, our pipeline has an option to merge the separated parts of the array (see the *Degenerated repeat and IS element search* Section in the Supplementary Materials for an example).

CRISPRIdentify Outperforms the Existing Tools thanks to Machine-learning Based Evaluation

We investigated different types of machine learning (ML) approaches, including Support Vector Machine, Naive Bayes, K-nearest neighbors, Fully Connected Neural Network, Decision Tree, Random Forest classifier and Extra-trees classifier, for the use in CRISPRIdentify. The performance analysis for the CRISPRIdentify pipeline consists of two parts. First, we determine which of the ML-approaches performs best by checking the classification performance of each. Then, we compare the selected, best ML-approach to the existing tools, using the built-in scoring function and thresholds of these tools as the classifier. To avoid any

bias from the training in the comparison with other tools, we split the positive and negative datasets into a **train&validate** dataset to determine the best ML-approach and parameters for CRISPRIdentify, and an independent **test** dataset that was only used for the comparison to other tools (see Table 1 and Methods for details on benchmark sets). With the **train&validate** dataset, we used a 10-fold cross-validation strategy to assess the accuracy of the different ML-approaches. As the **train&validate** dataset is balanced (see Table 1), reporting the accuracy is sufficient to assess the performance of each approach. As can be seen in Figure 3, ensemble methods, in particular, Extra Trees, which learn a whole ensemble of simple classifiers and output the average estimate for all of them as the overall classification result, clearly outperform the other classifiers with a median accuracy of 0.91. Thus, we employed Extra Trees for the further analyses.

Once we have selected and trained Extra Trees on the **train&validate** set, we used the untouched **test** set (see Table 1) to compare the performance of our method using the chosen classifier with those of other tools. Because our test set is unbalanced due to the smaller amount of negative data, we report, in addition to the accuracy (ACC), also the balanced accuracy (BACC) and the Mathews correlation coefficient (MCC, see Figure 3 and Table 2), as these prediction quality measurements are better suited in this case. The high true positive rate (TPR) demonstrates the high sensitivity of all approaches, with CRT having the lowest sensitivity (TPR=0.93) and our tool, CRISPRIdentify, having the highest sensitivity (TPR=0.99). The specificity, however, is drastically different between CRISPRIdentify and the existing tools. As indicated above, means to reduce the false positive rate (or equivalently, increase the true negative rate, as FPR=1-TNR) was not investigated in great depth for the existing tools. Thus, CRISPRIdentify has the highest possible TNR of 1, whereas CRT has a low TNR of 0.42, and CRISPRCasFinder and CRISPRDetect show intermediate specificities (TNR of 0.64 and 0.74, respectively). The relatively low specificity of the other tools is also manifested in their low balanced accuracy and Mathews correlation coefficient.

We then investigated the false positives in more detail to see whether a particular type of arrays was preferentially misclassified by a group of tools. Examination of the overlap between different tools shows that the majority of the false positives are unique for a specific tool (see Figure 4). From the 200 false arrays of our test set, all have been misclassified (that

is, recognized as CRISPR) by at least one of the existing tools, and none by CRISPRIdentify. A substantial majority of the false arrays (159, or 79.5%) were misclassified by exactly one tool, and only one array was misclassified by all three tools. In addition to excluding a specific bias, these observations support our selection of false arrays as true negatives. Among the tested tools, CRT had the highest false positive rate (117 false positive predictions (FP), FPR=58.5%), followed by CRISPRCasFinder (72 FP, FPR=36%) and CRISPRDetect (53 FP, FPR=26.5%) (See feature distribution of false predictions in Figure S12, Figure S13, Figure S14, Figure S15 and Figure S16). The CRISPRIdentify had, as stated before, a zero false positive rate.

We then investigated the true positive predictions, i.e., correctly reported CRISPR arrays, as well as false negative predictions, in more detail. We employed the full, comprehensive dataset of all **annotated arrays** from the most recent CRISPR classifications [2, 3, 18, 19] (see Methods) to obtain a wider picture of arrays that remained undetected by different tools. This dataset or a larger part of it were not used for training the parameters for several reasons: 1) the dataset would be highly imbalanced, 2) we would be unable to perform the same quality control, due to the dataset size, as for the **train&validate** dataset, and 3) the arrays are probably largely predicted with one of the existing tools. The last point is problematic as it would bias our tool to mimic the prediction tool that was mostly used to annotate these arrays. As shown in Figure 5, the sensitivity is high for all existing tools, but nevertheless, CRISPRIdentify performed the best (99.8%), followed by CRISPRCasFinder (98.9%), CRISPRDetect (94.5%) and CRT (93.5%). In this test, CRISPRIdentify could detect 147 previously annotated arrays [2, 3, 18, 19] that were not detected by any other approach. Overall, only 89 arrays, 7% or the total of 35310, were predicted by all tools, indicating considerable variation among the predictions. In summary, these results show that, in CRISPRIdentify, we increased the already high sensitivity of the existing approaches, while substantially improving the specificity as well.

Both Expected and Unexpected Features are Important for CRISPR Identification

To assess the importance of different features, we first performed feature selection to remove non-informative features. This procedure also helps avoid overfitting to the training data. Thus, we generated three subsets of features including 8, 9 and 10 features,

respectively, that are sufficient to reach the 99.8% accuracy. Although the feature combinations overlapped, they were not subsets of one another (see Methods for details, and Table S1 in Supplementary for the models and their feature subsets).

We then investigated how much each feature contributes to the prediction quality. Specifically, we used the ReBATE algorithm since it showed state of the art performance for the similar task [54]. We ran the algorithm on the training set with different numbers of the nearest neighbors within the interval [1, 100], to compute the relative importance of each feature. We then compared the importance values obtained with the training set with the feature importance of the trained model. To this end, we sampled 10,000 CRISPR candidates from the comprehensive dataset and re-ran the ReBATE algorithm with the same settings. Despite minor differences, both runs showed similar results with slightly different orderings of spacer similarity, spacer uniformity and AT-content (see Figure 6).

Overall, as could be expected, repeat similarity is clearly the most important feature. The MFE-feature is also expected to contribute substantially because the repeat forms a stem-loop structure that is required for processing in the vast majority of type I systems. The AT-content did not come across as a direct decision-making feature, i.e., there was no correlation between AT-content and array quality. Thus, this appears to be a supporting feature defining subclasses that can help the classifier to compensate for the differences between archaeal and bacterial arrays, probably, also in combination with the MFE feature. The blast score is important because similarity of the repeat sequence to a known repeat is a clear sign of a true array. The importance of this feature, however, is not high enough to jeopardize the generalization capacity of the trained model. Repeat length is also an important feature as the majority of known arrays in bacteria show a narrow distribution of the repeat length, with peaks around 28 and 36 nt, and in archaea, there are three peaks of repeat lengths (Figure 7). We also investigated the distribution the repeat lengths in the false arrays, finding that this distribution is somewhat broader than the distribution for the bona fide arrays, but covers a similar length range (Figure 7). This also shows that our set of false arrays shares basic properties with the true ones.

Whereas the importance of all repeat-related features could be readily expected, we found that, among all features, the second and third most important ones (second and fourth in the comprehensive data set, see Figure 6) are related to the spacers. Spacer similarity was one of the criteria used for constructing the negative set and is thus negatively correlated with the likelihood of the candidate being a true CRISPR array. The reason is simply that the presence of many (nearly) identical spacers in an array is strong evidence of a non-CRISPR repetitive element. However, we observed extensive overlap between spacer similarity in the positive and negative training set, showing that spacer similarity alone is not sufficient to discard an array (see Supplementary Section 3). This conclusion was supported by database analysis in which we found a few cases of spacer duplication in an array. Further, we found that the spacer lengths have a sharp peak around 30-35 for bacteria (see Supplementary Section 3), which is why spacer uniformity is an important feature as well.

In addition, we used the comprehensive dataset (see Supplementary material Section 12) for a quantitative comparison of bona fide CRISPR arrays and genomic regions that exhibit similarities to CRISPR arrays but are likely to be true negatives, that is, low scoring candidates detected by our pipeline. We found that such low scoring candidates tend to have much shorter repeat sequences, with the majority having a repeat length of about 20 nucleotides. In contrast, the vast majority of bona fide arrays had longer repeats consisting of more than 30 nucleotides. Another striking difference pertained to the number of repeats per array. Whereas low score candidates consistently contained 3-5 repeats per array, bone fide arrays were distributed more uniformly, often showing cases with large (more then 30) number of repeats. Further, the spacer length distribution of bona fide arrays was heavily concentrated within the interval between 27 and 50 nucleotides. Low scoring candidates, in contrast, showed a rather uniform distribution of spacer lengths between 0 and 60 nucleotides (See Figure S7, Figure S8, Figure S9, Figure S10 and Figure S11 in Supplementary materials)

Repeat and Spacer databases

In addition to CRISPR array identification, our tool also includes a feature that allows users to build databases that collate relevant information on the identified arrays. The advantage of this database is that it classifies CRISPR arrays into different categories (bona fide, alternative, possible, low score) and provides detailed information on each array, such as the repeat sequence, the spacer elements, the array score, its orientation, and other characteristics (see Methods for details).

CRISPRidentify can handle common challenges faced by existing tools

Multiple studies show that, in many functionally characterized CRISPR arrays, there is notable accumulation of point mutations in the repeats at the 3'-end [38, 39, 40]. To assess the frequency of such mutations, we investigate the bona fide arrays obtained by CRISPRidentify in a comprehensive genome dataset (see Supplementary Section 12). We filtered out cases with an overall repeat similarity below 0.8 and repeat number below 3, in order to focus on readily identifiable substitutions in the first or last repeat. The array orientation was determined using CRISPRstrand. In an effort to attain an even higher specificity, we introduced two groups of mutated terminal repeats: 1) damaged, where the number of substitutions in the first or last repeat of the array did not exceed 5, and 2) severely damaged (more than 5 mismatches with the consensus repeat sequence), where we suspect degenerated repeat sequences in the array [10, 11, 54] (here, using CRISPRidentify, we found that substitutions in 3'-end repeats were frequent, appearing in 10425 of the 39378 analyzed arrays (27%). We also noticed that severe damage occurred in nearly a half of the mutated repeats, 5125 (13%) (See Supplementary materials Table S8). We examined how the existing tools could handle the identification of the repeats with mismatches by checking for the predicted start and end positions for each array. Of the arrays with mismatches detected by CRISPRidentify, CRT, CRISPRCasFinder and CRISPRDetect reported 7974 (76%), 8523 (82%) and 387 (4%) cases in the damaged category, respectively, and 3518 (68%), 3714 (72%) and 8 (0.1%) cases in the severely damaged group, respectively. While CRT and CRISPRCasFinder yielded similar results, CRISPRDetect reported many fewer cases. This difference is due to the conservative approach adopted by CRISPRDetect. Despite its ability of tolerate mismatches, CRISPRDetect often does not include the damaged repeat in the candidate set. In contrast, CRT and CRISPRCasFinder can both add heavily damaged repeat candidates in the predicted

arrays (see Supplementary Section 16). Thus, regardless of the differences among the existing methods, CRISPRidentify

Another potential challenge for CRISPR array identification could be the presence of identical spacers in the array. In order to investigate this case, we examined bona fide CRISPR arrays obtained by CRISPRidentify from a comprehensive genome dataset (see Supplementary Section 12) as well as CRISPR arrays from the training and test sets for the classification. We took into account two possible configurations: first, a sequence of identical spacers forming a “cluster” of consecutive elements, and second, sparse identical spacers separated by other, distinct spacers. For the first case, our tool identified 2391 arrays, and for the second case, 1444 arrays of the total 39378 arrays analyzed (see Table S9 in Supplementary materials section 18, also see Table S10 and Table S11 for the analogous distribution on the train and test sets). The majority of the identified arrays with identical spacers contained less than 3 spacer duplicates. This result implies that, although our approach penalizes high similarity between spacers, it nevertheless can detect arrays with a limited number of repeated spacers, while discarding those in which many spacers are identical. In contrast, the existing tools do not evaluate the detected CRISPR arrays for spacer similarity and therefore can report spurious candidates with large clusters of identical spacers. Indeed, we examined the results of CRT, CRISPRCasFinder and CRISPRDetect, and found that all these tools reported some spurious CRISPR array candidates with large clusters of identical spacers (>5) (See Supplementary Section 18).

Finally, our approach is able to identify CRISPR array candidates with complete spacer deletion, in contrast, to CRT and CRISPRCasinder which do not support the search for such cases. We investigated the frequency of such cases in the comprehensive CRISPR array dataset collected with CRISPRidentify and found that 152 of the 39378 identified arrays contained spacer deletion (see Supplementary Section 12).

Tool Availability

The CRISPRIdentify is available as an open-source tool in Github

https://github.com/BackofenLab/CRISPRIdentify

It is implemented and tested in Python version 3.7.6 and ML models exported following the model persistence from scikit-learn.

https://scikit-learn.org/stable/modules/model_persistence.html

CRISPRIdentify accepts a complete or partial genome sequences input and identifies all possible arrays using a different mode of setting.

CONCLUSION

In this work, we present a pipeline for the detection of CRISPR arrays, CRISPRIdentify, which includes a highly sensitive method for the detection of array candidates, followed by a data-driven, machine-learning based approach for the evaluation of these candidates. We demonstrate the ability of CRISPRIdentify to accurately distinguish true CRISPR arrays from false ones. Although all existing CRISPR identification tools are highly sensitive, CRISPRIdentify outperforms them by this measure, while also being substantially more specific. Using a DNA sequence as input, our pipeline performs three steps: 1) detection of repetitive elements and construction of array candidates, 2) extraction of several array-related features, and 3) ML-based classification of array candidates based on the extracted features. The detected CRISPR arrays are then reported to the user accompanied by extensive annotation. In contrast to the existing tools, CRISPRIdentify not only provides the user with the basic statistics of the identified CRISPR arrays, but also outputs a confidence score as an easily interpretable measure of the likelihood that a given genomic locus encompasses a CRISPR array. Thus, users can immediately assess the confidence of the identification of each potential CRISPR array, which is a substantial advantage over the other CRISPR detection tools. Another key feature of our approach is its flexibility. Whereas other tools use pre-determined constant filtering criteria, CRISPRIdentify dynamically builds such criteria based on the data employed during training. Thus, CRISPRIdentify can adjust to new, previously unseen training data and use it to enhance the overall performance.

ACKNOWLEDGEMENTS

We thank the members of DFG SPP 2141 (“Much more than Defence: the Multiple Functions and Facets of CRISPR-Cas”) for constructive discussions.

FUNDING

This work was supported by the German Research Foundation(DFG), German Research Foundation (DFG) (grants BA 2168/23-1 SPP 2141 Much more than Defence: the Multiple Functions and Facets of CRISPR-Cas). The article processing charge was funded by the Baden-Wuerttemberg Ministry of Science, Research and Art and the University of Freiburg in the funding programme Open Access Publishing.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Barrangou, R. and van der Oost, J., (eds.) (2013) CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea, Heidelburg: Springer Press, pp. 1-129.
2. Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R., and Koonin, E. V. (November, 2015) An Updated Evolutionary Classification of CRISPR-Cas Systems. *Nature Reviews Microbiology,* **13**(11), 722–736.
3. Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R., and Koonin, E. V. (November, 2019) Evolutionary classification of CRISPRCas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology,* **xx**(xx), xx–xx.
4. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015 Apr 23;520(7548):505-510. doi: 10.1038/nature14302. Epub 2015 Apr 13. PMID: 25874675; PMCID: PMC4561520.
5. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJ. CRISPR-Cas: Adapting to change. Science. 2017 Apr 7;356(6333):eaal5056. doi: 10.1126/science.aal5056. Epub 2017 Apr 6. PMID: 28385959.
6. Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol. 2016 Feb;14(2):67-76. doi: 10.1038/nrmicro.2015.14. Epub 2016 Jan 11. PMID: 26751509.
7. Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.-V., Naismith, J. H., Spagnolo, L., and White, M. F. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol Cell,* **45**(3), 303– 13.

8. Deng, L., Kenchappa, C. S., Peng, X., She, Q., and Garrett, R. A. (2012) Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. *NAR,* **40**(6), 2470–80.
9. Shah, S. A., Hansen, N. R., and Garrett, R. A. (2009) Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochem Soc Trans,* **37**(Pt 1), 23–8.
10. Shah SA, Erdmann S, Mojica FJ, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol. 2013 May;10(5):891-9. doi: 10.4161/rna.23764. Epub 2013 Feb 12. PMID: 23403393; PMCID: PMC3737346.
11. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, Barrangou R, Beisel CL. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell. 2016 Apr 7;62(1):137-47. doi: 10.1016/j.molcel.2016.02.031. Epub 2016 Mar 31. PMID: 27041224; PMCID: PMC4826307.
12. Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev. 2015 May;39(3):428-41. doi: 10.1093/femsre/fuv023. Epub 2015 May 19. PMID: 25994611; PMCID: PMC5965381.
13. Hille F, Richter H, Wong SP, Bratovič M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018 Mar 8;172(6):1239-1259. doi: 10.1016/j.cell.2017.11.032. PMID: 29522745.
14. Jackson RN, Wiedenheft B. A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol Cell. 2015 Jun 4;58(5):722-8. doi: 10.1016/j.molcel.2015.05.023. Epub 2015 May 28. PMID: 26028539; PMCID: PMC4465340.
15. 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**(6), e60.
16. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., and Koonin, E. V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct,* **1**, 7.

17. Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J., and Koonin, E. V. (2011) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol,* **9**(6), 467–77.
18. Shmakov, S., Aaron, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O. O., Gootenberg, J. S., Makarova, K. S., Wolf, Y. I., Severinov, K., Zhang, F., and Koonin, E. V. (2017) Diversity and evolution of class 2 CRISPRCas systems. *Nature Reviews Microbiology,*.
19. Shah, S. A., Alkhnbashi, O. S., Behler, J., Han, W., She, Q., Hess, W. R., Garrett, R. A., and Backofen, R. (2018) Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. *RNA Biology,* **0**(0), 1–13 PMID: 29911924.
20. Shmakov, S., Makarova, K. S., Wolf, Y. I., Severinov, K., and Koonin, E. V. (2018) Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis.. *Proceedings of the National Academy of Sciences of the United States of America,*.
21. Cass, S. D. B., Haas, K. A., Stoll, B., Alkhnbashi, O., Sharma, K., Urlaub, H., Backofen, R., Marchfelder, A., and Bolt, E. L. (2015) The role of Cas8 in type I CRISPR interference. *Biosci Rep,* **35**(4), e00197.
22. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008 Aug 15;321(5891):960-4. doi: 10.1126/science.1159689. PMID: 18703739; PMCID: PMC5898235.
23. Plagens A, Richter H, Charpentier E, Randau L. DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. FEMS Microbiol Rev. 2015 May;39(3):442-63. doi: 10.1093/femsre/fuv019. Epub 2015 Apr 30. PMID: 25934119; PMCID: PMC5965380.
24. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. 2007;8(4):R61. doi:10.1186/gb-2007-8-4-r61

25. Horvath P, Romero DA, Coûté-Monvoisin AC, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol. 2008;190(4):1401-1412. doi:10.1128/JB.01415-07
26. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167-170. doi:10.1126/science.1179555
27. Reimann, V., Alkhnbashi, O. S., Saunders, S. J., Scholz, I., Hein, S., Backofen, R., and Hess, W. R. (2016) Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs. ,.
28. Nickel, L., Ulbricht, A., Alkhnbashi, O. S., Frstner, K. U., Cassidy, L., Weidenbach, K., Backofen, R., and Schmitz, R. A. (2018) Cross-cleavage activity of Cas6b in crRNA processing of two different CRISPR-Cas systems in Methanosarcina mazei G1. *RNA Biology,* **0**(0), 1–12 PMID: 30153081.
29. Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S., and Backofen, R. (January, 2013) CRISPRmap: An Automated Classification of Repeat Conservation in Prokaryotic Adaptive Immune Systems. *Nucleic Acids Research,* **41**(17), 8034–8044.
30. Alkhnbashi, O. S., Costa, F., Shah, S. A., Garrett, R. A., Saunders, S. J., and Backofen, R. (September, 2014) CRISPRstrand: Predicting Repeat Orientations to Determine the crRNA-Encoding Strand at CRISPR Loci. *Bioinformatics (Oxford, England),* **30**(17), i489–496.
31. Alkhnbashi, O. S., Meier, T., Mitrofanov, A., Backofen, R., and Voss, B. (2019) CRISPR-Cas bioinformatics. *Methods,*
32. Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., and Hugenholtz, P. (June, 2007) CRISPR Recognition Tool (CRT): A Tool for Automatic Detection of Clustered Regularly Interspaced Palindromic Repeats. *BMC Bioinformatics,* **8**, 209.
33. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., N?ron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., and Pourcel, C. (07, 2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.,* **46**(W1), W246–W251.
34. Edgar, R.C. (2007) PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8.

35. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C., and Brown, C. M. (2016) CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics,* pp. i356–i356.
36. Abouelhoda M., Kurtz S., Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. J. Discrete Algorithms.2004;2:53-86.
37. Alkhnbashi, O. S., Shah, S. A., Garrett, R. A., Saunders, S. J., Costa, F., and Backofen, R. (2016) Characterizing leader sequences of CRISPR loci. *Bioinformatics,* **32**(17), i576–i585.
38. Gudbergsdottir S, et al. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers, Mol. Microbiol., 2011, vol. 79 (pg. 35-49)
39. He J, Deem MW. Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats), Phys. Rev. Lett., 2010, vol. 105 pg. 128102
40. Weinberger AD, et al. Persisting viral sequences shape microbial CRISPR-based immunity, PLoS Comput. Biol., 2012, vol. 8 pg. e1002475
41. EM, et al. Recognition and maturation of effector RNAs in a CRISPR interference pathway, Nat. Struct. Mol. Biol., 2011
42. Juranek S, et al. A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs, RNA, 2012
43. Sternberg SH, Haurwitz RE, Doudna JA. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease , RNA , 2012
44. Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (Nov, 2011) ViennaRNA Package 2.0. *Algorithms Mol Biol,* **6**, 26.
45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (Oct, 1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**(3), 403– 410.
46. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (Mar, 2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics,* **11**, 119.
47. Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research,* **39**(suppl 2), W29–W37.

48. Saeys, Y., Inza, I., and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics,* **23**(19), 2507–17.
49. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research,* **12**(Oct), 2825–2830.
50. Zhang, Q., & Ye, Y. (2017). Not all predicted CRISPR-Cas systems are equal: isolated cas genes and classes of CRISPR like elements. BMC bioinformatics, 18(1), 92.
51. Dosztányi, Z., Mészáros, B., and Simon, I. (2009) ANCHOR: web server for predicting
52. Siguier, P., Gourbeyre, E., and Chandler, M. (09, 2014) Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews,* **38**(5), 865–891.
53. Enright, A. J., van Dongen, S., and Ouzounis, C. A. (April, 2002) An Efficient Algorithm for Large-Scale Detection of Protein Families. *Nucleic Acids Research,* **30**(7), 1575–1584.
54. Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH. Benchmarking relief-based feature selection methods for bioinformatics data mining. J Biomed Inform. 2018 Sep;85:168-188. doi: 10.1016/j.jbi.2018.07.015. Epub 2018 Jul 17. PMID: 30030120; PMCID: PMC6299838.
55. Swarts DC, et al. CRISPR interference directs strand specific spacer acquisition, PLoS One, 2012, vol. 7 pg. e35888
56. Yosef I, et al. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli, Nucleic Acids Res., 2012, vol. 40 (pg. 5569-5576)

Figure 1: **General architecture of the CRISPRIdentify approach.** The pipeline consists of two major steps. In the first step (Candidate Generation), candidate repeats are identified using Vmatch. Vmatch detects all potential pairwise repeats. This implies that leading and trailing nucleotides have an uncertain assignment as these repeated nucleotides could be part of the repeat, or part of a repetitive part of the spacer. For this reason, we first determine the minimal and maximal forms of the repeat by aligning the repeat candidates, and then consider all possible repeats between the minimal and maximal repeat variants. Mapping these repeat candidates back to the genome generates different array candidates, which are then evaluated using machine learning. For that purpose, in the second step (Candidate Evaluation) we extract features from the array such as repeat similarity, the evenness of spacer lengths, AT-content, minimum free energy of the repeat hairpin and others. We then use different machine learning (ML) approaches, learned on a large data sets of positive and negative examples, to classify them as CRISPR arrays. We use the scoring of the ML approaches to prioritize the different array candidates. Finally, we use different tools to annotate the array. Thus, we determine the strand orientation using CRISPRstrand and annotate the leader sequence using CRISPRleader. In addition, HMM and Prodigal to annotate *cas* genes.

Figure 2: **Example of a spurious detected array.** A) The repeat spacer structure is special in two aspects. First, the two spacers are identical (marked in yellow) with the exception of one position. Second, the third repeat has several mismatches (marked in red). B) The locus is overlapping a predicted protein on the opposite strand, which has an intrinsically disordered region predicted by several tools to overlap the array position. Furthermore, the three repeats overall exactly with three linear interacting peptides (LIPs) predicted by Anchor. In addition, the first LIP is terminated at the position where the last repeat (in opposite direction) has a mutated region. This together provides a repetitive protein sequence or an duplication event as an alternative explanation for the repetitive structure predicted as CRISPR array.

Table 1: **Generated Data Sets and their Combinations used in the Machine Learning Approach.**

Figure 3: **Performance of different Machine Learning classifiers.** We conducted the grid search of hyperparameters for each of the classifiers (See supplementary Materials Section 13). The initial experiment shows the dominance of the ensemble methods for CRISPR array classification. See the variety of Cas gene types in the Supplementary Materials Section 14

Table 2: **Performance of CRISPR detection tools on the test dataset, which was not used for training CRISPRIdentify**. As the test set is inbalanced, we report, beside the accuracy, also the balanced accuracy and the Matthews correlation coefficient

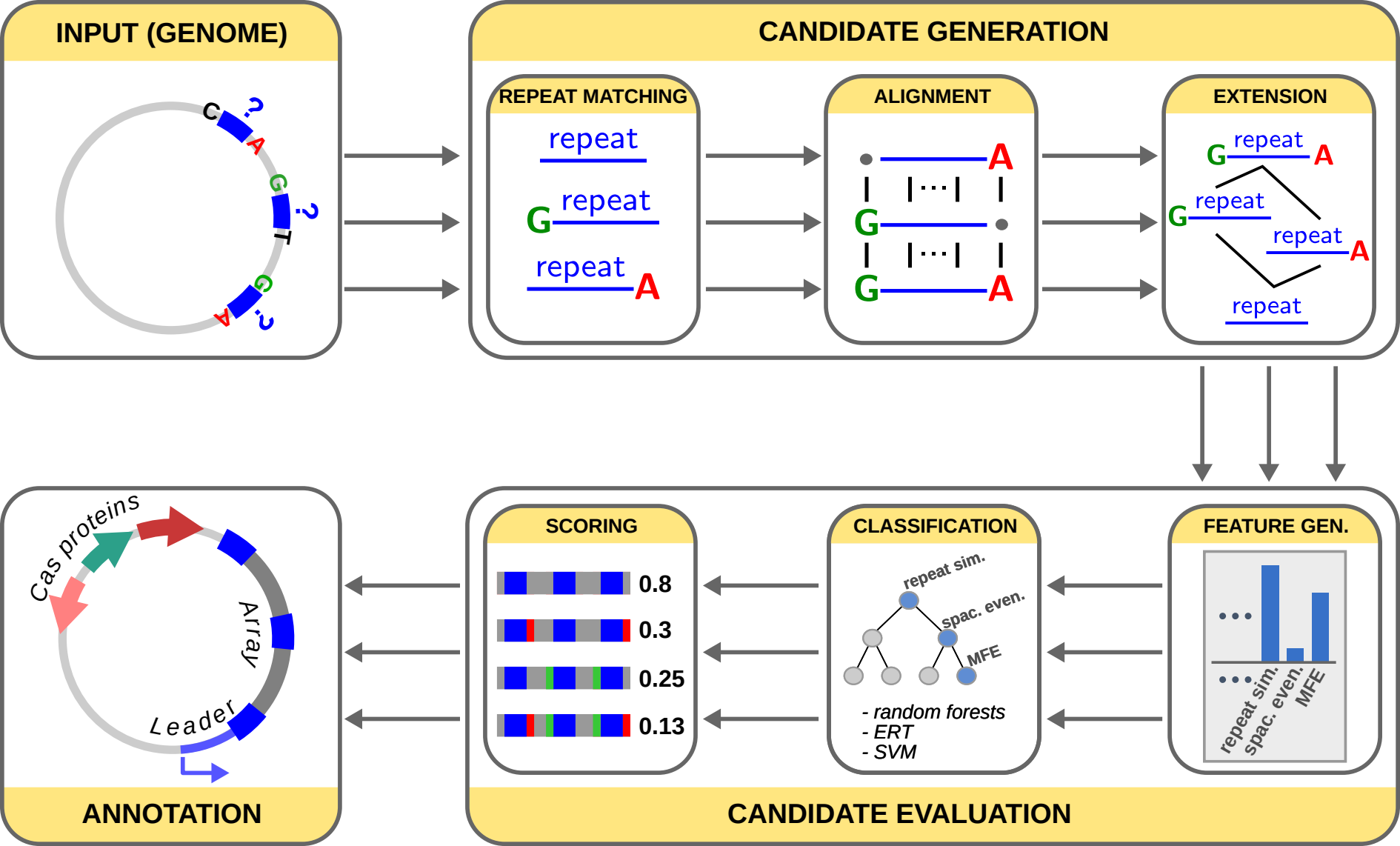
Figure 4: **Comparison of false positive predictions on the test dataset with 200 false arrays.** In summary, all false arrays were reported as a true array by at least one existing tool. In contrast, CRISPRIdentify predicts all 200 as true negatives. Even further, the majority (159 out of 200) of the arrays were reported as a true array by only one approach, validating the quality of the false data set all these 159 arrays where reported as false arrays by two existing tools. For more details about the datasets see corresponding statistics in the Supplementary Table S4 Section 10. Corresponding distribution plots are presented in Supplementary Section 19.

Figure 5: **CRISPRIdentify recall level in comparison with other tools on the annotated arrays dataset.** This dataset was compiled out of 987 archaeal and 27028 bacterial genomes obtained from NCBI. Only 89, 7% of the annotated arrays, are predicted by all tools. 147 arrays are only predicted by CRISPRIdentify, which is nearly three times as many as the 55 arrays predicted only by CRISPRCasFinder. Concerning the largest overlap between tools outside the commonly predicted 35310 arrays, 9.0% or 808 + 1095 + 3 arrays are predicted by both CRISPRCasFinder and CRISPRIdentify.

Figure 6: Feature importance distribution of 8 most influential features which was computed with ReBATE algorithm. Feature importance distribution on the train dataset described in table 1 and sampled cases from the comprehensive dataset classified with our approach. The sampled results contained 10,000 classified CRISPR array candidates.

Figure 7: **Distribution of repeat lengths in archaea and bacteria for positive and negative arrays**. CRISPR arrays were obtained by CRISPRIdentify on the **annotated arrays** dataset (see Table 1. *Bona fide Arrays* were defined as candidates that obtained certainty score (probability of being a CRISPR array) higher or equal to 0.75, *Low score Arrays* obtained certainty scores lower than 0.4). See separate distributions for archaea and bacteria in Supplementary Section 18 Distributions of arrays with different scores.

Table 3: **Spurious Arrays in the CRISPRCasFinder datasets [27]**. To detect arrays that are likely false positive candidates, we first filtered arrays that are not covered by the parameter setting of the different tools (i.e., <= 3 repeats, or repeat length < 23 or > 55). The remaining number of arrays is listed in the row *standard arrays*. Arrays that were labeled as *likely negative* (see Methods) if they had one of the following four defects: 1) damaged repeats, 2) high similarity between spacers, 3) uneven distribution of spacer lengths, and 4) high similarity between repeats and spacers. The following rows list how many of these likely negative arrays had specific deficiencies (see text for details). Note that some arrays can have several problems, hence there are multiple entries in these rows. The FDR is calculated considering all likely negatives as true negatives, as we do not have other information. This FDR values are, however, likely overestimates. Nevertheless, they show the dimension of the problem.

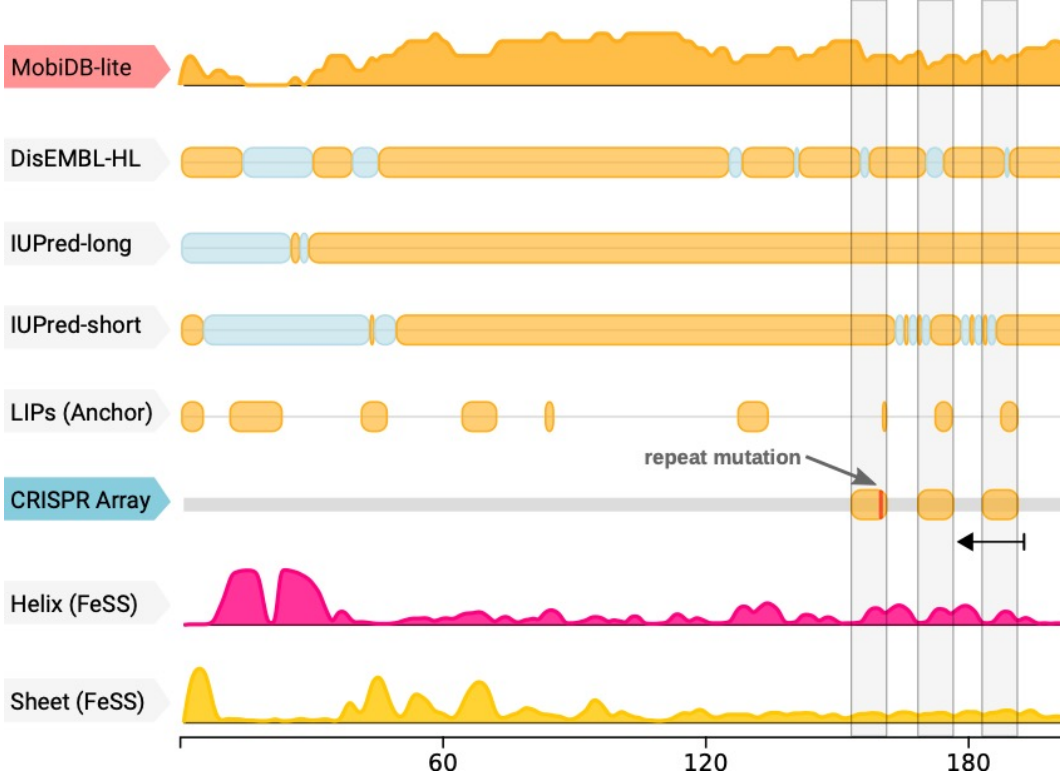

INPUT (GENOME)
CANDIDATE GENERATION
REPEAT MATCHING
repeat
G
repeat
repeat
A
ALIGNMENT
A
G
G
A
EXTENSION
G
repeat
A
G
repeat
repeat
A
repeat
FEATURE GEN.
repeat sim.
spac. even.
MFE
CLASSIFICATION
repeat sim.
spac. even.
MFE
- random forests
- ERT
- SVM
SCORING
0.8
0.3
0.25
0.13
CANDIDATE EVALUATION
Cas proteins
Array
Leader
ANNOTATION

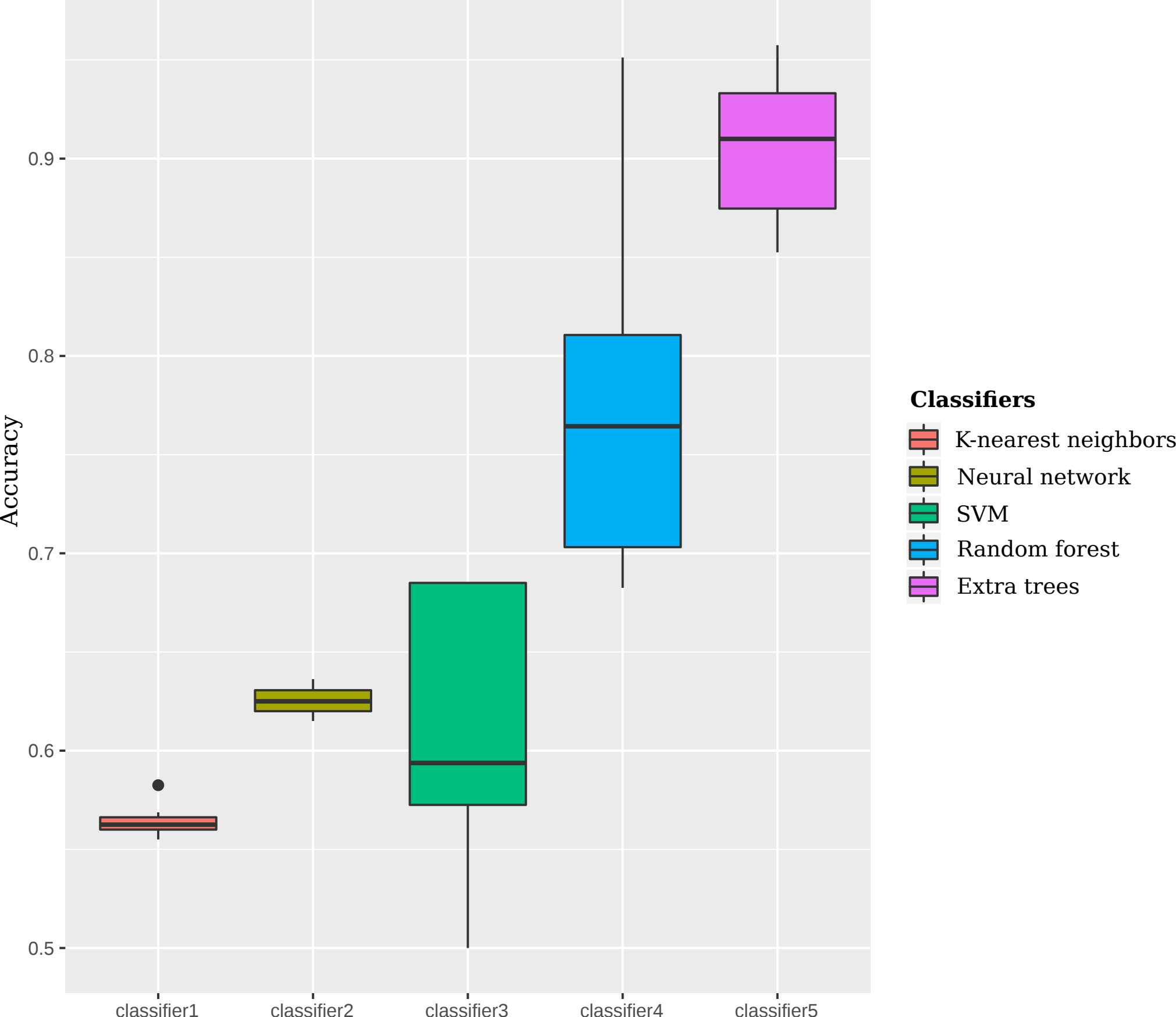
A)

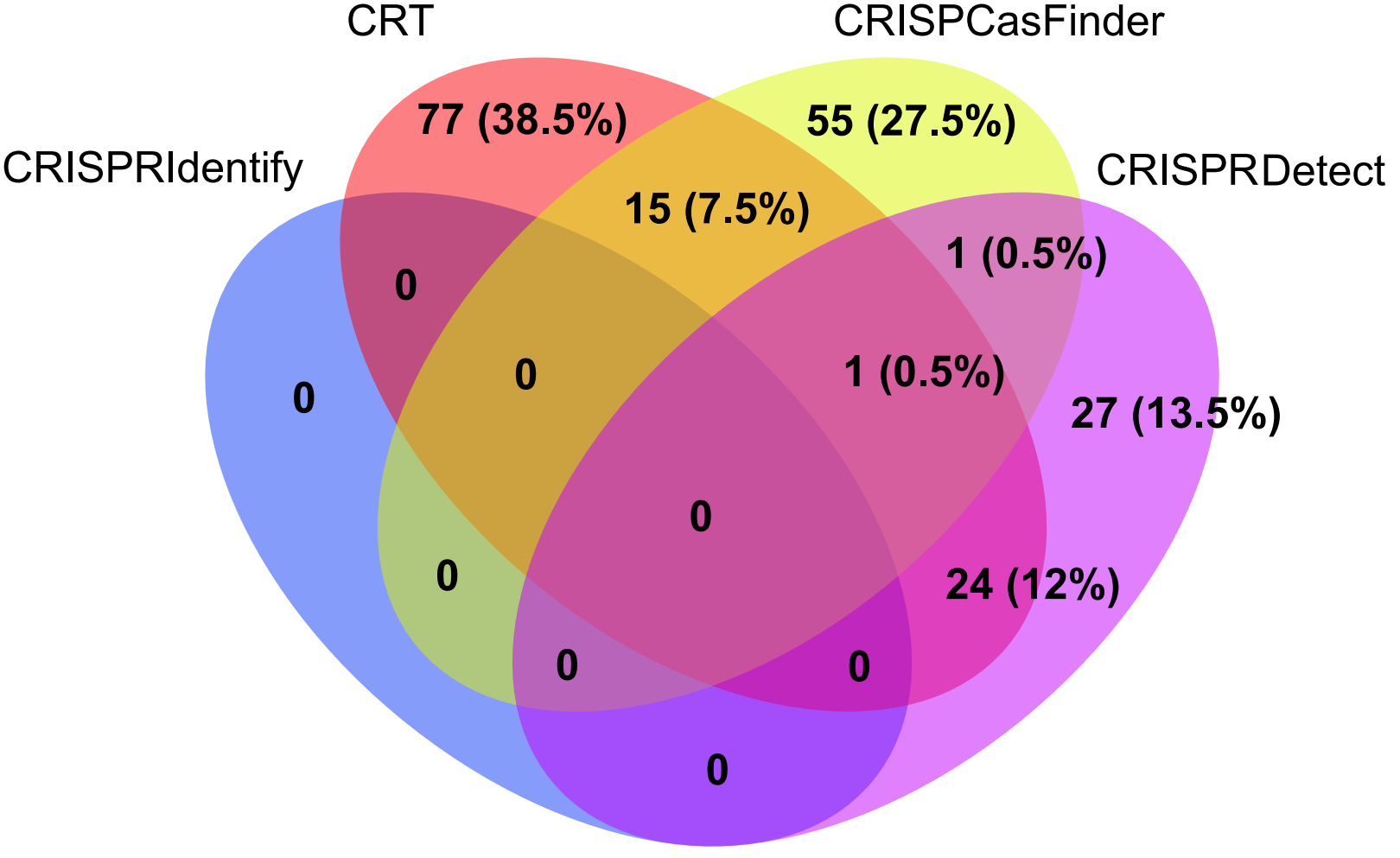
CRISPR 1 Range: 9 - 122

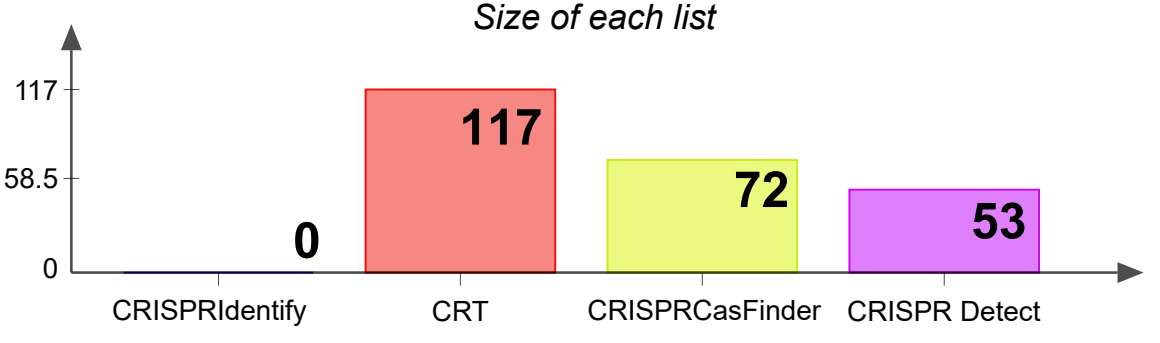
POSITION	REPEAT	SPACER
9	AAATTCTTGTATCGGTGAAACGGG	CATTTGCTTTTCCACTTTAGG
54	AAATTCTTGTATCGGTGAAACGGG	CATTTGCTTTTCTACTTTAGG
99	AAACTCCTGTATCGGTGAAACGGG	
Repeats: 3	Average Length: 24	Average Length: 21

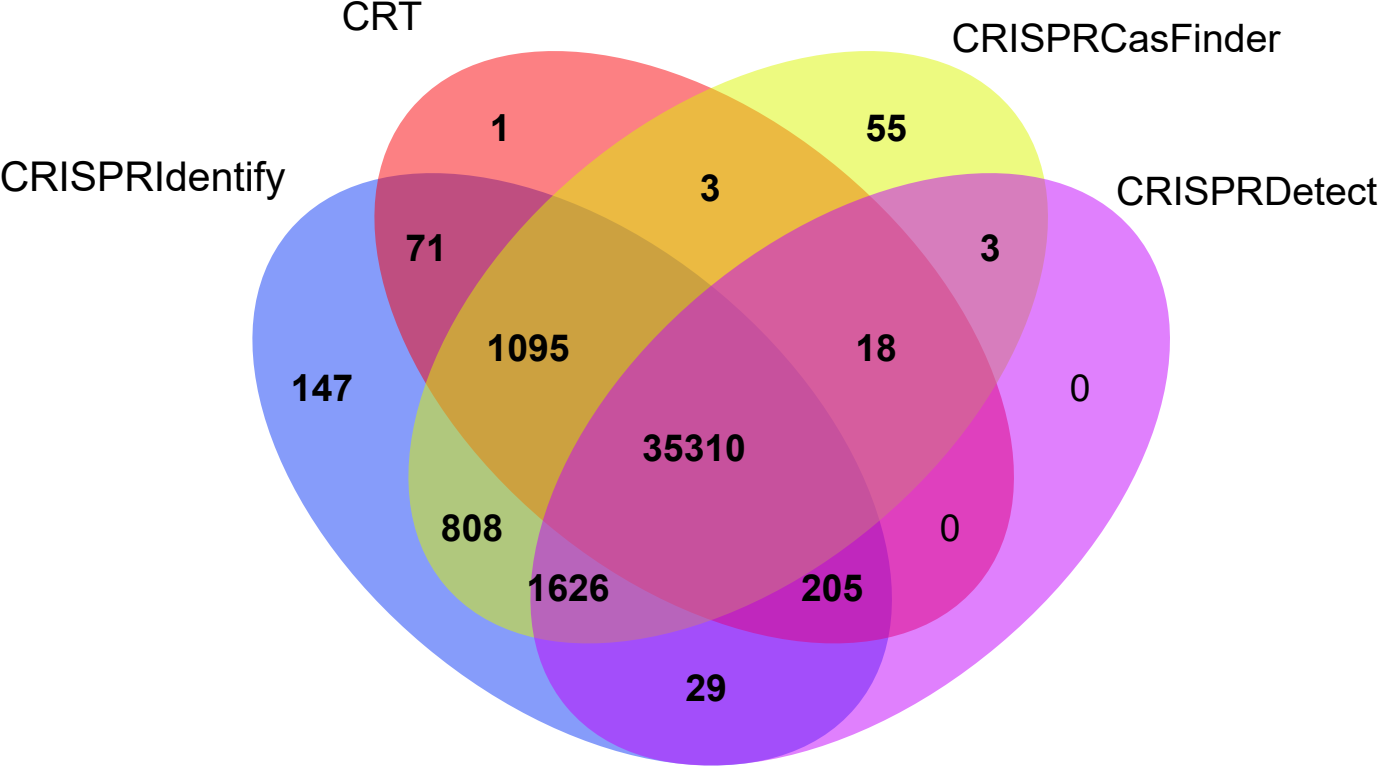
B)

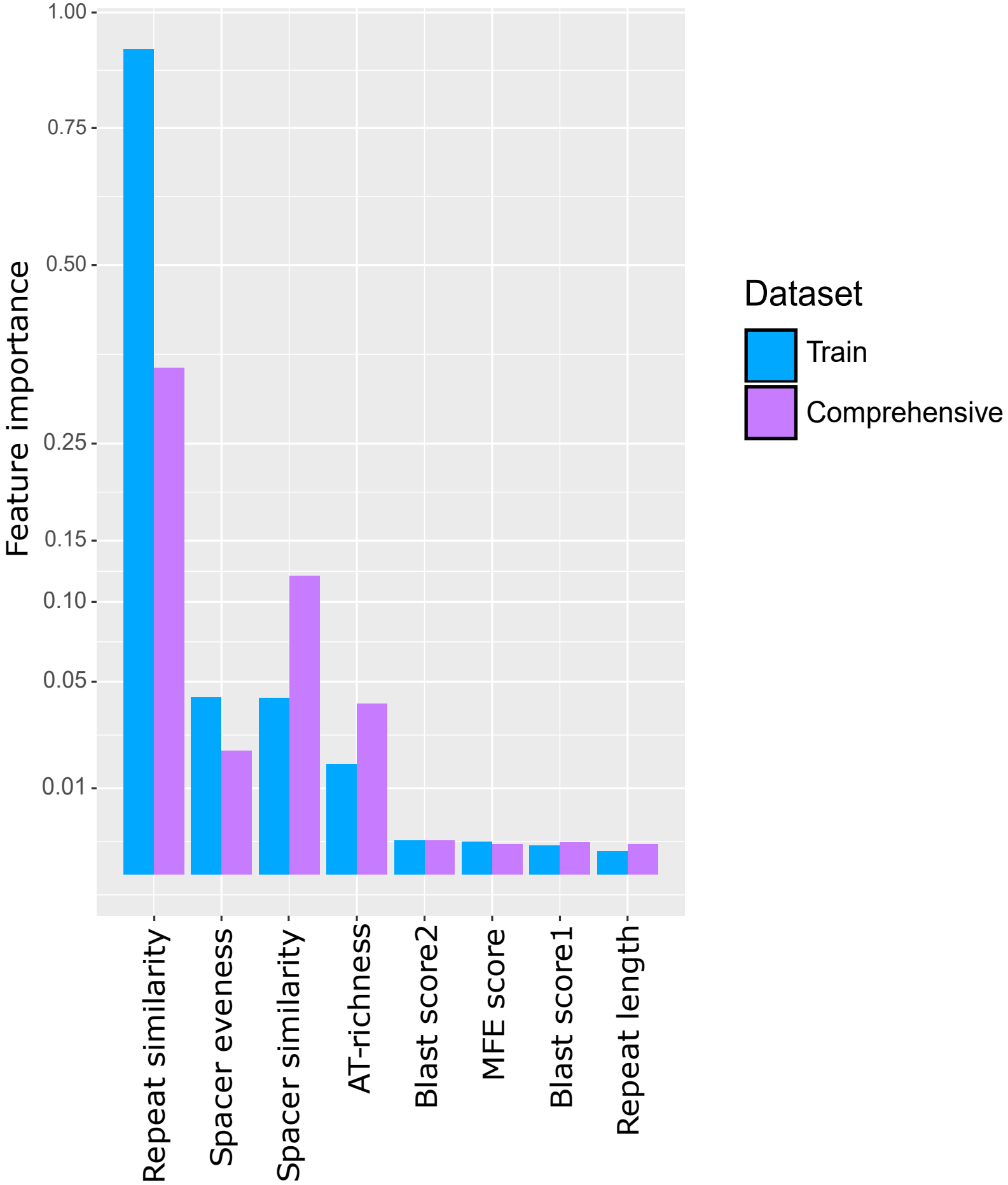


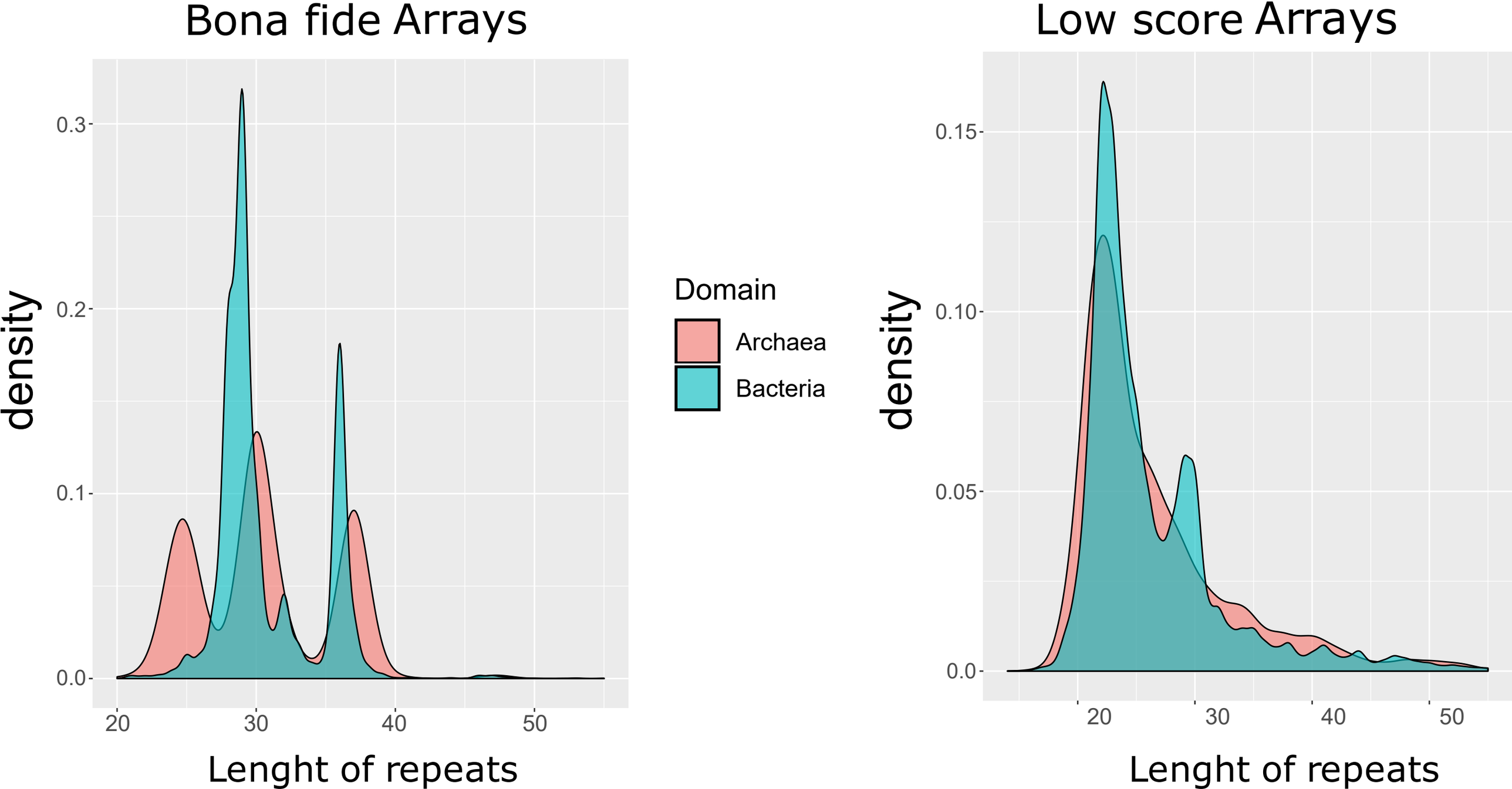

Accuracy
0.9
0.8
0.7
0.6
0.5
classifier1
classifier2
classifier3
classifier4
classifier5
Classifiers
K-nearest neighbors
Neural network
SVM
Random forest
Extra trees


CRT
CRISPCasFinder
CRISPRIdentify
CRISPRDetect
77 (38.5%)
55 (27.5%)
15 (7.5%)
1 (0.5%)
0
0
0
1 (0.5%)
27 (13.5%)
0
0
24 (12%)
0
0
0


Size of each list
117
58.5
0
0
117
72
53
CRISPRIdentify
CRT
CRISPRCasFinder
CRISPR Detect


CRT
CRISPRCasFinder
CRISPRIdentify
CRISPRDetect
1
55
3
71
3
1095
18
147
0
35310
808
0
1626
205
29


Feature importance
1.00
0.75
0.50
0.25
0.15
0.10
0.05
0.01
Repeat similarity
Spacer eveness
Spacer similarity
AT-richness
Blast score2
MFE score
Blast score1
Repeat length
Dataset
Train
Comprehensive


Bona fide Arrays
Low score Arrays
density
0.0
0.1
0.2
0.3
0.05
0.10
0.15
20
30
40
50
Lenght of repeats
Domain
Archaea
Bacteria

Name	Number comb.	Name	Number comb.	Name	Size
train & validate	2100	archaeal	800	archaeal arrays (positive)	400
				archaeal shuffled arrays (negative)	400
		bacterial	1200	bacterial arrays (positive)	600
				bacterial shuffled arrays (negative)	600
		false arrays, train (negative)			100
test	750	Cas-associated (positive)			550
		false arrays, test (negative)			200
CRISPRCasFinder					3263
annotated arrays (positive)					39371

	TPR	TNR	ACC	BACC	MCC
CRISPRIdentify	0.99	1	0.99	0.99	0.99
CRT	0.93	0.42	0.79	0.67	0.42
CRISPRCasFinder	0.96	0.64	0.87	0.80	0.67
CRISPRDetect	0.96	0.74	0.89	0.85	0.73

		CRT	Finder	Detect
arrays declared positive		1866	3263	764
standard arrays		1358	1395	346
likely negative arrays		602	628	110
of which are arrays	with damaged repeats	165	189	40
	with similar spacers	57	0	13
	with uneven spacers lengths	473	508	83
	with sim. repeats to spacers	23	4	11
FDR		44,3%	45,0%	31,8%
likely positive arrays		756	767	236