NOV 09, 2022

SHARE

WORKS FOR ME 1

Formatting Data for Use With ebFRET Hidden Markov Modeling Software

This protocol is published without a DOI.

Clark Fritsch[1]

[1]University of Pennsylvania

Clark Fritsch
University of Pennsylvania

COMMENTS 0

ABSTRACT

This protocol follows from the "Selecting a Region of Interest for Hidden Markov Modeling using ebFRET" protocol and is the second step towards analyzing your data using Hidden Markov Modeling. This protocol describes the steps you need to take to convert traces containing your region of interest (selected in the previous protocol) and formatting them for use with ebFRET, a software package that is capable of performing Hidden Markov Modeling on single-molecule FRET datasets.

PROTOCOL CITATION

Clark Fritsch 2022. Formatting Data for Use With ebFRET Hidden Markov Modeling Software
protocols.io
https://protocols.io/view/formatting-data-for-use-with-ebfret-hidden-markov-civaue2e

LICENSE

CREATED

Nov 05, 2022

LAST MODIFIED

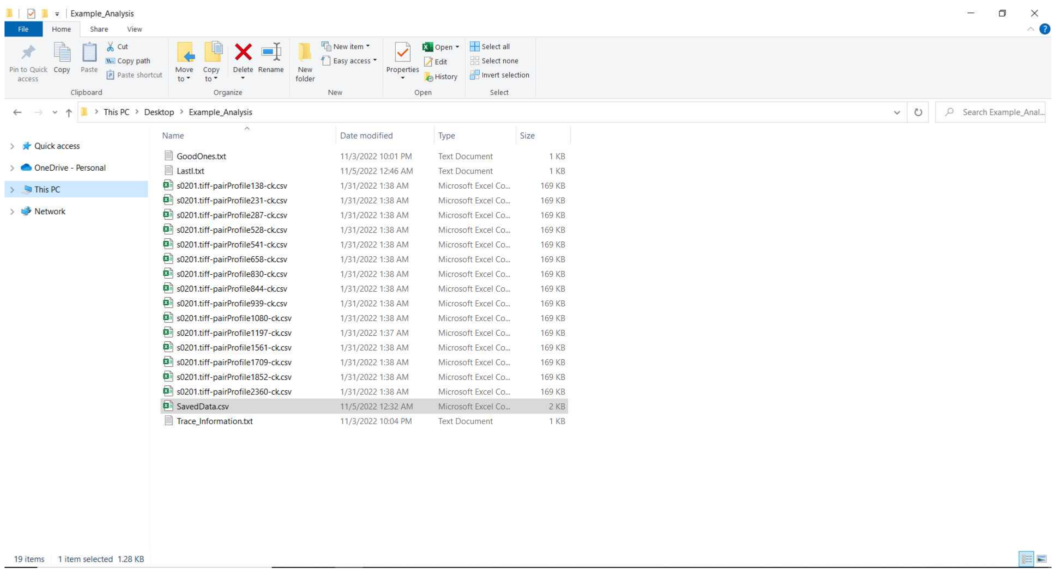
Nov 09, 2022

PROTOCOL INTEGER ID

72322

1 This protocol is a continuation of the "Selecting a Region of Interest for Hidden Markov Modeling using ebFRET" practice protocol.

2 To begin this protocol, you should have your "SavedData.csv" file and traces in one directory, as shown below:



Name	Date modified	Type	Size
GoodOnes.txt	11/3/2022 10:01 PM	Text Document	1 KB
LastI.txt	11/5/2022 12:46 AM	Text Document	1 KB
s0201.tiff-pairProfile138-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile231-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile287-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile528-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile541-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile658-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile830-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile844-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile939-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1080-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1197-ck.csv	1/31/2022 1:37 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1561-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1709-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1852-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile2360-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
SavedData.csv	11/5/2022 12:32 AM	Microsoft Excel Co...	2 KB
Trace_Information.txt	11/3/2022 10:04 PM	Text Document	1 KB

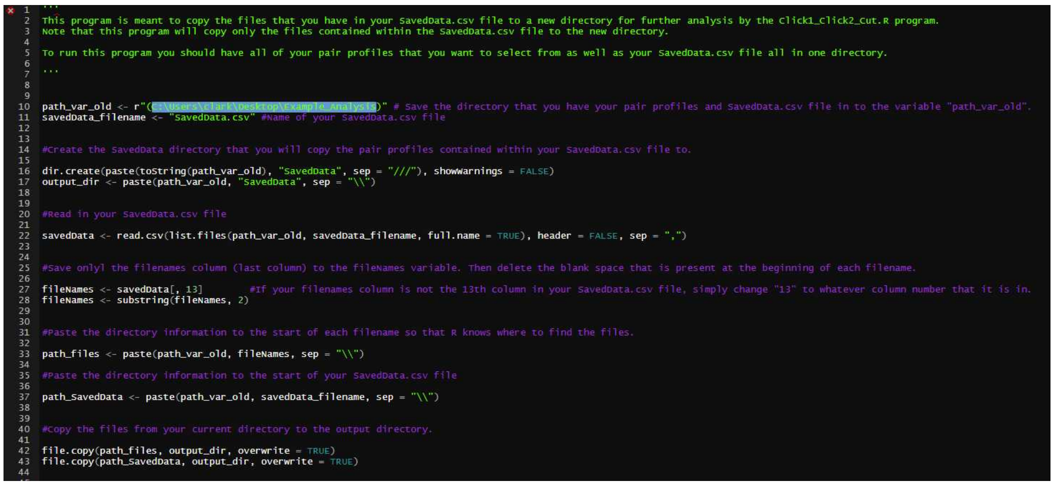
It is likely that when you are doing your real analysis that you don't include every trace that is included in your "GoodOnes.txt" file in your "SavedData.csv" file. For example, you might think that a trace is good at first and include it in your "GoodOnes.txt" file, but find upon closer inspection that the trace is not worthy of further analysis and therefore exclude it from your "SavedData.csv" file.

Because of this, you want to separate the traces that are included in your "SavedData.csv" file from the remainder of your traces.

3 To separate the traces included in your "SavedData.csv" file from the remainder of your traces, you can use the "SavedData_Copy_and_Transfer_v3.R" program that is attached below:

SavedData_Copy_and_Transfer_v3.R

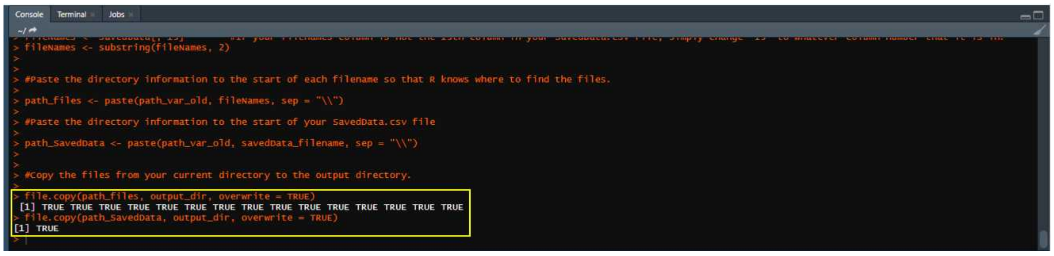
4 To use this program in R, simply copy the path to the directory that contains all of your traces and the "SavedData.csv" file and paste the path into the "path_var_old" variable in between the parentheses, as highlighted below:



```
'''
This program is meant to copy the files that you have in your SavedData.csv file to a new directory for further analysis by the Click1_Click2_Cut.R program.
Note that this program will copy only the files contained within the SavedData.csv file to the new directory.

To run this program you should have all of your pair profiles that you want to select from as well as your SavedData.csv file all in one directory.

'''


path_var_old <- r"(C:\Users\clark\Desktop\Example_Analysis)" # Save the directory that you have your pair profiles and SavedData.csv file in to the variable "path_var_old".
savedData_filename <- "SavedData.csv" #Name of your SavedData.csv file


#Create the SavedData directory that you will copy the pair profiles contained within your SavedData.csv file to.

dir.create(paste(toString(path_var_old), "SavedData", sep = "///"), showWarnings = FALSE)
output_dir <- paste(path_var_old, "SavedData", sep = "\\")


#Read in your SavedData.csv file

savedData <- read.csv(list.files(path_var_old, savedData_filename, full.name = TRUE), header = FALSE, sep = ",")


#Save onlyl the filenames column (last column) to the fileNames variable. Then delete the blank space that is present at the beginning of each filename.

fileNames <- savedData[, 13]        #If your filenames column is not the 13th column in your SavedData.csv file, simply change "13" to whatever column number that it is in.
fileNames <- substring(fileNames, 2)


#Paste the directory information to the start of each filename so that R knows where to find the files.

path_files <- paste(path_var_old, fileNames, sep = "\\")

#Paste the directory information to the start of your SavedData.csv file

path_SavedData <- paste(path_var_old, savedData_filename, sep = "\\")


#Copy the files from your current directory to the output directory.

file.copy(path_files, output_dir, overwrite = TRUE)
file.copy(path_SavedData, output_dir, overwrite = TRUE)
```

Then press "Ctrl A" to highlight all of the code and press "Ctrl Enter" to run the code.

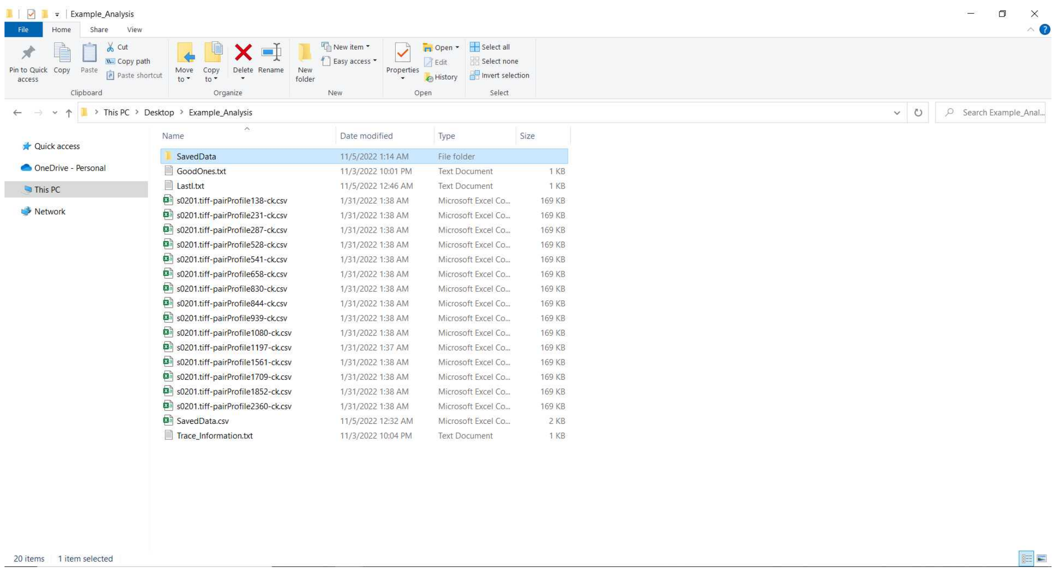
5 After running the code, you will see that the following information has been outputted to the console (boxed in yellow):



```
> fileNames <- substring(fileNames, 2)
>
>
> #Paste the directory information to the start of each filename so that R knows where to find the files.
>
> path_files <- paste(path_var_old, fileNames, sep = "\\")
>
> #Paste the directory information to the start of your SavedData.csv file
>
> path_SavedData <- paste(path_var_old, savedData_filename, sep = "\\")
>
>
> #Copy the files from your current directory to the output directory.
>
> file.copy(path_files, output_dir, overwrite = TRUE)
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> file.copy(path_SavedData, output_dir, overwrite = TRUE)
[1] TRUE
>
```

This output simply shows that each trace (TRUE) has been transferred successfully to the output directory. If a trace is not transferred successfully, it was output FALSE. Additionally, your "SavedData.csv" file is transferred to the output directory.

6 After running the program successfully, you will see that the "SavedData" output directory has been created at the path directory that you inputted into the program:

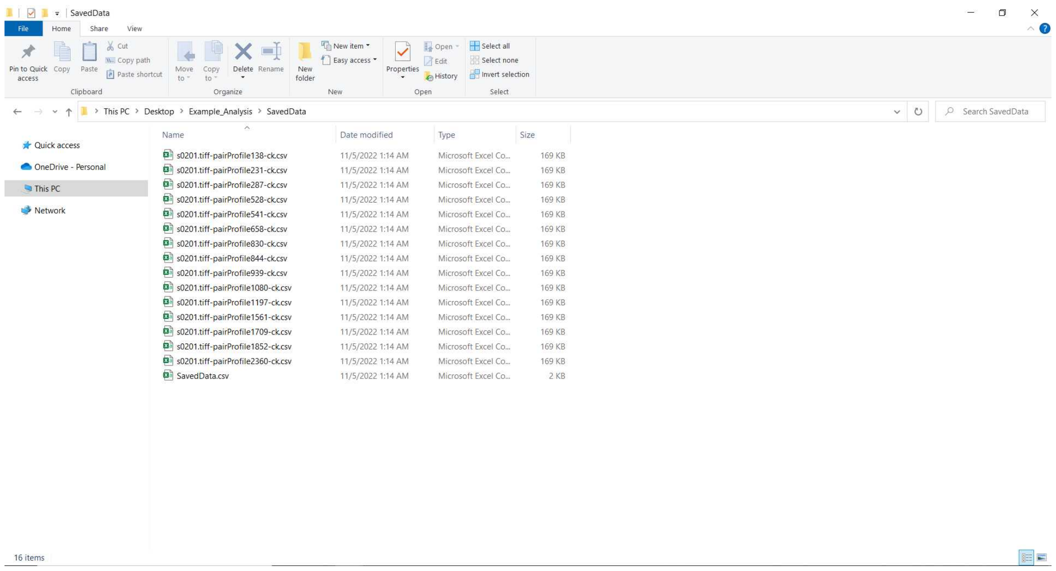


This PC > Desktop > Example_Analysis

Name	Date modified	Type	Size
SavedData	11/5/2022 1:14 AM	File folder	
GoodOnes.txt	11/3/2022 10:01 PM	Text Document	1 KB
LastI.txt	11/5/2022 12:46 AM	Text Document	1 KB
s0201.tiff-pairProfile138-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile231-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile287-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile528-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile541-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile658-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile830-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile844-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile939-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1080-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1197-ck.csv	1/31/2022 1:37 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1561-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1709-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1852-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile2360-ck.csv	1/31/2022 1:38 AM	Microsoft Excel Co...	169 KB
SavedData.csv	11/5/2022 12:32 AM	Microsoft Excel Co...	2 KB
Trace_Information.txt	11/3/2022 10:04 PM	Text Document	1 KB

20 items 1 item selected

This "SavedData" directory will contain only your "SavedData.csv" file and the traces that were contained within your "SavedData.csv" file, as shown below:

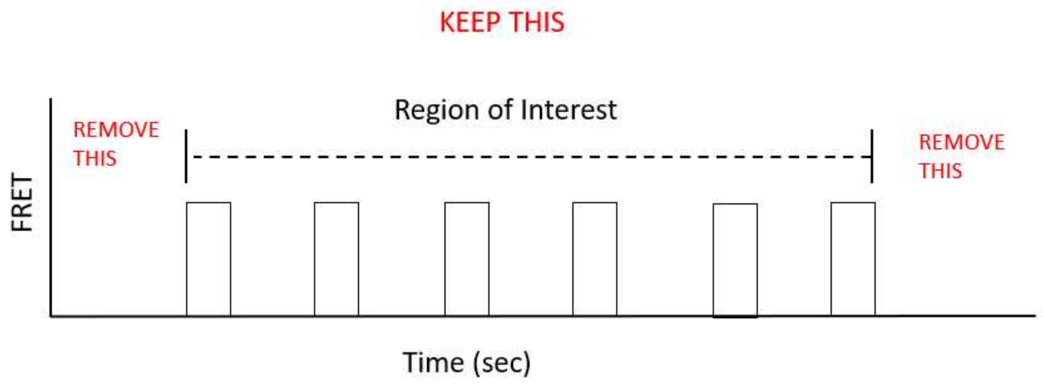


This PC > Desktop > Example_Analysis > SavedData

Name	Date modified	Type	Size
s0201.tiff-pairProfile138-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile231-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile287-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile528-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile541-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile658-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile830-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile844-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile939-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1080-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1197-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1561-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1709-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1852-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile2360-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
SavedData.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	2 KB

16 items

7 Next, you want to cut out the parts of each of your traces that are not included within the range set by Click 1 and Click 2 that is included in your "SavedData.csv" file.

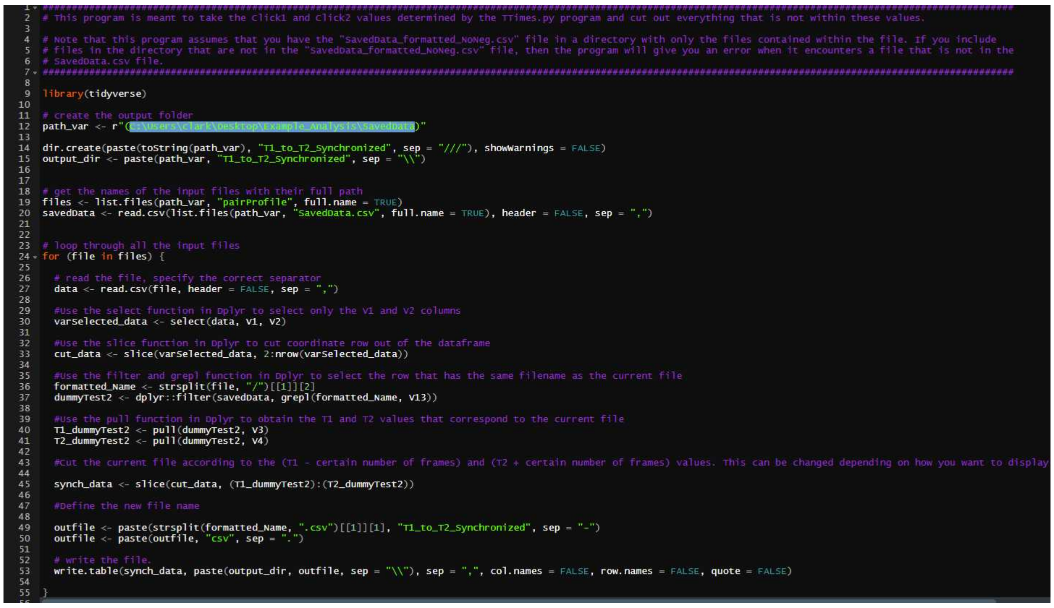


You can do this using the "Click1_Click2_Cut.R" program that is included below:

Click1_Click2_Cut.R

8 The "Click1_Click2_Cut.R" program, as described above, simply takes the values for Click1 (leftmost boundary of your ROI) and Click2 (rightmost boundary of your ROI) and removes the frames from each trace that are not included between Click1 and Click2.

To use this program in R, simply copy the path to "SavedData" directory and paste the path into the "path_var" variable in between the parentheses, as highlighted below:

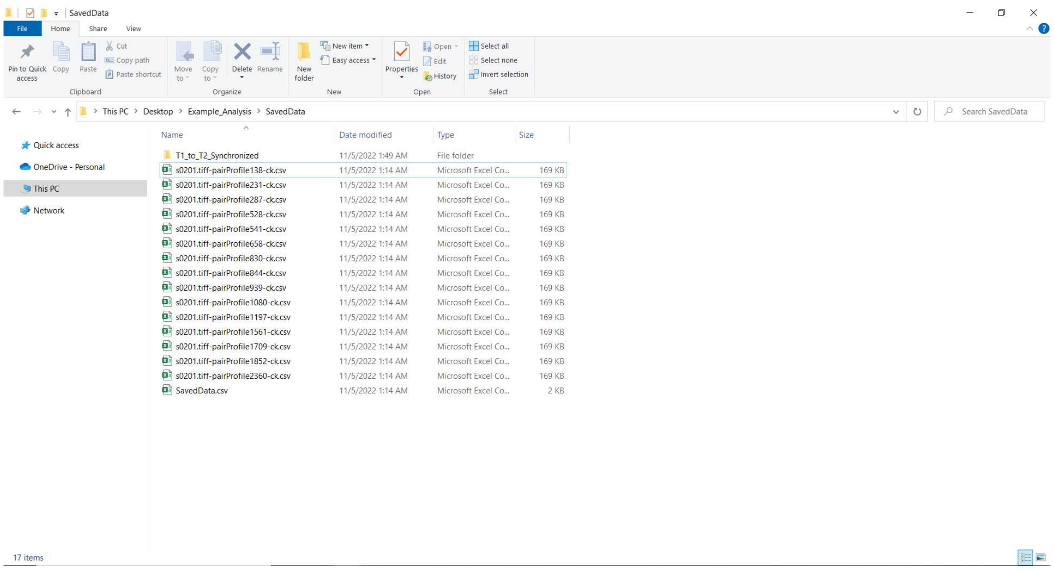


```r
##########################################################################################################################################
# This program is meant to take the Click1 and Click2 values determined by the TTimes.py program and cut out everything that is not within these values.

# Note that this program assumes that you have the "SavedData_formatted_NoNeg.csv" file in a directory with only the files contained within the file. If you include
# files in the directory that are not in the "SavedData_formatted_NoNeg.csv" file, then the program will give you an error when it encounters a file that is not in the
# SavedData.csv file.
##########################################################################################################################################

library(tidyverse)

# create the output folder
path_var <- r"(C:\Users\clark\Desktop\Example_Analysis\SavedData)"

dir.create(paste(toString(path_var), "T1_to_T2_Synchronized", sep = "///"), showWarnings = FALSE)
output_dir <- paste(path_var, "T1_to_T2_Synchronized", sep = "\\")


# get the names of the input files with their full path
files <- list.files(path_var, "pairProfile", full.name = TRUE)
savedData <- read.csv(list.files(path_var, "SavedData.csv", full.name = TRUE), header = FALSE, sep = ",")


# loop through all the input files
for (file in files) {

  # read the file, specify the correct separator
  data <- read.csv(file, header = FALSE, sep = ",")

  #Use the select function in Dplyr to select only the V1 and V2 columns
  varSelected_data <- select(data, V1, V2)

  #Use the slice function in Dplyr to cut coordinate row out of the dataframe
  cut_data <- slice(varSelected_data, 2:nrow(varSelected_data))

  #Use the filter and grepl function in Dplyr to select the row that has the same filename as the current file
  formatted_Name <- strsplit(file, "/")[[1]][2]
  dummyTest2 <- dplyr::filter(savedData, grepl(formatted_Name, V13))

  #Use the pull function in Dplyr to obtain the T1 and T2 values that correspond to the current file
  T1_dummyTest2 <- pull(dummyTest2, V3)
  T2_dummyTest2 <- pull(dummyTest2, V4)

  #Cut the current file according to the (T1 - certain number of frames) and (T2 + certain number of frames) values. This can be changed depending on how you want to display

  synch_data <- slice(cut_data, (T1_dummyTest2):(T2_dummyTest2))

  #Define the new file name

  outfile <- paste(strsplit(formatted_Name, ".csv")[[1]][1], "T1_to_T2_Synchronized", sep = "-")
  outfile <- paste(outfile, "csv", sep = ".")

  # write the file.
  write.table(synch_data, paste(output_dir, outfile, sep = "\\"), sep = ",", col.names = FALSE, row.names = FALSE, quote = FALSE)

}
```

Then press "Ctrl A" to highlight all of the code and press "Ctrl Enter" to run the code.

9 Once you have run the "Click1_Click2_Cut.R" program, a new directory called "T1_to_T2_Synchronized" that will contain your new traces, each of which has been cut to contain

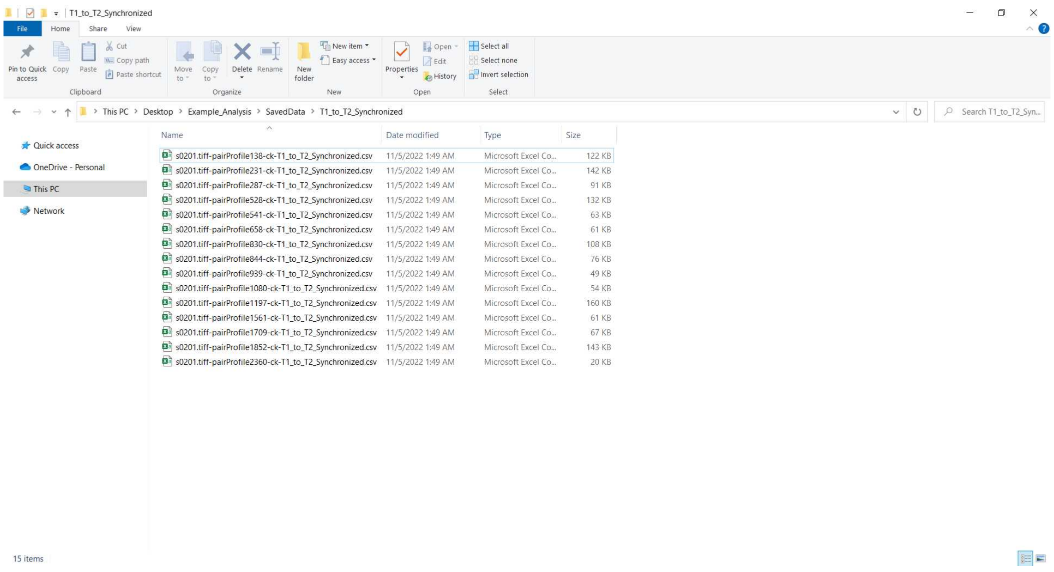
only the frames between your Click1 and Click2 values:


This PC > Desktop > Example_Analysis > SavedData

Name	Date modified	Type	Size
T1_to_T2_Synchronized	11/5/2022 1:49 AM	File folder	
s0201.tiff-pairProfile138-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile231-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile287-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile528-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile541-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile658-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile830-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile844-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile939-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1080-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1197-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1561-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1709-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile1852-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
s0201.tiff-pairProfile2360-ck.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	169 KB
SavedData.csv	11/5/2022 1:14 AM	Microsoft Excel Co...	2 KB

17 items

Each of these new "trimmed" traces can be distinguished from the original traces because the trimmed traces contain "-T1_to_T2_Synchronized" appended to the end of each filename:


This PC > Desktop > Example_Analysis > SavedData > T1_to_T2_Synchronized

Name	Date modified	Type	Size
s0201.tiff-pairProfile138-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	122 KB
s0201.tiff-pairProfile231-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	142 KB
s0201.tiff-pairProfile287-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	91 KB
s0201.tiff-pairProfile528-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	132 KB
s0201.tiff-pairProfile541-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	63 KB
s0201.tiff-pairProfile658-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	61 KB
s0201.tiff-pairProfile830-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	108 KB
s0201.tiff-pairProfile844-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	76 KB
s0201.tiff-pairProfile939-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	49 KB
s0201.tiff-pairProfile1080-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	54 KB
s0201.tiff-pairProfile1197-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	160 KB
s0201.tiff-pairProfile1561-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	61 KB
s0201.tiff-pairProfile1709-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	67 KB
s0201.tiff-pairProfile1852-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	143 KB
s0201.tiff-pairProfile2360-ck-T1_to_T2_Synchronized.csv	11/5/2022 1:49 AM	Microsoft Excel Co...	20 KB

15 items

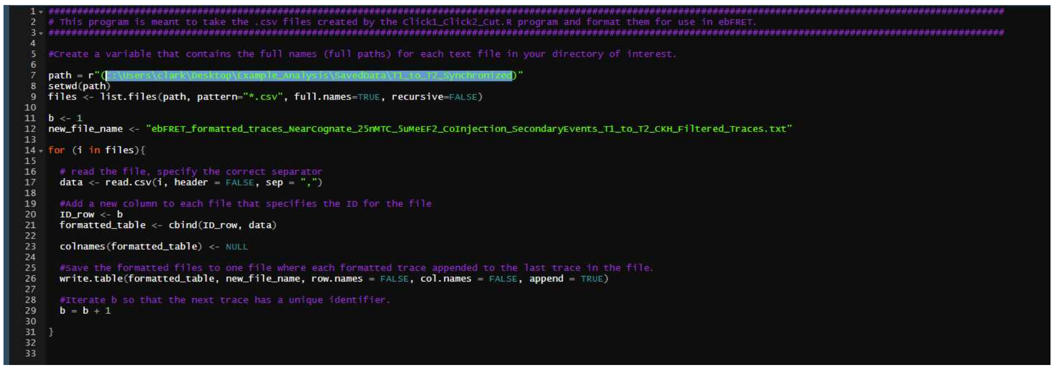
10 Now that you have your trimmed traces prepared, you need to format the traces so that they can be run using the ebFRET Hidden Markov Modeling program.

To do this, you can use "RegionOfInterest_to_ebFRET.R" program that is attached below:

RegionOfInterest_to_ebFRET.R

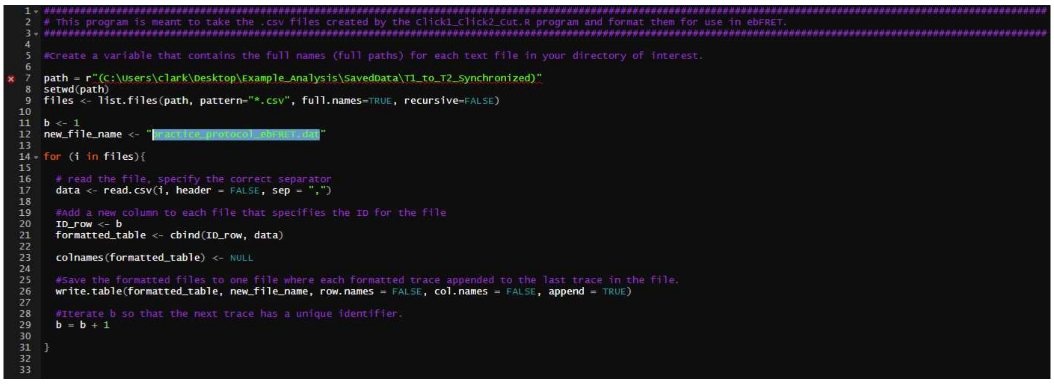
11 To use ebFRET, your traces of interest need to be stitched together with each trace being associated with a unique ID number. Additionally, the donor and acceptor fluorophore intensity for each timepoint in each trace needs to be included in the file.

To use the "RegionOfInterest_to_ebFRET.R" program to format your traces in this way, simply copy the path that contains your trimmed traces into the "path" variable in the program, as highlighted below:



```r
##############################################################################################################
# This program is meant to take the .csv files created by the Click1_Click2_Cut.R program and format them for use in ebFRET.
##############################################################################################################

#Create a variable that contains the full names (full paths) for each text file in your directory of interest.

path = r"(C:\Users\clark\Desktop\Example_Analysis\SavedData\T1_to_T2_Synchronized)"
setwd(path)
files <- list.files(path, pattern="*.csv", full.names=TRUE, recursive=FALSE)

b <- 1
new_file_name <- "ebFRET_formatted_traces_NearCognate_25nMTC_5uMeEF2_CoInjection_SecondaryEvents_T1_to_T2_CKH_Filtered_Traces.txt"

for (i in files){

  # read the file, specify the correct separator
  data <- read.csv(i, header = FALSE, sep = ",")

  #Add a new column to each file that specifies the ID for the file
  ID_row <- b
  formatted_table <- cbind(ID_row, data)

  colnames(formatted_table) <- NULL

  #Save the formatted files to one file where each formatted trace appended to the last trace in the file.
  write.table(formatted_table, new_file_name, row.names = FALSE, col.names = FALSE, append = TRUE)

  #Iterate b so that the next trace has a unique identifier.
  b = b + 1

}
```

Additionally, you need to name the formatted file that your trace information will be outputted to by entering the new filename into the "new_file_name" variable, as shown below:



```r
##############################################################################################################
# This program is meant to take the .csv files created by the Click1_Click2_Cut.R program and format them for use in ebFRET.
##############################################################################################################

#Create a variable that contains the full names (full paths) for each text file in your directory of interest.

path = r"(C:\Users\clark\Desktop\Example_Analysis\SavedData\T1_to_T2_Synchronized)"
setwd(path)
files <- list.files(path, pattern="*.csv", full.names=TRUE, recursive=FALSE)

b <- 1
new_file_name <- "practice_protocol_ebFRET.dat"

for (i in files){

  # read the file, specify the correct separator
  data <- read.csv(i, header = FALSE, sep = ",")

  #Add a new column to each file that specifies the ID for the file
  ID_row <- b
  formatted_table <- cbind(ID_row, data)

  colnames(formatted_table) <- NULL

  #Save the formatted files to one file where each formatted trace appended to the last trace in the file.
  write.table(formatted_table, new_file_name, row.names = FALSE, col.names = FALSE, append = TRUE)

  #Iterate b so that the next trace has a unique identifier.
  b = b + 1

}
```

Note that the new file name must have the ".dat" extension, as this is the only text file format that ebFRET will recognize.

Press "Ctrl A" to highlight all of the code and press "Ctrl Enter" to run the code.

12 Once you have run the program, your ebFRET formatted file will appear in your "T1_to_T2_Synchronized" directory. It is now ready to input into ebFRET for Hidden Markov Analysis.