CdiffFragR Manual

C. difficile Fluorescent PCR Ribotyping v 12.7.21

CdiffFragR was developed to match query *Clostridioides difficile* chromatograms (.fsa file format) generated on capillary electrophoresis sequencers to a database of know *C. difficile* ribotypes. CdiffFragR is an R tool for DNA fragment analysis developed and maintained by the Walk Lab at MSU Bozeman. CdiffFragR is compatible with both MacOS X and Windows 10.

DOWNLOADING SOFTWARE

1.) DOWNLOAD R
If you do not already have R installed on your computer:

For Windows, go to:
https://cran.r-project.org/bin/windows/base/

For MacOS, go to:
https://cran.r-project.org/bin/macosx/

and download/install at least version 4.0. If you already have R on your computer, please upgrade to version 4.0 or more recent. We cannot guarantee backwards compatibility.

2.) DOWNLOAD RTOOLS
If you do not already have Rtools installed on your computer:

For Windows, go to:
https://cran.r-project.org/bin/windows/Rtools/

For MacOS:
Rtools is not needed for MacOS (the required tools are available in base R).

and download/install at least version 4.0. If you already have R on your computer, please upgrade to version 4.0 or more recent. We cannot guarantee backwards compatibility.

3.) DOWNLOAD RSTUDIO
If you do not already have RStudio on your computer:

For both Windows and MacOS, go to:
https://www.rstudio.com/products/rstudio/download/

and download/install at least version 1.4. If you already have Rstudio on your computer, please upgrade to version 1.4 or more recent. We cannot guarantee backwards compatibility.

4.) DOWNLOAD CdiffFragR
For both Windows and MacOS, go to:
       http://github.com/nvpinkham/CdiffFragR
and download the program by clicking the <mark>Code</mark> button and the <mark>Download Zip</mark> option. Save this to your computer.
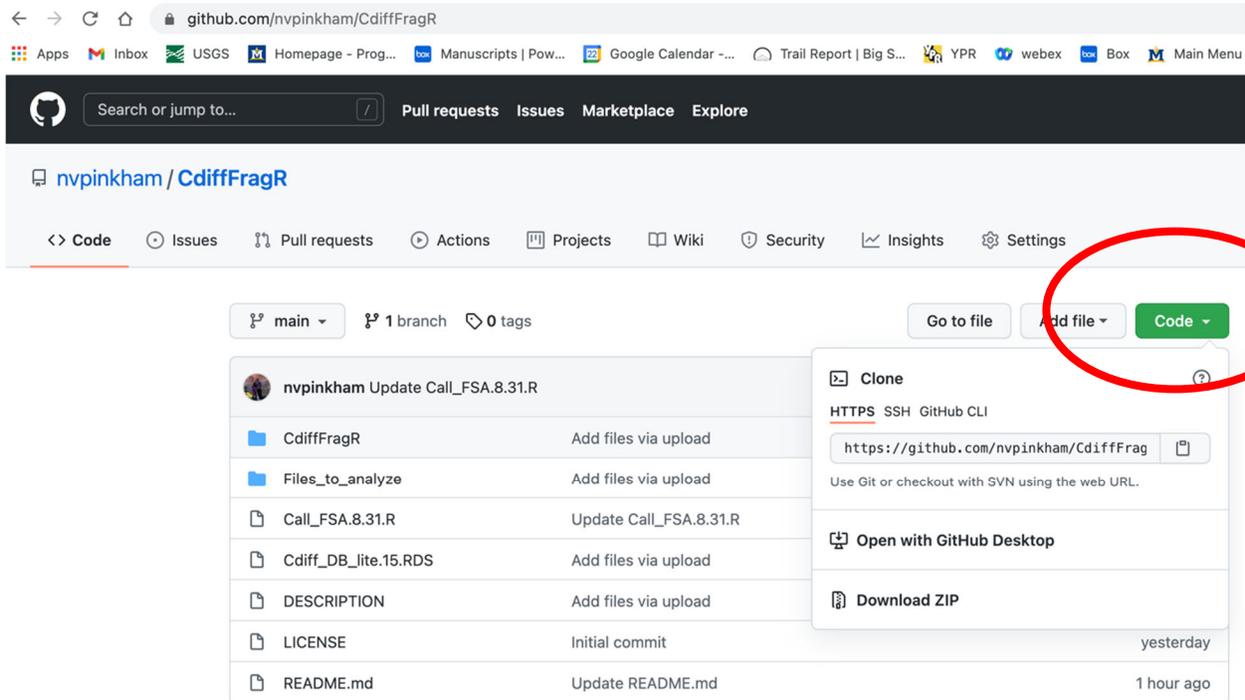
Following download, extract the contents of the zip file to a directory on your computer.

        Windows – Right click on <mark>CdiffFragR-main.zip</mark> and select <mark>Extract All…</mark> from the dropdown window. When prompted, click the <mark>Browse</mark> button and select a location (e.g. <mark>Desktop</mark>).

        Mac – Uncompress "CdiffFragR-main.zip"

**5.) RUNNING CdiffFragR FOR THE FIRST TIME**

CdiffFragR uses functions in the bayestestR, vegan, and zip packages. If these packages are not currently installed, they will be when running the program for the first time (i.e. dependencies are required and will be taken care of the first time the program is executed).

We set up the download process to run a test on previously generated datafiles before running new datafiles. If you look in the directory "Files_to_analyze", you will see some example .fsa files. The following is a test run for these files:

Open "CdiffFragR-main.Rproj" with Rstudio by double-clicking the program.

Once the script is loaded, highlight all the text in the script window (upper-left) and click the <mark>Run</mark> button.
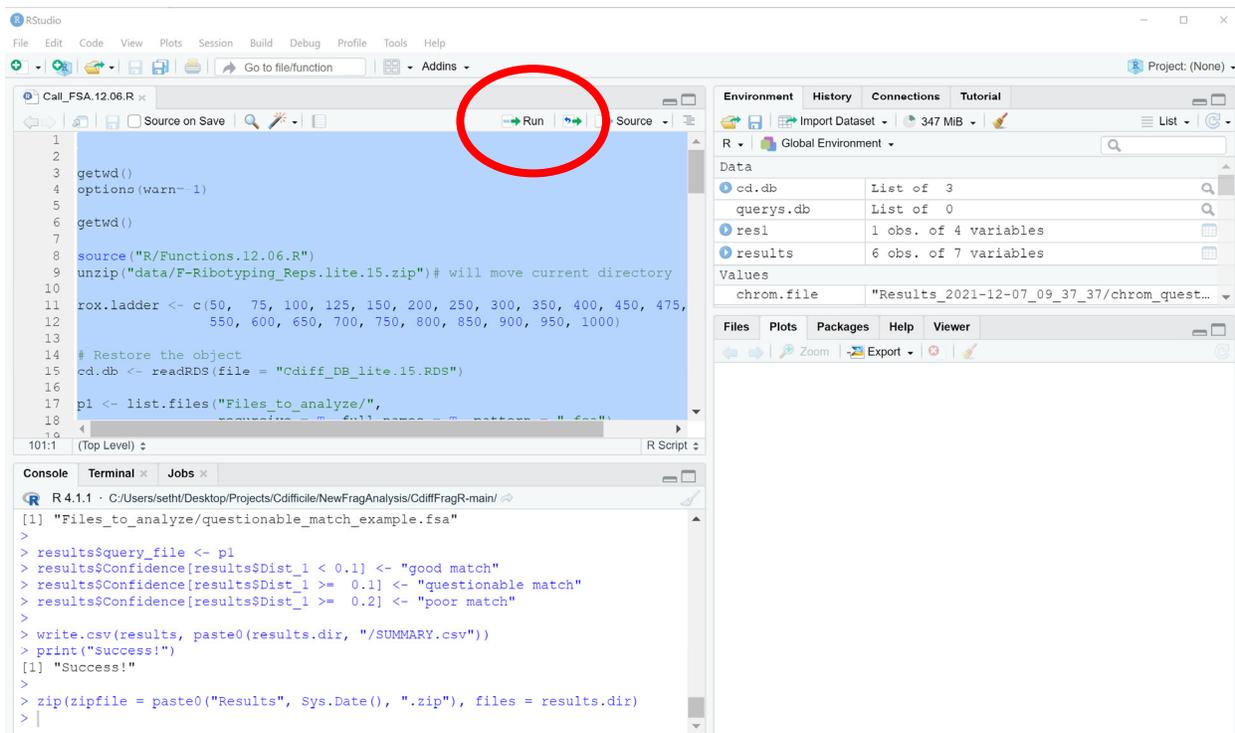
The program will call peaks and process the .fsa chromatogram files, moving them from the "Files_to_analyze" directory to a new directory it creates, called "Files_analyzed". The results will populate a new directory called "Results_YYYY-MM-DD_Hour_Min_Sec". The script generates both jpeg summaries of chromatograms (filenames begin with "chrom_") and overlay plots comparing queries against the closest (best) match to the database (filenames begin with "hits_"). A "SUMMARY.csv" table is also produced, summarizing the results as follows.

Column Name	Definition
query_file	Directory and filename of query file
Confidence	Computationally defined match
Hit_1	Directory and filename of the closest match to the database
ribo_1	Ribotype of closest database match

Dist_1	Bray-Curtis dissimilarity between query and database chromatogram
hit_1_jpeg	Directory and filename of query-match overlay plot
chromatogram_jpeg	Directory and filename of chromatogram

6.) INTERPRETING RESULTS

1) Open the SUMMARY.csv file and look at the Dist_1 column. Numbers in this column refer to the Bray-Curtis dissimilarity between the query (your file) and hit (database file). Bray-Curtis values range between 0 and 1, and the closer the number is to 0, the better the match. Based on tens of thousands of analyzed .fsa files, the "Confidence" of each match falls into one of three categories, as follows:

Dist_1	Confidence
<0.10	"good match"
0.10 – 0.20	"questionable match"
>0.20	"poor match"

2) Our interpretation:
   Good match – This is a solid match to the database with a very low chance of a false positive.

   Poor match – This is not a match to the database. Take a look at the chromatogram and hit images. It is likely that either 1) the chromatogram looks fine, and the ribotype is not yet in our database, or 2) something went wrong with the amplification or capillary electrophoresis and the sample should be re-analyzed.

   Questionable match – This may or may not be a match to the database. Take a look at the hit image and see whether the peaks line up (peaks within 1-2 base pairs should be considered the same peaks).

   Examples:
   From the SUMMARY.csv file, you will see that one of the queries
   ("questionable_match_example.fsa")
   had a "questionable match" in the "Confidence" column, meaning the Bray-Curtis dissimilarity to the database was 0.10-0.20. If you open up the corresponding chromatogram
   ("chrom_ questionable_match_example.jpeg")
   you will see that the peaks are quite small compared to those of the size standard. This isolate should be re-analyzed.

From the SUMMARY.csv file, you will see that one of the queries ("poor_match_example.fsa")
had a "poor match". If you open up the corresponding chromatogram ("chrom_ poor_match_example.jpeg")
you will see that the peaks are nice and tall compared to those of the size standard, indicating that there was plenty of signal. Next, if you open up the corresponding query-match overlay plot ("query_ poor_match_example.jpeg")
you will see that the query contains at least two peaks that the best match to the database does not and is missing at least two peaks that the best match to the database has. This is likely a new ribotype. COOL!

For reference, queries producing typical matches ("good_match_example.fsa" and all other .fsa files) are provided. All .fsa files currently in the database (F-ribotyping_Reps.lite.15) are available from the Walk Lab upon request.

New ribotypes:
If you believe you have identified a new ribotype, it would be best to re-analyze and confirm the result. Next, please contact the Walk Lab to see if it can be added to the database. We would love to continue developing this resource for the scientific community and adding new ribotypes as they are observed is essential to this resource.

RUNNING CdiffFragR AGAIN

Each time you wish to run CdiffFragR, simply drop in .fsa files into the "Files_to_analyze" directory. The program simply looks for files there and generates a time-stamped results directory ("Results_YYYY-MM-DD_Hour_Min_Sec") on every Run.

***From time to time, please check the Walk Lab website (thewalklab.com/tools) to make sure you're working with the most recent version of the database. We will periodically update ribotype calls based on comparative genomics datasets and versions will be listed on the website.

To update, simply delete the current CdiffFragR version and download the updated version from:
   http://github.com/nvpinkham/CdiffFragR