W246–W251 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 22 May 2018*
doi: 10.1093/nar/gky425

CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins

David Couvin[1], Aude Bernheim[2,3], Claire Toffano-Nioche[1], Marie Touchon[2,3], Juraj Michalik[4], Bertrand Néron[5], Eduardo P. C. Rocha[2,3], Gilles Vergnaud[1], Daniel Gautheret[1] and Christine Pourcel[1,*]

[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France, [2]Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue du Docteur Roux, 75015, Paris, France, [3]CNRS, UMR3525, 25-28 rue du Docteur Roux, 75015, Paris, France, [4]Université Lille 1, CRIStAL, équipe Bonsai, Cité Scientifique Bat M3, 59655 Villeneuve d'Ascq Cedex, France and [5]Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, Institut Pasteur, 25–28 rue du Docteur Roux, 75015, France

Received January 31, 2018; Revised May 02, 2018; Editorial Decision May 03, 2018; Accepted May 09, 2018

ABSTRACT

CRISPR (clustered regularly interspaced short palindromic repeats) arrays and their associated (Cas) proteins confer bacteria and archaea adaptive immunity against exogenous mobile genetic elements, such as phages or plasmids. CRISPRCasFinder allows the identification of both CRISPR arrays and Cas proteins. The program includes: (i) an improved CRISPR array detection tool facilitating expert validation based on a rating system, (ii) prediction of CRISPR orientation and (iii) a Cas protein detection and typing tool updated to match the latest classification scheme of these systems. CRISPRCasFinder can either be used online or as a standalone tool compatible with Linux operating system. All third-party software packages employed by the program are freely available. CRISPRCasFinder is available at https://crisprcas.i2bc.paris-saclay.fr.

INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) and associated proteins (Cas) form the CRISPR-Cas systems. CRISPRs consist of a succession of 24–50 bp long direct repeats or 'repeats' separated by similarly sized unique sequences called spacers. They are transcribed from a promoter present in the leader (often a 100–200 bp AT-rich sequence) and therefore CRISPR arrays are functionally oriented (1). A community effort resulted in the classification of CRISPR-Cas systems into two classes, six types and 22 subtypes, according to their Cas proteins (2,3). Since the development of genome editing technologies based on elements of the CRISPR-Cas systems, these genomic entities have attracted a lot of attention. Indeed, components of these biological systems present in about 80% archaea and half of bacteria can be used in multiple applications in genetic engineering (4,5). The rapid rate of evolution of certain CRISPR arrays also allows their effective use in typing bacterial isolates (6). Several programs have been developed to identify CRISPR arrays in genomic sequences, the most frequently cited being CRISPRFinder (7), CRT (8) and PILER-CR (9). Additional programs such as CRISPRDetect (10), CRISPRdigger (11) and CRF (12) are also available. Three programs have been proposed for CRISPR array strand prediction based on the characteristics of the CRISPR repeat and the leader: CRISPRDirection using CRISPRDetect (10,13), CRISPRstrand (14,15) and CRISPRleader (16). Cas proteins and systems can be identified using the program Macromolecular System Finder (MacSyFinder), which has a dedicated module (CasFinder) (17). The program is based on the search of protein similarity using Hidden Markov Models (HMM) and a model of genetic composition and organization of the identified components. HMMCAS is a web tool that can be queried online to identify Cas proteins (18). The search for CRISPRs and Cas in user-submitted data can be done on the web using CRISPRone (19) or locally using the CRISPRdisco pipeline (20).

Here, we present CRISPRCasFinder, which is an updated, improved, and integrated version of CRISPRFinder and CasFinder with freely available third-party software dependencies. CRISPRCasFinder now includes a standalone version, and presents enhanced CRISPR detection performance.

*To whom correspondence should be addressed. Tel: +33 1 69 15 30 01; Fax: +33 1 69 15 72 96; Email: christine.pourcel@u-psud.fr

RESULTS

Availability and implementation

The CRISPRCasFinder web server is based on independent front and back ends. The front end was implemented as a user-friendly web application using .NET Core (dotnet core) development platform and C# (C sharp) programming language. The Bootstrap framework was used to design the web application. CRISPRCasFinder is also available as a standalone program compatible with Linux (including Windows Subsystem for Linux) and MacOS systems. The program was written in Perl. A workflow is shown on Figure 1 and Supplementary Figure S1, and details on dependencies are provided in Supplementary Material.

Input

The web server currently accepts (multi-)Fasta DNA sequence files of size up to 50 Mb including up to 100 sequences. The standalone application has no pre-defined input size limit and is only limited by the available computer memory.

Output

The web server produces a summary table with an overview of the results (Figure 2A) and the possibility to visualize each array separately (Figure 2B). CRISPR arrays and Cas protein analyses are returned as .xls, GFF3, JSON, TSV and Fasta formatted files. The standalone program returns the same files as well as optional files (see Supplementary Material for further details).

Improved detection of CRISPR arrays and evidence level rating

To identify CRISPR arrays CRISPRCasFinder uses CRISPRFinder v4.2 which is itself based on Vmatch version 2.3 (21) (http://www.vmatch.de/) to identify the CRISPR repeats. CRISPRFinder v4.2 has evolved from the first version described by Grissa *et al*. (7) and the differences are listed in Supplementary Material. In order to help the user to discriminate spurious CRISPR-like elements from true CRISPRs, we included a rating system based on several criteria. Short candidate arrays made of one to three spacers often do not correspond to CRISPRs (22) and are therefore given the lowest evidence level (rated **1**). Evidence levels **2**–**4** are attributed based on combined degrees of similarity of repeats and spacers. In the majority of cases, repeats are very well conserved and can be defined as a stretch of sequence with a 100% similarity inside the CRISPR array when excluding the distal truncated/diverged repeat. Arrays showing repeats heterogeneity often correspond to coding sequences with a repeated element and are rarely real CRISPRs (23). In contrast, spacers are not expected to show a significant degree of similarity, except in the case of rare recombination or duplication events (24). Therefore, the degree of similarity between spacers is expected to be very low in *bona fide* CRISPRs. We thus implemented an algorithm to measure CRISPR repeat conservation based on Shannon's entropy (Supplementary Material, Table S1, Figures S2–4) and produce an EBcons (entropy-based conservation) index. We empirically determined EBcons thresholds based on the analysis of 128 CRISPR arrays from 128 genomes in CRISPRdb (23) (See Supplementary Material and Supplementary dataset 1). Putative CRISPR arrays with at least four spacers are assigned to levels **2**–**4** as follows: repeats EBcons < 70 (level **2**), repeats EBcons ≥ 70 and spacers overall percentage identity $> 8\%$ (level **3**); repeats EBcons ≥ 70 and spacers overall percentage identity $\leq 8\%$ (level **4**). CRISPR arrays having evidence-levels **3** and **4** may be considered as highly likely candidates, whereas evidence-levels **1** and **2** indicate potentially invalid CRISPR arrays. The ambiguous notion of 'confirmed' or 'hypothetical' CRISPR array (associated with CRISPRFinder v1.0) is no longer used in CRISPRFinder v4.2. We used a panel of 400 genomic sequences (260 bacteria and 140 archaea) from different species (Supplementary dataset 2) taken in alphabetical order, to evaluate the distribution of CRISPR arrays in the four different evidence-level groups. Out of 3251 arrays, there were respectively 1969, 63, 76 and 1143 arrays with evidence level 1, 2, 3 or 4. The identification of false-positive arrays when they possess less than four spacers is not an easy task and some of the evidence-level 1 arrays may in fact be real CRISPRs (see Supplementary Material for an example). Therefore we give the possibility either to view all the detected CRISPR arrays or to hide those with evidence-level 1. When a short CRISPR array has the same consensus repeat as an evidence-level **4** array, it can be considered as a level **4** CRISPR. Applying this rule would upgrade about 5% of the level 1–3 arrays in the test panel to level 4 (163 arrays out of 2108). This scoring correction will be automatically applied in the future when the CRISPR database is integrated in the system. In addition, evidence-level 1 arrays which are not associated to *cas* genes will be deleted (42 arrays in 26 genomes out of the 400 test genomes).

Orientation of the CRISPR array

CRISPRFinder provides two indicators of CRISPR arrays orientation. First, orientation was predicted by CRISPRDirection for a curated dataset of consensus CRISPR repeats and this result is shown for CRISPR arrays with a matching repeat. We provide an additional method that does not require the existence of previously oriented homologous systems and is based on the AT% in the 100 bp region flanking the array on both sides. As the 5′ region of an oriented array is often AT-rich (25), the flanking side showing the higher AT% is used as a second indicator of orientation. An option in the standalone program allows users to determine the length of the flanking region to be analyzed. The result of both tests is sometimes different as illustrated on Figure 2 with the genome of *Bacillus halodurans*, showing that additional developments are still necessary to orientate CRISPR arrays with accuracy. The search for an AT-rich region flanking the CRISPR array has been used by different authors to orientate the CRISPR array (e.g. with CRISPRmap or CRISPRDirection) but it is not relevant for all genomes, particularly those which are globally AT-rich. In addition Alkhnbashi *et al*. (16) showed that 13% of

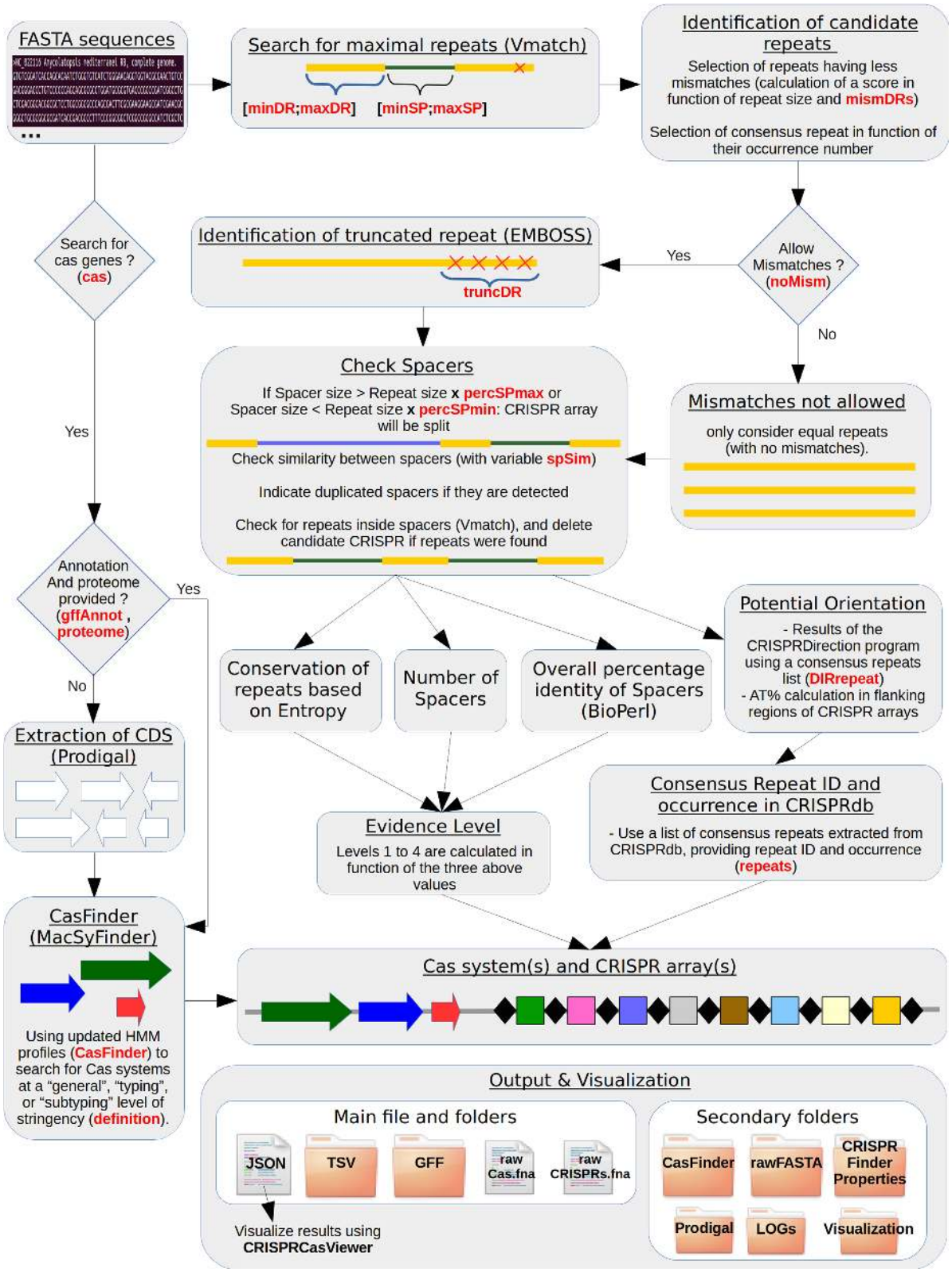

Figure 1. CRISPRCasFinder workflow.

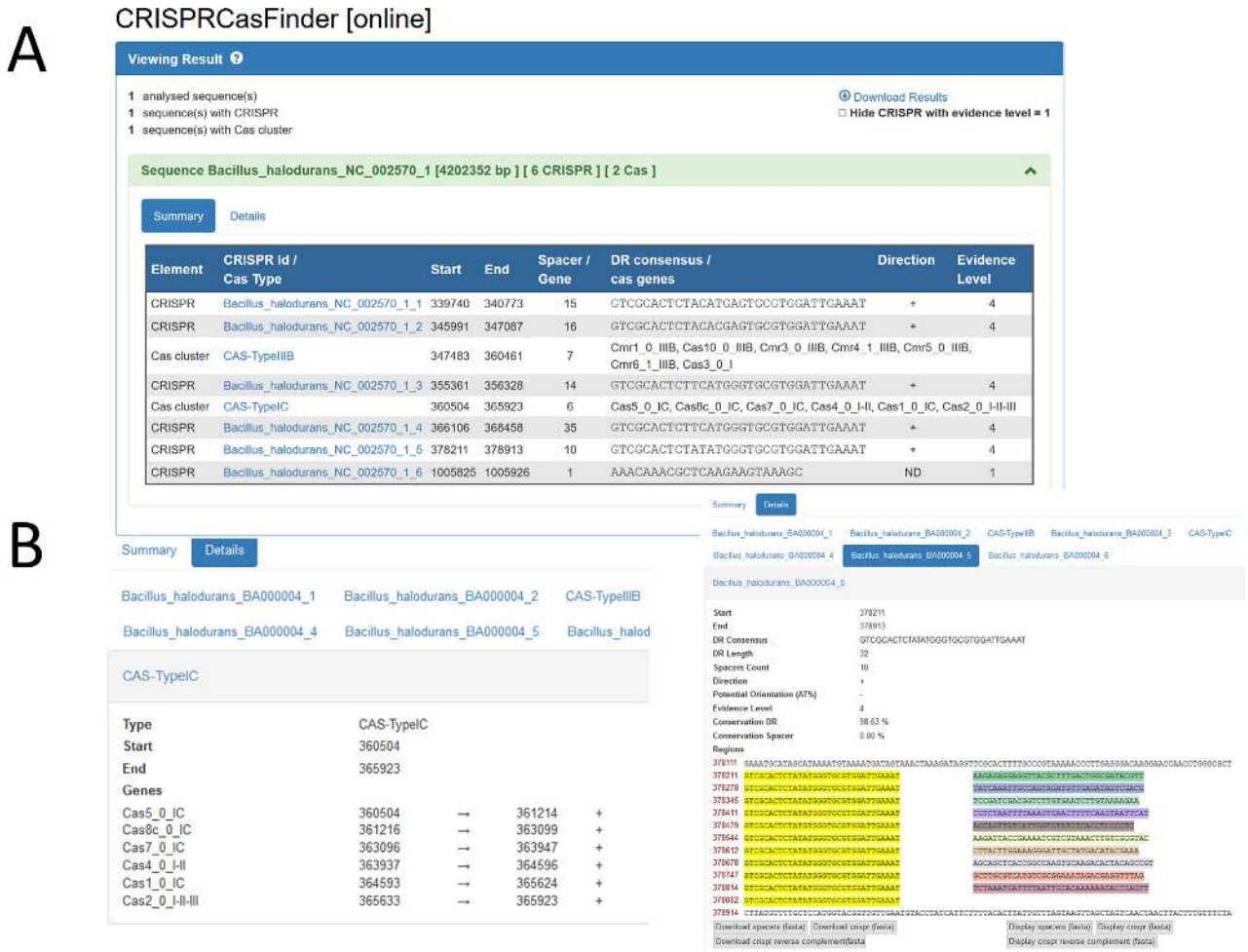

Figure 2. Output of CRISPRCasFinder. (**A**) The summary displays information on CRISPR arrays and cas gene clusters in the order in which they lie along the chromosome. "Direction" is the proposed orientation of the CRISPR cluster according to the CRISPRDirection program. (**B**) Details on individual CRISPR arrays and cas gene clusters can be viewed.

980 archaeal CRISPR loci, and 24% of 2852 bacterial loci were leaderless.

Update of the detection and typing of Cas proteins

CRISPRCasFinder identifies and types Cas systems based on known Cas protein sequences to increase sensitivity and specificity. Putative coding sequences (CDSs) are identified using Prodigal v2.6.3 (26) in the input nucleotide sequences. Translated CDSs are analyzed by CasFinder to identify the systems and their components. CasFinder is based on MacSyFinder (17), which provides a flexible framework to identify systems in genomes using a model describing their genetic architecture and the minimally sufficient number of protein components. Putative Cas proteins are searched by sequence similarity using HMM protein profiles. The assignment of a protein to a given system is decided based on its compliance with the content and organization defined in the model. Components can be defined as mandatory, accessory or forbidden. The latter (proteins that cannot belong to a defined subtype) are useful for discriminating between systems' types or sub-types. Cas systems can be identified at three levels of precision ('General', 'Typing' and 'SubTyping') using increasingly specific descriptions of the systems. This allows dealing with the trade-off between the detection of a maximum of Cas clusters by using the more sensitive General model, and a precise identification of each cluster, using the more stringent sub-type model. The General Cas model was designed to detect any cluster of at least three Cas proteins (class 1 systems), or containing at least Cas9, Csn2, Cas12 or Cas13 for class 2 systems. The first version of CasFinder contained a set of models to identify and classify CRISPR-Cas systems in three types and thirteen subtypes (17). The last update (CasFinder version 2.0) fits with the new community-based classification of CRISPR-Cas systems (2,3). The program distinguishes class 1 from class 2 systems, detects three new types and nine new subtypes of CRISPR-Cas systems (a description of the sub-typing models is provided in Supplementary Figure S5). Moreover, as 394 new profiles were published with the new classification, previously existing models and protein profiles were revised to improve speed, limit the redundancy between protein profiles, and enhance subtyping accuracy (Supplementary Figure S6). As a result, the new ver-

sion contains 43 additional protein profiles, for a total of 120 profiles (Supplementary dataset 3).

Evaluation of CRISPR and Cas detection by CRISPRCasFinder

To test the accuracy of the novel CRISPR detection method we used the test panel of 400 genomic sequences. Predictions were compared to the expert-curated annotations in CRISPRdb (used as a reference-set) and to three CRISPR detection programs PILER-CR, CRT and CRISPRDetect. Only CRISPR arrays having at least four spacers were taken into account in the evaluation. In CRISPRdb, 1263 arrays were displayed for the above mentioned set of 400 genomes after manual curation (Supplementary dataset 4). Precision, recall and F-measure metrics were used to compare the three programs with the reference-set (Supplementary Tables S2 and 3). These metrics showed that the four programs performed similarly, with precision from 0.921 to 0.982, recall from 0.935 to 0.987, and F-measure from 0.932 to 0.9776. Detailed validation procedures are provided in Supplementary Material. Predictions of CRISPRCasFinder are visible in Supplementary dataset 5.

A comparison between the detection of CasFinder v.2.0 and the summary presented with the new classification in (2) revealed few differences (Jacquard coefficients between 77 and 96%, Supplementary Figure S7). Overall, CasFinder v2.0 is more conservative, finds fewer systems, because it requires at least three genes in the Class 1 Cas system, whereas the previous study only required two genes. We opted for the conservative approach because, to the best of our knowledge, no study identified a fully functional Class 1 Cas system capable of adaptation and interference with only two genes. Expert users can change the underlying models to identify Cas systems in the standalone version and lower the threshold. However the use of the 'General' model already allows the identification of relevant small clusters such as in *Melioribacter roseus* (NC_018178) which possesses an evidence-level 4 CRISPR with a 46-bp repeat located immediately adjacent to two class 2 *cas* genes (cas2_TypeI-II-III, cas1_TypeII).

We evaluated the performance of CRISPRCasFinder, CRISPRDetect and CRISPRone online using a set of 30 genomes (Supplementary dataset 6) selected because they possess particular sets of CRISPRs observed while curating CRISPRdb (Supplementary Table S5). As compared to CRISPRCasFinder and CRISPRone, CRISPRDetect proposes options to edit the CRISPR array and provides a directional analysis based on seven characteristics of the leader and the CRISPR repeat (10), however it relies on NCBI annotations to identify Cas protein and therefore often fails to produce a result. CRISPRone performs an HMM search to identify Cas proteins but the method is less stringent than CasFinder v2.0, and Cas-like proteins are frequently displayed (see Supplementary Material for selected examples). Online the duration of an analysis was variable with the three programs presumably depending on the capacity and the workload of the associated servers. Single bacterial genomes of 4–6 Mbases (Mb) were analyzed by CRISPRCasFinder in 1–2 min whereas a 50Mo file (the current limit on the web server) of 10 fasta files made by concatenating a 5-Mb genome (containing two Cas and seven CRISPRs loci) ran in 5 min. The same 50Mo file split into 100 fasta files ran at the same speed. Runtimes and memory usage were calculated with the standalone versions of CRISPRCasFinder and CRISPRDetect using four genomes of 0.5, 5, 10 and 53 Mb. The results showed that runtimes were similar between the two programs but CRISPRCasFinder tends to require more memory (Supplementary Table S6). At last, we believe that the output of CRISPRCasFinder in the form of a clear and compact summary is an advantage over the other programs.

Use case studies

The simultaneous search for CRISPR and Cas by CRISPRCasFinder greatly facilitates the evaluation of tentative CRISPR and *cas* loci and this is further exemplified in several cases. Some genes possess tandem repeats which can be misidentified as CRISPRs. For example, analysis of the *Pantoea ananatis* LMG20103 genome (NC_013956) reveals the existence of a putative CRISPR array with a 23-bp repeat and 26 spacers (Supplementary Figure S8). The evidence level of this array is 2, with repeats and spacers conservations of 57 and 12%, respectively, and no Cas protein detected in the genome. In fact this sequence is part of an Ice nucleation protein gene. An opposite situation is that of *Streptococcus sanguinis* SK36 (CP000387) which displays a cluster of Type III-A *cas* genes intermixed with two evidence-level 2 CRISPR arrays showing highly dissimilar repeat sequences (Supplementary Figure S9). In both cases, CRISPRone (19) and CRISPRCasFinder were in agreement. Another interesting feature of CRISPRCasFinder is the possibility to compare the repeat of short arrays with evidence-level 1 to that of larger arrays present in the same genome, allowing to confirm the small size loci as valid CRISPRs such as in the genome of *Methanosarcina thermophila* MT-1 (AP017646) (Supplementary Figure S10).

DISCUSSION AND CONCLUSION

The updated CRISPRCasFinder shows enhanced performance and capabilities to identify both CRISPR arrays and Cas proteins, improving the previously existing separate tools CRISPRFinder and CasFinder. In addition CRISPRCasFinder and CRISPRCasViewer (Supplementary Figure S11) are available as standalone programs for users willing to analyze large volumes of sequences (see Supplementary Material for details). We are developing a dedicated tool for the analysis of large metagenomic datasets, allowing a simpler and faster CRISPR array and Cas protein detection. CRISPRCasFinder will continue to evolve, notably by providing a better prediction of CRISPR array orientation using curated data from the currently developed database. Key criteria for array orientation are the presence of a leader/promoter sequence immediately before the first repeat, the existence of a diverged/truncated repeat at the 3′ end, the nature of the repeat sequence and its secondary structure, and the position of the *cas* genes cluster. The program will also be updated to match novel typing methods for Cas systems if and when sufficient examples become available.

At last, using extra information can improve the ability to distinguish small CRISPRs from false positives, including the existence of a similar repeat in a larger CRISPR array, the presence of *cas* genes, generally situated upstream the CRISPR array (27) or of a characteristic leader sequence. The next version of CRISPRFinder will incorporate these elements for improved CRISPR classification.

CRISPRCasFinder will be part of a new integrated CRISPR-Cas analysis system, eventually replacing CRISPRdb and associated tools (CRISPRtionary, CRISPRcompar, MyCRISPRdb), which were not designed as actual web services.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Stefan Kurtz for making Vmatch an open source project without licensing requirements. We thank Christine Drevet, Maud Pupin, Luis M. Rodriguez and Lee S. Katz for their help in the previous developments of CRISPRFinder. We are particularly grateful to the SICS team (I2BC), and especially to Cyrille Petat and Pierre-Albert Charbit for the development of the web application. We also acknowledge the constant help and support by Arnaud Martel and Anne-Pascale Jaudier. We thank Mélina Gallopin, David Christiany, Nicolas Villeriot, Nicolas Maillet, Fabrice Leclerc and Emilie Drouineau for helpful comments and for testing the standalone software.

FUNDING

Institut Français de Bioinformatique (IFB) [ANR–11–INSB–0013, in part]; European Defense Agency Research and Technology project JIP-ICET2 A-1341-RT-GP, called Bioforensics for Biodefence (B2-forensics). Funding for open access charge: ANR Grant.
Conflict of interest statement. None declared.

REFERENCES

1. Tang,T.H., Bachellerie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Huttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 7536–7541.
2. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S., Costa,F., Shah,S.A., Saunders,S.J., Barrangou,R., Brouns,S.J., Charpentier,E., Haft,D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
3. Shmakov,S., Abudayyeh,O.O., Makarova,K.S., Wolf,Y.I., Gootenberg,J.S., Semenova,E., Minakhin,L., Joung,J., Konermann,S., Severinov,K. *et al.* (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell*, **60**, 385–397.
4. Barrangou,R. and Doudna,J.A. (2016) Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*, **34**, 933–941.
5. Hille,F. and Charpentier,E. (2016) CRISPR-Cas: biology, mechanisms and relevance. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **371**, 20150496.
6. Grissa,I., Vergnaud,G. and Pourcel,C. (2009) Clustered regularly interspaced short palindromic repeats (CRISPRs) for the genotyping of bacterial pathogens. *Methods Mol. Biol.*, **551**, 105–116.
7. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
8. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
9. Edgar,R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
10. Biswas,A., Staals,R.H., Morales,S.E., Fineran,P.C. and Brown,C.M. (2016) CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics*, **17**, 356.
11. Ge,R., Mai,G., Wang,P., Zhou,M., Luo,Y., Cai,Y. and Zhou,F. (2016) CRISPRdigger: detecting CRISPRs with better direct repeat annotations. *Sci. Rep.*, **6**, 32942.
12. Wang,K. and Liang,C. (2017) CRF: detection of CRISPR arrays using random forest. *PeerJ*, **5**, e3219.
13. Biswas,A., Fineran,P.C. and Brown,C.M. (2014) Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics*, **30**, 1805–1813.
14. Lange,S.J., Alkhnbashi,O.S., Rose,D., Will,S. and Backofen,R. (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044.
15. Alkhnbashi,O.S., Costa,F., Shah,S.A., Garrett,R.A., Saunders,S.J. and Backofen,R. (2014) CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, **30**, i489–i496.
16. 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**, i576–i585.
17. Abby,S.S., Neron,B., Menager,H., Touchon,M. and Rocha,E.P. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
18. Chai,G., Yu,M., Jiang,L., Duan,Y. and Huang,J. (2017) HMMCAS: a web tool for the identification and domain annotations of Cas proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi:10.1109/TCBB.2017.2665542.
19. Zhang,Q. and Ye,Y. (2017) Not all predicted CRISPR-Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics*, **18**, 92.
20. Crawley,A., Henriksen,J.R. and Barrangou,R. (2018) CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas systems. *CRISPR J.*, **1**, http://doi.org/10.1089/crispr.2017.0022.
21. Abouelhoda,M., Kurtz,S. and Ohlebusch,E. (2004) Replacing suffix trees with enhanced suffix arrays. *J. Discrete Algorithms*, **2**, 53–86.
22. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.
23. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
24. Kupczok,A., Landan,G. and Dagan,T. (2015) The contribution of genetic recombination to CRISPR array evolution. *Genome Biol. Evol.*, **7**, 1925–1939.
25. Jansen,R., Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
26. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
27. Koonin,E.V., Makarova,K.S. and Zhang,F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.*, **37**, 67–78.