***READ ME***

## 1. HOW TO USE

After downloading the code from the Sabatinos lab Github, open the R code in an R code editor. In our case, R studio was used. With the code now open in the editor, follow the “read me” instructions and adjust the code to your computer. Changing the parameters at this time won’t matter, as you would focus on the initial scatter plot, which isn’t based on parameters- changing the parameters would be more important in future runnings of the program after a preliminary baseline threshold is established. Next, to run the code (in R studio- this may change in other editors):

1. Save your changes. This can be done with ctrl+ ‘s’, or pressing the save button shown as a floppy disk button, or going to file> save. These options are shown in the red squares in Fig 1.
2. Select the whole code. This could be done with ctrl + ‘a’ in Windows, or command + ‘a’ in Mac.
3. Run the code while the whole code is selected. This can be done with ctrl + “enter button”, or the top left run button can be pressed. These options are shown in the blue squares in Fig 1.

The output should occur in a folder that is specified in the location you wrote in the code (line 61), this output would be generated in less than a couple of minutes.

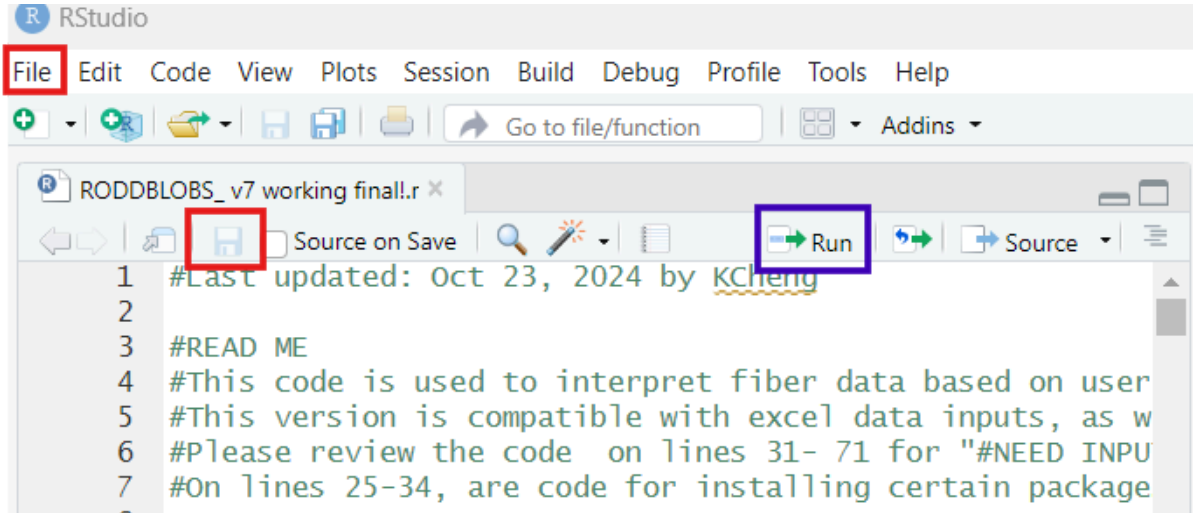



**Figure 1 Important Buttons When Running the Code.** This shows the location and placement of buttons useful when running the code. The buttons shown in the red

square highlight methods to save changes in the code. The blue square highlights the button to use to run the code.

## 2. DATA FILES

Data files currently compatible with the Oct. 23, 2024 version of R-ODD-BLOBS are Excel files (.xlsx and .csv). Multiple sheets can be tolerated within one workbook. The information within these files should have the channels' pixel intensity values to be analyzed. Your channel names should start on the row you specified in line 37 of the code (defined as `xlstart_row <- 6`), here you can see that the starting row is 6 in the code, this would mean your column names would be in row 6, and your data would follow as seen in Fig 2. Your first channel should be your background DNA signal, your second channel should be your newly synthesized DNA channel, and your two final channels should be for protein.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | INFO | | | | | |
| 2 | Fri Jan 27 18:16:00 2012 | | | | | |
| 3 | Segment | Z Section | Time Point | Time (second | Angle | Segment Len |
| 4 | 1 | 1 | 1 | 0 | -18.869 | 9.1167 |
| 5 | | | | | | |
| 6 | Channel 452 | Channel 523 | Channel 594 | Channel 676 | | |
| 7 | 134 | 0 | 0 | 0 | | |
| 8 | 135 | 142 | 0 | 0 | | |
| 9 | 136 | 0 | 0 | 0 | | |
| 10 | 143 | 147 | 0 | 0 | | |
| 11 | 136 | 0 | 0 | 0 | | |
| 12 | 134 | 0 | 0 | 0 | | |

**Figure 2 Sample of Input Excel Sheet for R-ODD-BLOBS.** This highlights how you should format your data.

In addition to your input files, there will be some basic configurations required before running. This is mainly to ensure that R has proper libraries to run the code, as well as, ensure that the output file path exists on your computer. Similarly, to a map route app (eg. Google Maps), you need to give R a path direction for it to follow to deliver your results to the correct place. Further instructions are posted from lines 3- 20 at the top of the code.

## 3. ANALYSIS FILES

The output files are organized into a folder and placed into the path depending on the user (this can be adjusted in line 61 of the code), this contains:

1. Individual scatterplot of initial intensity data for each sheet
2. Individual violin plot comparing mean lengths of each column of each sheet in the workbook.
3. Binned mean lengths of each column in a bar graph
4. Individual sheet protein colocalization plots of both proteins (from columns 3 and 4)
5. A binned protein colocalization plot of both proteins *
6. An output Excel sheet (further explained in the next section- section 4)

* ”binned” is defined as the grouping of sheets. This would be useful when grouping replicates. The keywords for the bins are defined at the beginning of the script in lines 54-59, this would allow the program to search the sheet names and find matching sheet names that contain the keyword. For example, a keyword of “wt” would bin sheet names like “wt1” and “redo_wt2” and bin the results together. Another thing to note is that you have to ensure the keywords are unique to the bins, otherwise, double counting may occur.

One final thing to note is that currently with the output Excel file, you may receive an error code with the code from the Oct. 23, 2024 version of R-ODD-BLOBS. This is likely to due with the summary page. This a known error, that deletes the second column in the summary page. We have mitigated it by deleting important information by inserting a buffer column called “diff” in affect data so that the p-value won’t be deleted.

## 4.NAVIGATING THE EXCEL OUTPUT FILE

There’s three different types of sheets in the workbook:

1. The “Threshold parameters” sheet, this summarizes all the parameters you used when running the program, and the time and date. As well as the name of the file you ran. This acts as a housekeeping file to ensure you have a record of what was imputed in to the program and how it was ran.
2. The “Thres.Result. - *original file name*” shows the dataframes and how R-ODD-BLOBS processes you data during analysis This is shown with the red, yellow and blue boxes shown in Fig 3A. The red box shows the initial pass through determining if the value is above or below the threshold value defined by the user. TRUE means its above threshold and FALSE means it is below threshold. The yellow box indicates the data after smoothing, gaps of FALSEs inbetween two TRUEs that are less than the user defined smoothing values are

turned into TRUE values. The blue box shows the counts of lengths that the program will take for the mean lengths. The second half of analysis shown in Fig 3B, which is found through scrolling right of the same sheet. The green box shows the labelled tract regions and protein colocalization. The magenta box above the green box shows the number of pixel counts of each region. Specifically for forks this comprised of “UNR FORK”, “R FORK” and “FORK”. The orange box is shows all the lengths counted, this shows the frequency of certain lengths and is used for the violin plots in the main 2024 R-ODD-BLOBS publication. The light blue box shows the average of the lengths found in each column of the orange box.The brown box shows the plots for the individual sheets, this includes, the initial scatter plot, and protein colocalization.

3. The “Summary” sheet is comprised of the binned data, the violin plot comparing all sheets, and the statistical analysis results of the plots. In Fig 4A this shows the outputs of the “Summary” sheet. In the red box, this shows the violin plots comparing the mean lengths of each sheet per column, the bar graph of mean lengths per bin, and the protein colocalization for each bin. In the blue box this shows tukey test results (what is being compared and the p-value) for each violin plot comparison for each sheet, they print it in the same order as the columns, in this case the one from the most left is DNA, next is BrdU, rad51 and the most right is cdc45. The yellow box indicates F-test of the percentage of protein colocalization per region- one thing to note, is that this result was removed on the graph due to potential instability. The final tukey test is shown in Fig 4B, this is shown in the green box and this is shown underneath the results of the mean lengths sheet comparison tukey test shown previously in Fig 4A (the blue box).

A.

B.

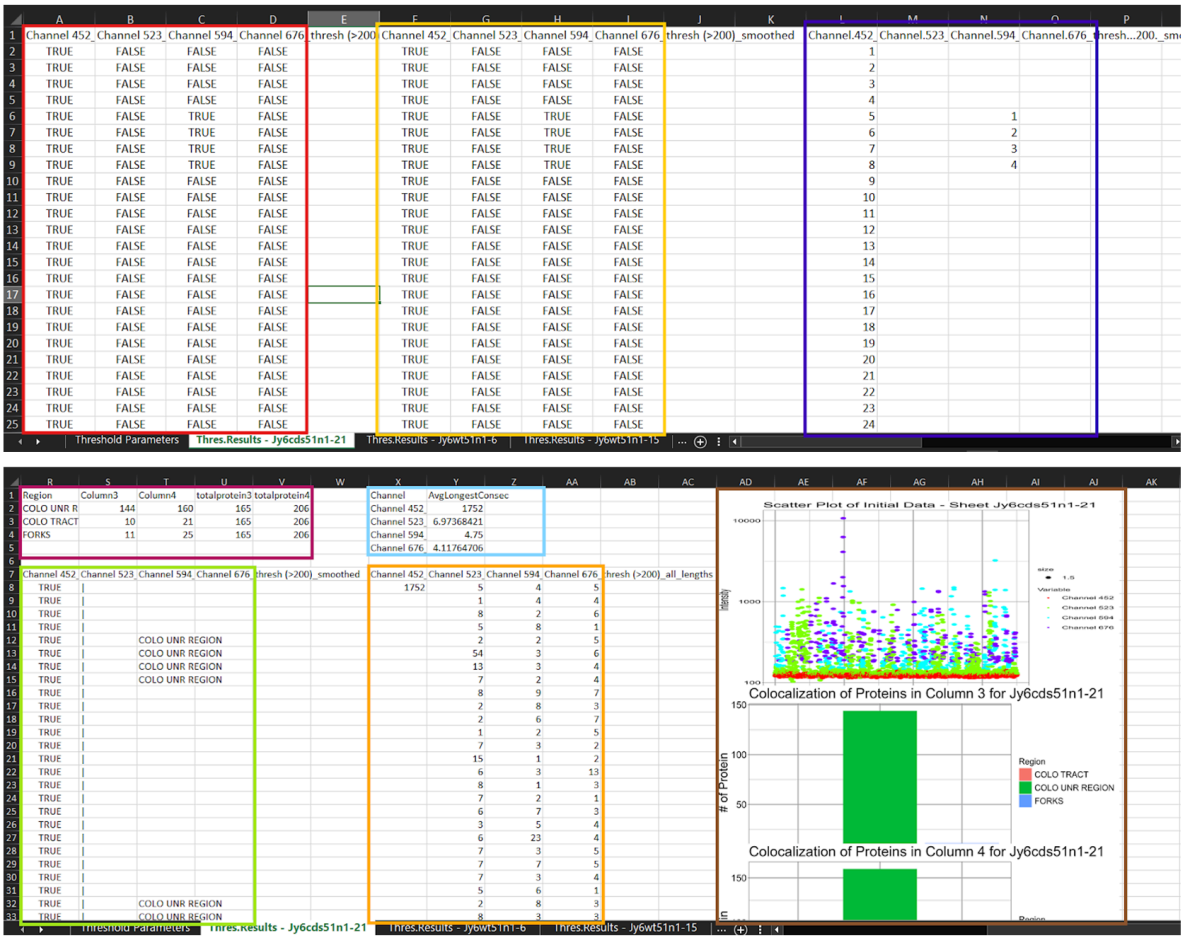
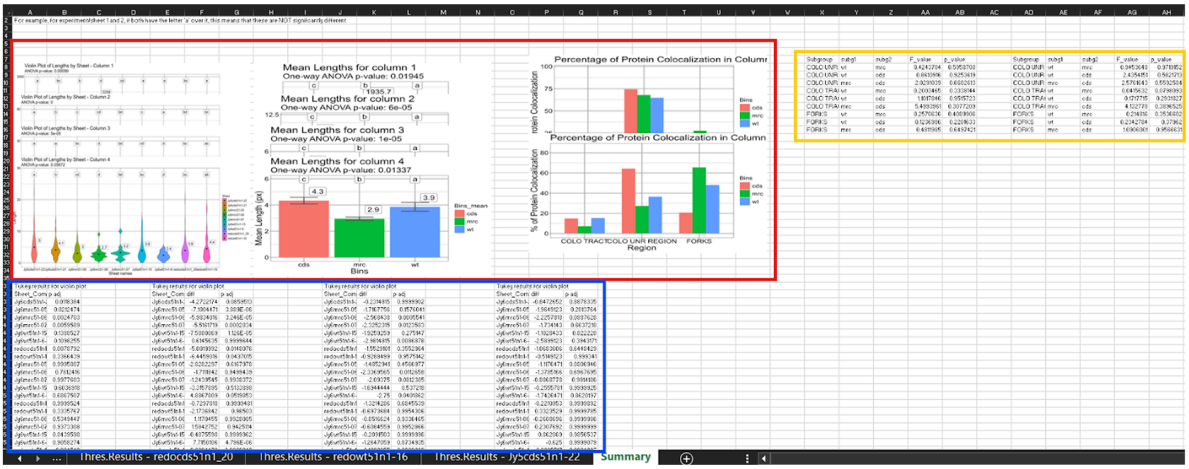
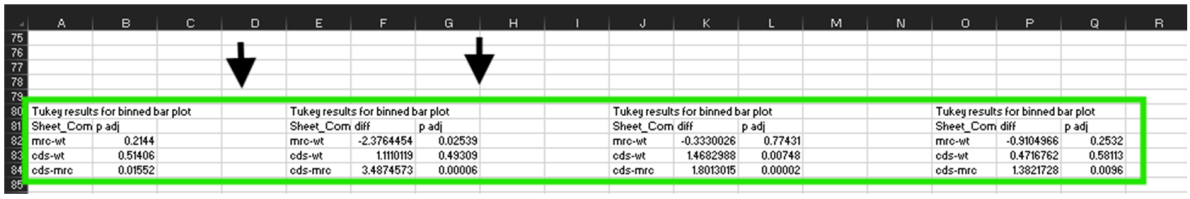



**Figure 3 The Thres.Results Page Navigation of Data Outputs.** The shows what is found on the thres.results sheet. This page is created for each sheet that is found in the original file. (A) Shows the data processing of the program, this includes the initial determination if it's above or below threshold, the smoothing results, and the counts of tracts. (B) Shows the results of labelling regions and colocalization, the counts of colocalization, the lengths found, the average lengths, and the plots created for from this individual file.

A.




B.



| Tukey results for binned bar plot | | Tukey results for binned bar plot | | | Tukey results for binned bar plot | | | Tukey results for binned bar plot | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sheet_Com | p adj | Sheet_Com | diff | p adj | Sheet_Com | diff | p adj | Sheet_Com | diff | p adj |
| mrc-wt | 0.2144 | mrc-wt | -2.3764454 | 0.02539 | mrc-wt | -0.3330026 | 0.77431 | mrc-wt | -0.9104966 | 0.2532 |
| cds-wt | 0.51406 | cds-wt | 1.1110119 | 0.49309 | cds-wt | 1.4682988 | 0.00748 | cds-wt | 0.4716762 | 0.58113 |
| cds-mrc | 0.01552 | cds-mrc | 3.4874573 | 0.00006 | cds-mrc | 1.8013015 | 0.00002 | cds-mrc | 1.3821728 | 0.0096 |

**Figure 4 The Summary Page Navigation of Data Outputs.** The shows what is found on the Summary sheet. This is comprised of the plots created that is from binning or comparing sheets. (A) Shows the plots created from binning mean lengths and protein colocalization, as well as the mean lengths compared between sheets. It also shows partial statistical test, the tukey test for the mean lengths compared between sheets, and the now exempt f-test for the percentage of protein colocalization per region. (B) Is the continuation, and is underneath the blue box of part A, this shows the tukey test of binned mean lengths.