CRISPRCasFinder/CRISPRCasViewer manual version 1.0 february 2018

Table of content

Introduction

The standalone CRISPRCasFinder version allows the flexible adjustment of options and generates additional output files as compared to the CRISPRCas++ website. CRISPRCasViewer is useful to view results produced by the standalone.

Installation guide for standalone CRISPRCasFinder

The program CRISPRCASFinder.pl has been tested on 64-bit Ubuntu versions (14.04 and 16.04), including the Windows10 Subsystem for Linux (WSL) and on Mac OS X 10.11 El Capitan.

Software Dependencies

(Mac OS X operating system: Xcode https://developer.apple.com/xcode/ programming environment (version 5.0 or upper) library must be installed)

Users must have rights on the folder containing the files. Ubuntu users must have sudo rights.

The following package managers are required for the installation:

- apt-get https://wiki.debian.org/apt-get (Ubuntu, "sudo" command is also needed)
- brew http://brew.sh/ (Mac OS X, “sudo” command with administrator rights)

WARNING: the installer will install the most recent version of the following applications except indicated otherwise. Please check that this is compatible with your current usages.

- **Vmatch** version 2.3.0 (http://www.vmatch.de/download.html)
- **EMBOSS** version 5.0.0 or upper (http://emboss.sourceforge.net/)
- **Prodigal** version 2.6.3 (https://github.com/hyattpd/Prodigal)
- **MacSyFinder** version 1.0.5 (https://github.com/gem-pasteur/macsyfinder)
- **ClustalW** (version 2.1) (http://www.clustal.org/clustal2/)
- **Muscle** (version 3.8.31) (http://www.drive5.com/muscle)
- **Perl** (https://www.perl.org/). The installer_MAC.sh will install perl5.
- **BioPerl** version 1.6.2 or upper (http://bioperl.org/)
- installer_MAC.sh will also install **prokka-1.12** and **tbl2asn**

The following BioPerl or Perl modules will be installed: Class::Struct, Bio::DB::Fasta, Bio::Tools::Run::Alignment::Clustalw, Bio::Tools::Run::Alignment::Muscle, Date::Calc, File::Copy, Bio::Seq, Bio::SeqIO, and JSON::Parse. See https://www.cpan.org/modules/ for further information.

Content of the CRISPRCasFinder archive

- **CRISPRCasFinder.pl** (main program)
- **CasFinder-2.0** folder (HMM profiles and models of Cas genes used by MacSyFinder http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0110726)
- **installer_MAC.sh** (Mac OS X bash script used for installing the program in order to run CRISPRCasFinder.pl from the CRISPRCasFinder folder)
- **installer_UBUNTU.sh** (Ubuntu bash script used for installing the program in order to run CRISPRCasFinder.pl from the CRISPRCasFinder folder)
- **CRISPRCasFinder.patch** (needed for installer_MAC.sh)
- **install_test** folder (data allowing to test the installation) including:
  1. **sequence.fasta** (example of FASTA file)
  2. **Crisprs_REPORT.tsv** (CRISPR predictions on sequence.fasta)
  3. **Cas_REPORT.tsv** (Cas prediction on sequence.fasta)
- **supplementary_files** folder including:
  1. **CRISPR_crisprdb.csv** (CSV file containing summary information on CRISPR arrays extracted from CRISPRdb during its last update)
  2. **repeatDirection.tsv** (file containing orientation of CRISPRdb consensus repeats using CRISPRDirection https://academic.oup.com/bioinformatics/article/30/13/1805/2422159)
  3. **Repeat_List.csv** (CSV file containing consensus DRs from CRISPRdb, their ID and their number of occurrence in the database)
  4. **crispr.css** (CSS file allowing to visualize results as in current CRISPRdb)

Launching the installation

MACOSX

run installer:

```
./installer_MAC.sh
```

Ubuntu

i) pre-processing stage

before running installer_UBUNTU.sh, allow apt-get with both the multiverse and the restricted repositories as follows:

Edit file /etc/apt/sources.list & uncomment multiverse and backports lines

(multiverse is needed for clustalw installation and backports is needed for prodigal installation)

ii) run installer

```
bash installer_UBUNTU.sh
```

iii) set environment variables

```
source ~/.profile
```

iv) restore the /etc/apt/sources.list to initial commented file if needed.

Check the install:

The installation will have created the following folders or files:

-**bin** folder with binary files (clustalw2, MacSyFinder, mkvtree2, vmatch2, vsubseqselect2)

-**src** folder with source files (Vmatch)

-**macsyfinder-1.0.5** folder (MacSyFinder)

-**sel392v2.so** file (needed for Vmatch)

With installer_UBUNTU.sh, the installation will have defined $MACSY_HOME and redefined $PATH (check with echo command e.g. `echo $MACSY_HOME`). With installer_MAC.sh, MacSyFinder is installed in /usr/local.

functional test

Once the installation is complete, launch one functional test:

```
perl CRISPRCasFinder.pl -cf CasFinder-2.0 -def General -cas -i install_test/sequence.fasta -out Results_test_install -keep
```

and check that the two output files conform to expectation:

```
diff Results_test_install/TSV/Cas_REPORT.tsv install_test/Cas_REPORT.tsv
```

```
diff Results_test_install/TSV/Crisprs_REPORT.tsv install_test/Crisprs_REPORT.tsv
```

The diff commands will return the list of differences if any. Some minor differences could be observed in function of the version of EMBOSS that has been installed (e.g. version 5.0.0 or version 6.6.0).

Using CRISPRCasFinder

Input FASTA file

See https://en.wikipedia.org/wiki/FASTA_format for FASTA format definition.

The fasta file should meet additional constraints:

- must be identified/named (the ID follows character “>”, and a description could be added after a space character),
- the ID should not contain special characters such as “$%”,
- The | sign can be used as field separator
- the file must contain nucleotides (not amino acids),
- the file may contain several sequences in FASTA format,
- each ID must be unique,
- the ID and the file name must not be too long (a limit of 30 is advised),
- the ID will be used for output.

List of CRISPRCasFinder options

The list of options provided below is indicative. The reference list of available options can be accessed via the ‘perl CRISPRCasFinder.pl –help’ command.

Options waiting for a given parameter (filename, text, or number) are followed by symbols "[XXX]". Other options could be considered as booleans (yes or no, 1 or 0).

[Input/Output and -so]

- **-in** or **-i** [XXX] Input Fasta file (recognised extensions: .fasta, .fna, .mfa, .fa, .txt)
- **-outdir** or **-out** [XXX] Output directory. Default output will begin by "Result_", the input filename, followed by current date and time (day, month, year, hours, minutes, and seconds)
- **-LOG** or **-log** Option allowing to write LOG files
- **-keepAll** or **-keep** Option allowing to keep temporary folders/files (Prodigal, Cas-finder, rawFASTA, Properties)
- **-HTML** or **-html** Option allowing to display results as a static HTML web page. The web page created (index.html) will be dependent of the CSS file provided in supplementary_files named crispr.css.
- **-copyCSS** [XXX] Option allowing to copy provided CSS file into "Visualization" repository if option -HTML is set (default: 'supplementary_files/crispr.css' (Ubuntu) or ‘/usr/local/share/CRISPRCasFinder/crispr.css’ (Mac OS X))

[Detection of CRISPR arrays]

- **-mismDRs** or **-md** [XXX] Percentage of mismatches allowed between DRs (default: 20)
- **-truncDR** or **-t** [XXX] Percentage of mismatches allowed for truncated DR (default: 33.3)
- **-minDR** or **-mr** [XXX] Minimal size of DRs (default: 23)
- **-maxDR** or **-xr** [XXX] Maximal size of DRs (default: 55)
- **-minSP** or **-ms** [XXX] Minimal size of Spacers (default: 25)
- **-maxSP** or **-xs** [XXX] Maximal size of Spacers (default: 60)

- **-noMism** or **-n** Option used to do not allow mismatches (default value is 1 when this option is not called. i.e. mismatches are allowed by default)
- **-percSPmin** or **-pm** [XXX] Minimal Spacers size in function of DR size (default: 0.6)
- **-percSPmax** or **-px** [XXX] Maximal Spacers size in function of DR size (default: 2.5)
- **-spSim** or **-s** [XXX] Maximal allowed percentage of similarity between Spacers (default: 60)
- **-DBcrispr** or **-dbc** [XXX] Option allowing to use a CSV file of all CRISPR candidates contained in CRISPRdb (from last update) (default: 'supplementary_files/CRISPR_crisprdb.csv' (Ubuntu) or '/usr/local/share/CRISPRCasFinder/CRISPR_crisprdb.csv' (Mac OS X))
- **-repeats** or **-rpts** [XXX] Option allowing to use a consensus repeats list such as generated by CRISPRdb in order to assign IDs and occurrence (default: 'supplementary_files/Repeat_List.csv' (Ubuntu) or '/usr/local/share/CRISPRCasFInder/Repeat_List.csv' (Mac OS X))
- **-DIRrepeat** or **-drpt** [XXX] Option allowing to use a file file containing repeat IDs and orientation according to CRISPRDirection (default: 'supplementary_files/repeatDirection.tsv' (Ubuntu) or '/usr/local/share/CRISPRCasFinder/repeatDirection.tsv' (Mac OS X))
- **-flank** or **-fl** [XXX] Option allowing to set the size of flanking regions in base pairs (bp) for each analyzed CRISPR array (default: 100)
- **-levelMin** or -**lMin** Option allowing to choose the minimum evidence-level corresponding to CRISPR arrays to display (default: 1)

[Detection of Cas clusters]

- **-cas** or **-cs** Search corresponding Cas genes using MacSyFinder (default value is 0, i.e. Cas genes will not be searched)
- **-ccvRep** or **-ccvr** Option used to write the CRISPR-Cas_systems_vicinity.tsv report in the TSV output folder (CRISPRs and Cas) if option -cas is set (default: 0)
- **-vicinity** or **-vi** [XXX] Option used combined to –ccvRep to define number of nucleotides separating a CRISPR array from its neighboring Cas system (default: 600)
- **-CASFinder** or **-cf** [XXX] Option allowing to change the CasFinder models (default: 'CasFinder-2.0' or '/usr/local/share/macsyfinder/CasFinder-2.0')
- **-cpuMacSyFinder** or -**cpuM** [XXX] Option allowing to set number of CPUs to use for MacSyFinder (default: 1)
- **-rcfowce** Option allowing to run CasFinder only when a CRISPR is detected (default: 0) (set if -cas is set)
- **-definition** or **-def** [XXX] Option allowing to specify Cas-finder definition (if option -cas is set) to be more or less stringent (default: 'General' or 'G'). Other allowed parameters are 'Typing' (or 'T'), and 'SubTyping' (or 'S'). For more information, see MacSyFinder documentation
- **-gffAnnot** or **-gff** [XXX] Option allowing to provide an annotation GFF file (if options -cas and -faa are set) (default: '')
- **-proteome** or **-faa** [XXX] Option allowing to provide a proteome file '.faa' (if options -cas and -gff are set) (default: '')
- **-cluster** or **-ccc** [XXX] Option allowing to constitute clusters or groups of CRISPR or Cas systems given a determined threshold e.g. 20000 bp (default: 0). The output file CRISPR-Cas_clusters.tsv will be created in the TSV output folder.
- **-getSummaryCasfinder** or **-gscf** Option allowing to create file Casfinder_summary_#sequenceID.tsv in TSV output folder (default: 0)
- **-geneticCode** or **-gcode** [XXX] Option allowing to modify the genetic code (translation table) for CDS annotation (default: 11)

Examples of command lines

In most cases the parameters indicated in the examples below are the default values.

(1) The minimal command line (Mac OS: do not include '.pl'):

```
CRISPRCasFinder.pl –in sequence.fasta
```

By default, the result folder will be in the directory named:

"Result_sequence_DD_MM_YYYY_h_m_s" (day, month, year, hours, minutes, and seconds)

(2) full list of options modifying the CRISPR identification parameters

```
CRISPRCasFinder.pl -in sequence.fasta -md 20 -t 33.3 -mr 23 -xr 55 -ms 25 -xs 60 -pm 0.6 -px 2.5 -s 60
```

(3) options allowing to change source data locations

```
CRISPRCasFinder.pl -in multifasta.fna -drpt supplementary_files/repeatDirection.tsv -rpts
supplementary_files/Repeat_List.csv -cas -ccvr -dbc supplementary_files/CRISPR_crisprdb.csv -html
```

(4) options including the allocation of more processing capacity to MacSyFinder

```
CRISPRCasFinder.pl -in multifasta.fna -cas -rcfowce -log -out Results_multifasta -cpuMacSyFinder 8
```

(5) Getting CRISPR arrays and/or Cas systems organized as clusters containing elements having at most 20000 bp of difference between them

```
CRISPRCasFinder.pl -in multifasta.fna -cas -cf CasFinder-2.0 -ccc 20000 -def SubTyping -out
Results_with_clusters_and_Cas_subtyping_level
```

(6) Providing proteome and annotation (GFF) files for searching Cas genes

```
CRISPRCasFinder.pl -in sequence.fasta -cas -cf CasFinder-2.0 -gff path/to/sequence.gff -proteome
path/to/sequence.faa -keep
```

(7) Displaying the current version of the program

```
CRISPRCasFinder.pl -v
```

Output Files

Files/folders generated when using specific standalone options are indicated by an asterisk *). #sequenceID" represents the ID of analyzed sequence(s). "DD_MM_YYYY_h_m_s" represents the current date (day, month, and year) and time (hours, minutes, and seconds).

File or Folder	Description
result.json	JSON file containing main information on detected CRISPRarrays and Cas genes.
TSV	Folder containing .tsv and .xls files.
*Casfinder_summary_#sequenceID.tsv	Summary file generated by MacSyFinder. Contains information on Cas proteins. The option '-gscf' must be set to get this file.
Cas_REPORT.tsv	File containing information on detected Cas systems and genes. This file is also available in Excel format.
*CRISPR-Cas_clusters.tsv	File containing information on CRISPR arrays and Cas systems that are close to each other (e.g. separated by less than 20000 bp). The option '-ccc' is needed to get this file.
CRISPR-Cas_summary.tsv	File containing summary information on both CRISPRs and Cas. This file is also available in Excel format.
*CRISPR-Cas_systems_vicinity.tsv	File containing information on CRISPR arrays and their nearest Cas system, in function of a maximal distance defined with option '-vicinity'. The option '-ccvr' is needed to get this file.
*crisprs_orientations_count.tsv	File showing a comparison between predictions made using precomputed results from CRISPRDirection versus AT% calculation in flanking regions. The option '-drpt' must be set to get this file.
Crisprs_REPORT.tsv	File containing information on detected CRISPR arrays. This file is also available in Excel format.
GFF	Folder containing GFF3 files.
annotation_#sequenceID.gff	File generated by Prodigal, containing CDS of the given sequence.
#sequenceID.gff	File generated by CRISPRCasFinder, containing annotations of CRISPR arrays (repeats and spacers) as well as their flanking sequences.
***LOGs**	
*logFile_DD_MM_YYYY_h_m_s.txt	File containing the output that CRISPRCasFinder produced during its run. This file is created if option “-log” is set.
*logSequences_DD_MM_YYYY_h_m_s.tsv	File containing some statistics related to the computation time for CRISPR and Cas searches and other features for each sequence. This file is created if option “-log” is set.
***Visualization**	Folder containing an HTML page and a CSS file allowing to visualize CRISPRs and Cas with the same design as CRISPRdb (these files are generated if options "-html" and "-copyCSS" were set).
The four folders listed below are kept if the option "-keep" is set:	
CasFinder	Folder containing files generated by MacSyFinder.
CRISPRFinderProperties	Folder containing files generated by CRISPRFinder (CRISPR properties, FASTA files of DRs and spacers, and alignments).
Prodigal	Folder containing files generated by Prodigal
rawFASTA	Folder containing FASTA sequences of all detected CRISPR arrays ("rawCRISPRs.fna") and Cas genes ("rawCas.fna"). FASTA files of analyzed sequences are also contained in this folder.

CRISPRCasViewer

The archive contains a **CRISPRCasViewer** folder (a standalone viewer of CRISPRCasFinder.pl results). This folder contains:

1. **CRISPRCasViewer.html** (main HTML page of the standalone viewer that can be opened in a web browser) and one html file for each of the viewing options
2. **css** folder (CSS files)
3. **scripts** folder (JavaScript files)
4. **result.json** (JSON example result file)

Installing the Viewer

CRISPRCasViewer v-1.0.1 is best used with Firefox, Chrome and Safari. It does not work with the tested versions of Internet Explorer or Edge. The viewer is launched by opening the “CRISPRCasViewer.html” file or by dragging the file on the web browser window. A result.json file as produced by CRISPRCasFinder needs to be selected. A sequence is then selected within the list of previously analyzed sequences.

Three different views can be shown, linear, circular, and a plot of the start and end points of each locus detected. Each of these views can be zoomed in, by selecting a scale and dragging the ruler (linear view), using the mouse wheel (circular plot) or selecting a portion of the chart (scatter plot view). Different image formats can be exported.

Summary tables on CRISPR arrays and Cas systems are presented under each of the charts. Information on individual loci is provided when passing the mouse over.

Frequently Asked Questions

- Why do I need the ".so" file ("sel392v2.so") to launch CRISPRCasFinder.pl?

This ".so" file is mandatorily required by Vmatch and dependencies. See Vmatch documentation and the option "-so" for more details. With Ubuntu, the file should be present in the same folder as CRISPRCasFinder.pl. With Mac OS, the file is present in ‘/usr/local/lib’.

- Why do I need the "CasFinder-2.0" folder when I search for CRISPR associated Cas genes?

This folder ("CasFinder-2.0") is needed because it contains HMM profiles and Cas models allowing definition of Cas genes by MacSyFinder. With Mac OS X, the folder is located in /usr/local/share/macsyfinder. Do not rename included files and subfolders.