Manual of SwimR

Jing Wang, Andrew Hardaway, Bing Zhang, Randy Blakely

April 24, 2017

1 Introduction

The nematode Caenhorhabditis elegans offers great power for the identification and characterization of genes that regulate many behaviors, from locomotion to learning and memory. To more precisely quantify these behaviors, analytical methods are required that provide dimensional analysis of subcomponents of behavior. Thus, we developed the package SwimR, to illustrate and quantify C. elegans Swip, which can reveal novel kinetic alterations in swimming produced by these manipulations that can be of use in dissecting the differential control of swimming behavior by converging signaling pathways.

2 Environment

The R version is at least 3.0.0, which can be downloaded in the website http://www.r-project.org/. The installation process of SwimR is as follows.

>source("http://bioconductor.org/biocLite.R")

>biocLite("SwimR")

3 Creation of frequency matrix

After building up the basic environment mentioned above, the users can install SwimR package and create frequency matrix and annotation file for the example files returned by Tracker program.

```
> library("SwimR")
> inputPath <- system.file("extdata", "trackerFiles", package="SwimR")
> outputPath <- getwd()
> freMat <- createFrequencyMatrix(inputPath, outputPath, method = "Extrema", Threshold = 0.6,
+ DeltaPeakDt = 1.6, MinFrameBtwnMax = 4, MinDelta = 2.5, longPeriod = 5, AvWindowSize = 10,
+ fps = 15, ZP_Length = 100, WindowSize = 30, MaxCompWin = 2, minTime = 0, maxTime = 600)

Processing...

File:  N2_M9_10-14-11_1-1.txt is in process...

File:  N2_M9_10-14-11_2-1.txt is in process...

File:  N2_M9_10-14-11_3-1.txt is in process...

File:  N2_M9_10-14-11_4-1.txt is in process...

File:  N2_M9_10-14-11_5-1.txt is in process...

File:  dat-1_(ok157)_1-10-11_1-1.txt is in process...
```

```
File:  dat-1_(ok157)_1-10-11_2-1.txt is in process...

File:  dat-1_(ok157)_1-11-11_1-1.txt is in process...

File:  dat-1_(ok157)_1-11-11_4-1.txt is in process...

File:  dat-1_(ok157)_1-12-11_3-1.txt is in process...
Processing completed!

Please see the frequency matrix
and detailed information for each animal in the outputPath directory!
```

3.1 Input

Here is a description of all the arguments needed to get the frequency matrix and annotation file:

1. *inputPath* is a directory which contains the files returned by the Tracker program. If you are using a common directory as described above, you may type in inputPath <-"folder name" where folder name is a subfolder containing the tracker files to be analyzed. Because annotation file is generated by extracting the genotype information from tracker file names, the user should use dashes to denote genotypes and separate the date in the file name like Genotype_Drug(if used)_Dose(if used)_Date_#. The following are some examples of tracker file names: N4_AMPH_100uM_2-11-14_1 (genotype is N4), dat-1ok157_IMI_10uM_12-12-12_7 (genotype is dat-1ok157), cat-2e1112dat-1ok157_2-20-09_2 (genotype is cat-2e1112dat-1ok157) and dat-1ok157dop-3vs106_5-17-06_4 (genotype is dat-1ok157dop-3vs106).

2. *outputPath* is a directory which saves the plots and files returned by the function.

3. The function provides four different counting methods: "FFT" (Fast Fourier Transform), "Extrema", "PeakDet" (peak delta) and "RT+GP" (Get Peaks plus Racetrack Filter) and the users can select one of them to output the corresponding frequency matrix. The default outputted *method* is "Extrema".

4. *Threshold* is the amount of degrees (in radians) required to count at as thrash and the default is 0.6.

5. *DeltaPeakDt* is the threshold for "peak delta" algorithm, similar to *Threshold* and the default is 1.6.

6. *MinFrameBtwnMax* is the minimum number of frames between maxima and the default is 4.

7. *MinDelta* is similar to *DeltaPeakDet* and the default is 2.5.

8. *longPeriod* is the longest period cycle which is not zerod and the default is 5.

9. *AvWindowSize* is the length of the average window in seconds and the default is 10 seconds.

10. *fps* is the frame per second and the default is 15.

11. *ZP_Length* is the Zero-padding length and the default is 100.

12. *WindowSize* is the size of window for computing the Fast Fourier Transform and the default is 30.

13. *MaxCompWin* is the window size on deciding if the current angle is a maxima and the default is 2.

14. *minTime* is the minimum threshold of time points for the following analysis and the default is 0 second.

15. *maxTime* is the maximum threshold of time points for the following analysis and the default is 600 seconds.

3.2 Output

The createFrequencyMatrix function outputs six files:

1. *outputDescription_createFrequencyMatrix.html* contains a summary of all output files. 2. *XFig.jpg* is the image of scatter plot of one animal plotted as "Frequency vs Time(min)" with all four counting methods overlaid (see Figure 1 for file dat-1_(ok157)_1-10-11_2-1.txt). "X" of "XFig.jpg" represents the input file names (see Figure1 as an example).

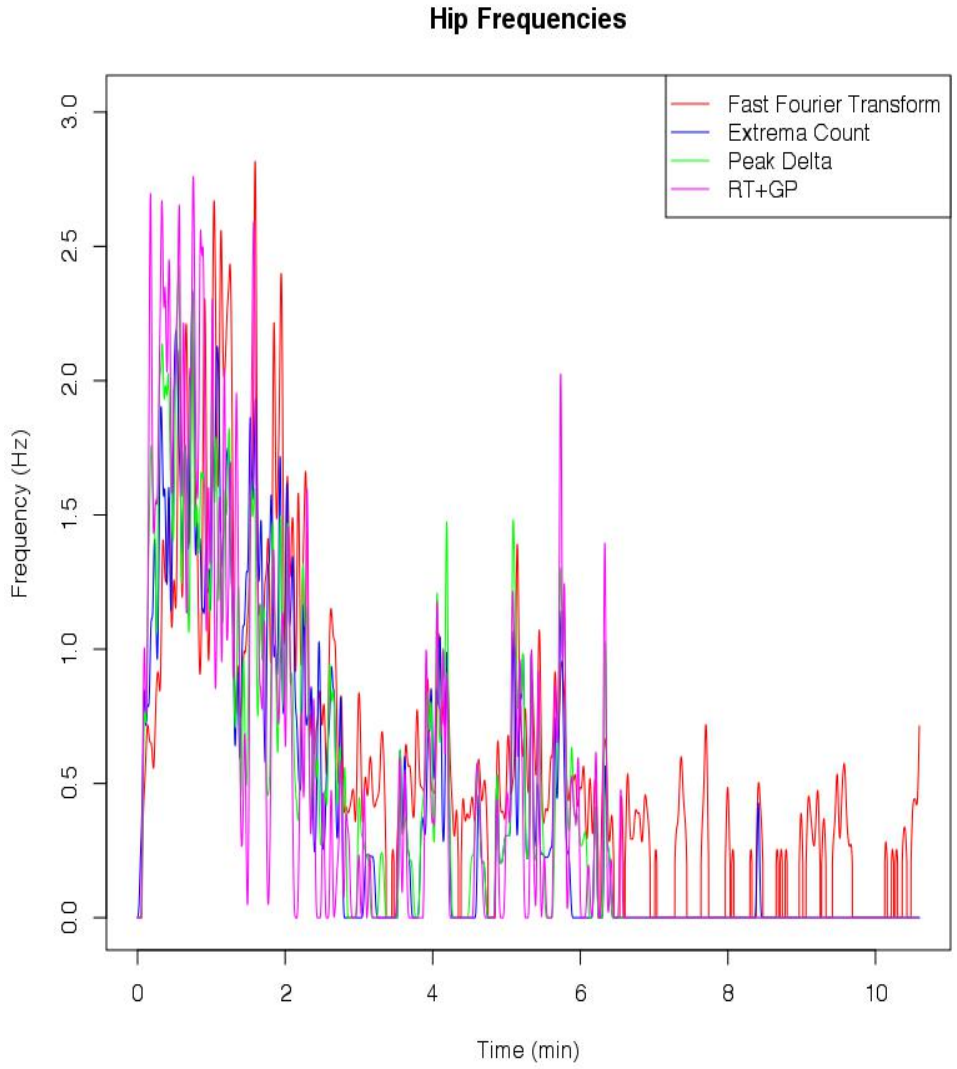

Figure 1: *A scatter plot of three methods for file dat-1_(ok157)_1-10-11_2-1.txt*

3. *XFigSub.jpg* is the same as *XFig.jpg* except counting methods are broken up into four different plots. This is very helpful in checking through a video to make sure that Tracker tracked the worm properly. Bad contrast can be a problem with Tracker missing the worm and these files help to identify troublesome videos for retracking or discarding (see Figure2 as an example).

4. *XFreq.csv* is the CSV file of raw data organized by column, where column one represents frequency as counted by FFT, column two represents frequency calculated by Extrema, column three represents frequency calculated by PeakDt , column four represents frequency as counted by RT+GP and column five represents time in seconds.

5. *frequencyMatrix.txt* is a TXT file which combines the analysis results of the frequency of worm thrashing over time for all Tracker files in the *inputPath*.

6. *annotationfile.txt* is a TXT file which contains all genotype information extracted from file names of all Tracker files in the *inputPath*.

The createFrequencyMatrix function also returns a list object which contains all information of output files.

4 SwimR

SwimR can analyze and visualize worm swimming data returned by the above function. It places a particular emphasis on identifying paralysis and quantifying the kinetic elements of paralysis during swimming.

```
> expfile <- system.file("extdata", "SwimExample", "SwimR_Matrix.txt", package="SwimR")
> annfile <- system.file("extdata", "SwimExample", "SwimR_anno.txt", package="SwimR")
> projectname <- "SwimR"
> outputPath <- getwd()
> result <- SwimR(expfile, annfile, projectname, outputPath, color = "red/green",
+ data.collection.interval = 0.067, window.size = 150, mads = 4.4478,
+ quantile = 0.95, interval = 20, degree = 0.2, paralysis.interval = 20,
+ paralysis.degree = 0.2, rev.degree = 0.5)

Processing...
Processing completed!
Please see the detailed information in the outputPath directory!
```

4.1 Input

Here is a description of all the arguments for SwimR:

1. *expfile* is the path of the frequency matrix returned by the *createFrequencyMatrix* function.

2. *annfile* is the path of annotation file returned by the *createFrequencyMatrix* function.

3. *projectname* is the name of the project.

4. *outputPath* is a directory which saves the plots and files returned by the function.

5. The function provides four colors to plot the heat map figure:"red/green", "red/blue", "yellow/blue" and "white/black". The default *color* is "red/green".

6. *data.collection.interval* is the time interval between two points and the default is 0.067.

7. *window.size* is the size of the window for the running average that is calculated to smooth the data. The default is 150.

8. *mads* is the number of median absolute deviations that a given animal must deviate from the median sum of frequencies to be called an outlier. The default is 4.4478.

9. *quantile* is the proportion of data points that are used in calculating the color scheme for the heat map and the default is 0.95.

10. *interval* is the minimum time that a given animal must lie below a threshold to be called a paralyzed worm for the first calculation and the default is 20.

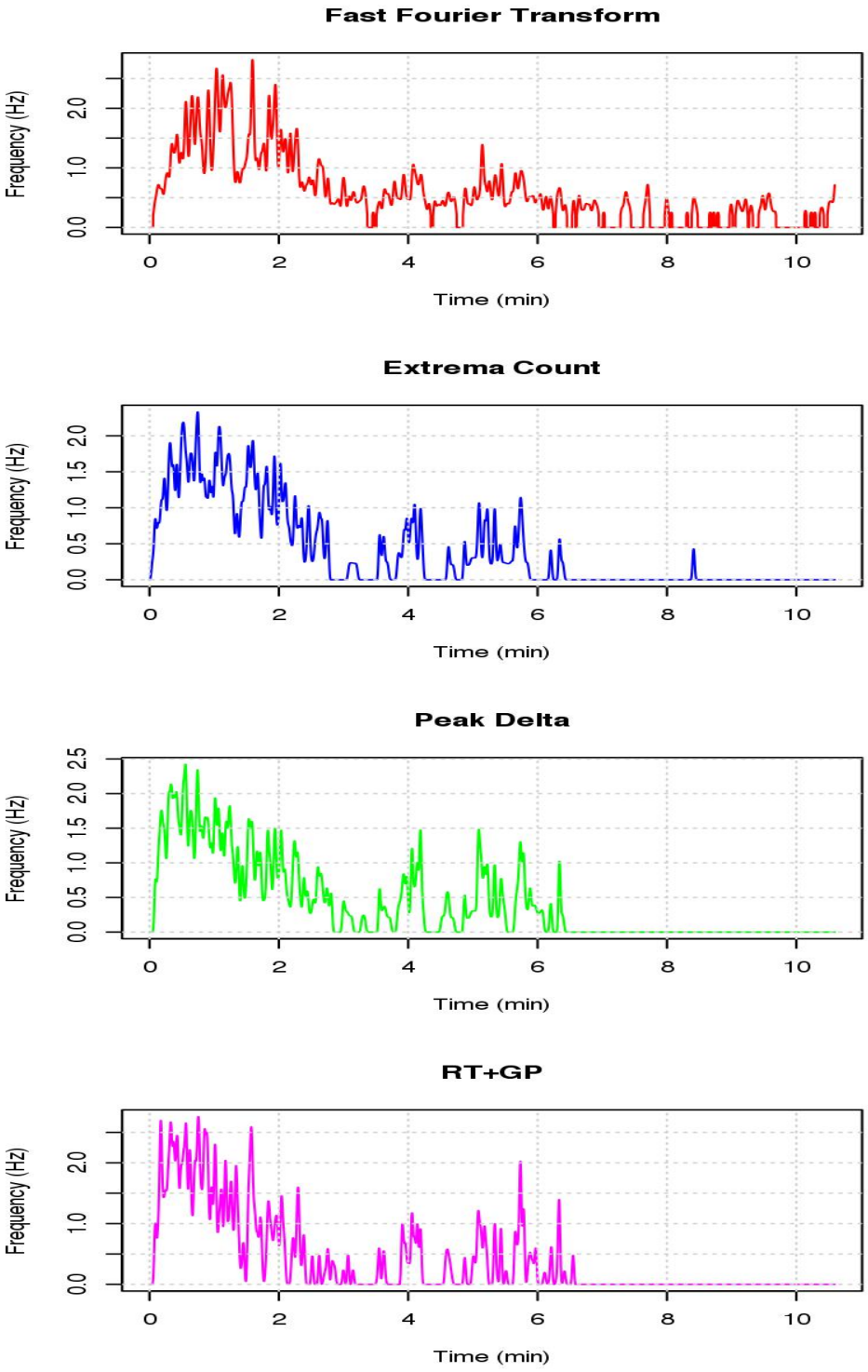

Figure 2: *Three separate scatter plots of four methods for file dat-1_(ok157)_1-10-11_2-1.txt*

11. *degree* is the paralytic degree for the first calculation and the default is 0.2.

12. *paralysis.interval* is the same as *interval* but for the second calculation and the default is 20.

13. *paralysis.degree* is the paralytic degree for the second calculation and the default is 0.2.

14. *rev.degree* is the threshold that an animal must cross to be called a revertant and the default is 0.5.

4.2 Output

The SwimR function outputs 13 files: 1. *output_SwimR.html* contains a summary of all output files.

2. *P_sample_t_half.txt* is the TXT file of each animal and their corresponding latency to paralyze. For non-paralyzers, N/A will be listed. "P" of "P_sample_t_half.txt" is the *projectname* inputted by users.

3. The columns of *P_group_data.txt* is defined as follow. "freq_max_mean": Mean maximal swimming frequency; "freq_max_sd": Standard deviation of Mean maximal swimming frequency; "freq_min_mean": Mean minimum swimming frequency; "freq_min_sd": Standard deviation of Mean minimum swimming frequency; "freq_range_mean": Mean range between maximum and minimum; "freq_range_sd": Standard deviation of Mean range between maximum and minimum; "paralytic_count": The number of paralyzed animals amongst the samples; "non-paralytic_count": The number of non-paralyzed animals amongst the samples; "t_half_-mean": Mean latency to cross the paralytic threshold set by the users (default is 20% of Frequency range) and stay below it for the interval specified interval (default is 20 seconds); "t_half_sd": Standard deviation of *t_half_mean*; "t_p_start_mean": The mean time point (in seconds) at which each animal crosses a frequency that is min+paralytic threshold and stays below that threshold for the paralytic interval; "t_p_start_sd": Standard deviation of *t_p_start_mean*; "t_p2end_mean": The average range of time after paralysis; "t_p2end_-sd": Standard deviation of *t_p2end_mean*; "rev_count": The number of revertants amongst the samples as defined by the threshold set by the user (default is animals have to recross 50% of their frequency range for any length of time; "rev_percent": The number of revertants; "rev_frequency_mean": The number of reversion events; "t_p2r_mean": Mean time between 1st reversion and *t_p_start_mean*; "t_p2r_sd": Standard deviation of *t_p2r_mean*; "t_r_total_mean": Mean of total time spent in reversion for all revertants; "t_r_total_sd": Standard deviation of *t_r_total_mean*; "t_r_average_mean": Mean length of an individual reversion event; "t_r_average_sd": Standard deviation of *t_r_average_mean*; "r_amp_mean": Mean of total amplitude of reversion for all revertants, where amplitude is defined by the area beyond the reversion threshold set by user (default is 50% Freq range) during reversion, calculated by summing up discrete values for each measurement (same unit as frequency); "r_amp_sd": Standard deviation of *r_amp_mean*.

4. *P_heatmap_withingroup_ordered_globalcentering.jpg* is a JPEG of the heat map of all samples included in the data matrix after outlier exclusion, smoothing, ordering based on the latency to paralyse, and centering the color based on the quantile percent that can be set by the user in the parameters section of SwimR (see Figure 3 as an example).

5. *P_heatmap_withingroup_ordered.txt* is a TXT file of the raw data used to plot the heat map after outlier exclusion, smoothing and ordering based on the latency to paralyze.

6. After exclusion and smoothing, *P_histogram.nooutliers.smoothed.jpg* is a JPEG file of all frequency data points broken up into increasing 0.1 Hz bins and then plotted as the fraction of the total as a histogram (see Figure 4 as an example).

7. *P_histogram.nooutliers.smoothed.data.G.txt* is a TXT file of the raw data used to plot the histogram. "G" in the "P_histogram.nooutliers.smoothed.data.G.txt" is the genotype in the annotation file.

8. *P_individual_data.txt* is a TXT file that returns reversion information for individual animals. The definitions are identical to the P_group_data.txt file, but "R_count" is the number of reversion events for that animal. If there is no paralyzed animal, this file will not be outputted.

9. *P_individual_data1.txt* is a TXT file. For animals that paralyzed: The R_instances row tells the user exactly when the animal reverted. For animals that did not revert, N/A will be listed. If there is no paralyzed animal, this file will not be outputted.

10. *P_intermediate.results.txt* describes some key features of your samples after running SwimR, and is a great way to get a quick look at the incidence of paralysis amongst your samples. At the top of the file, it lists the parameters used in the subsequent calculations. Below that, it lists the summed frequency values for each of the animals included in the sample. And then the p value of the bimodal test for each genotype

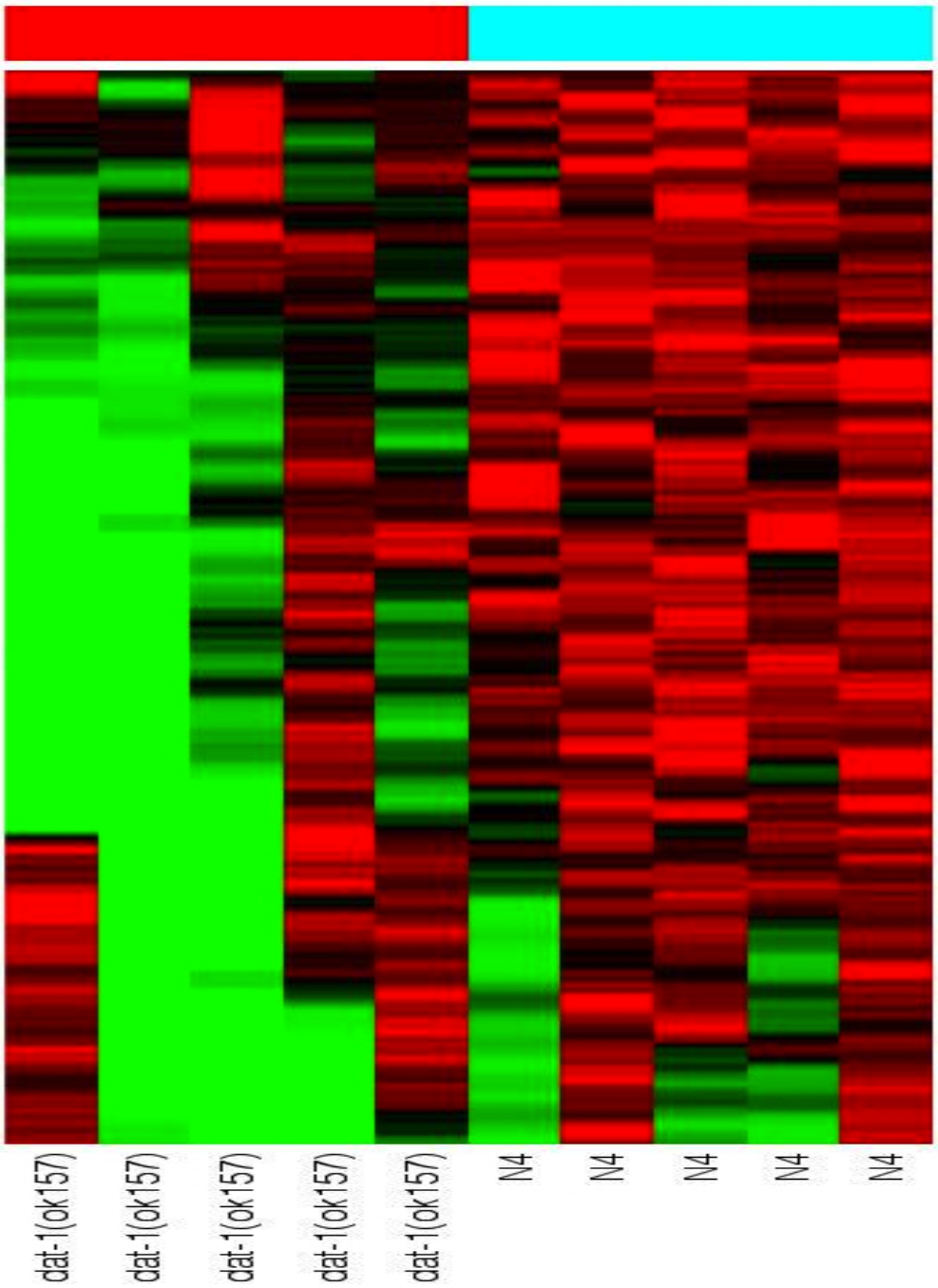

Figure 3: *Heat map of all samples after outlier exclusion, smoothing, ordering based on the latency to paralyse*

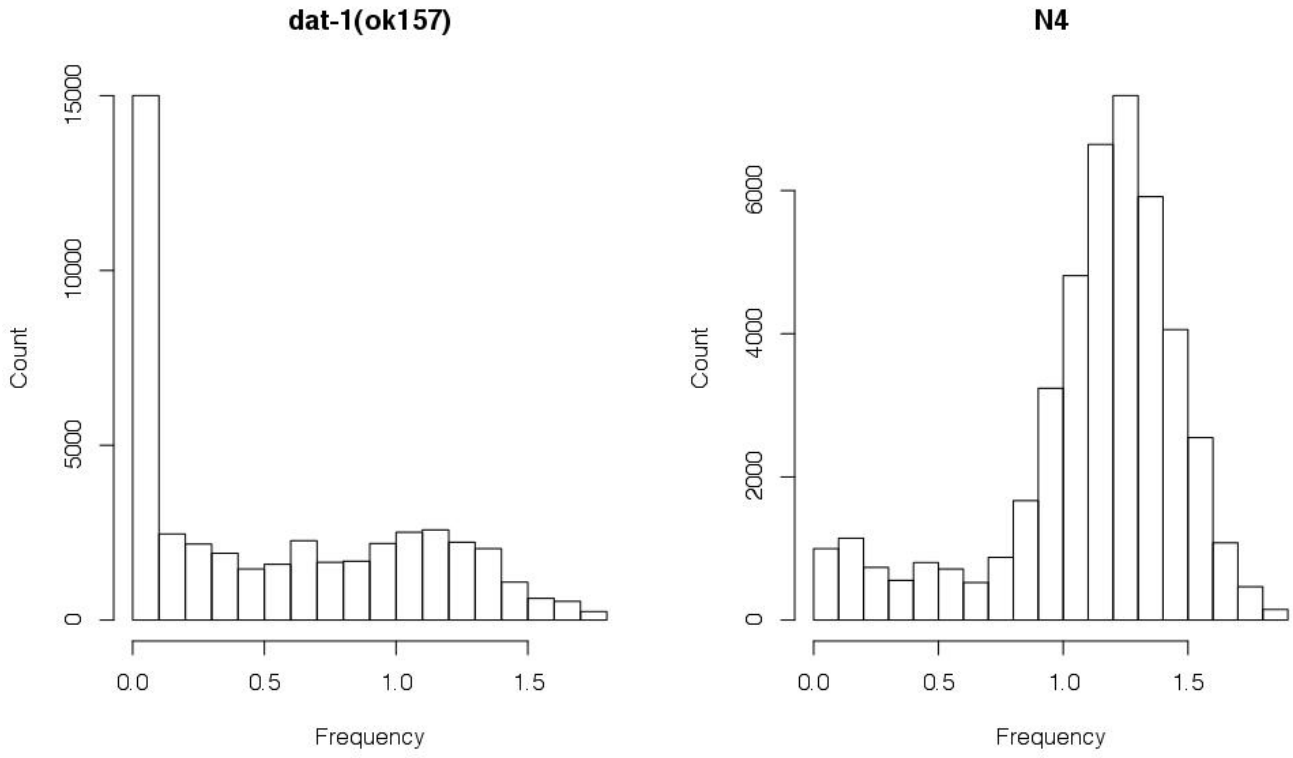

Figure 4: *Histogram of all frequency data points broken up into increasing 0.1 Hz bins*

was listed. Below that, it lists each of the animals included and excluded after outlier detection. After that, it lists which animals were considered paralyzed and which not. For paralyzed animals, it then lists which of them were called revertants.

11. *P_scatter.jpg* is a JPEG image of the average frequency plotted against time after outlier exclusion, but no smoothing (see Figure 5 as an example).

12. *P_nooutliers_smoothed_scatter.jpg* is a JPEG image of the average frequency plotted against time after outlier exclusion and smoothing (see Figure 6 as an example).

13. *P_nooutliers_smoothed_scatter_data.txt* is a TXT file of the raw data used to plot the smoothed scatter plot. Column one is time, Column two is average frequency and Column three is standard deviation.

The SwimR function also returns a list object which contains all information of output files.

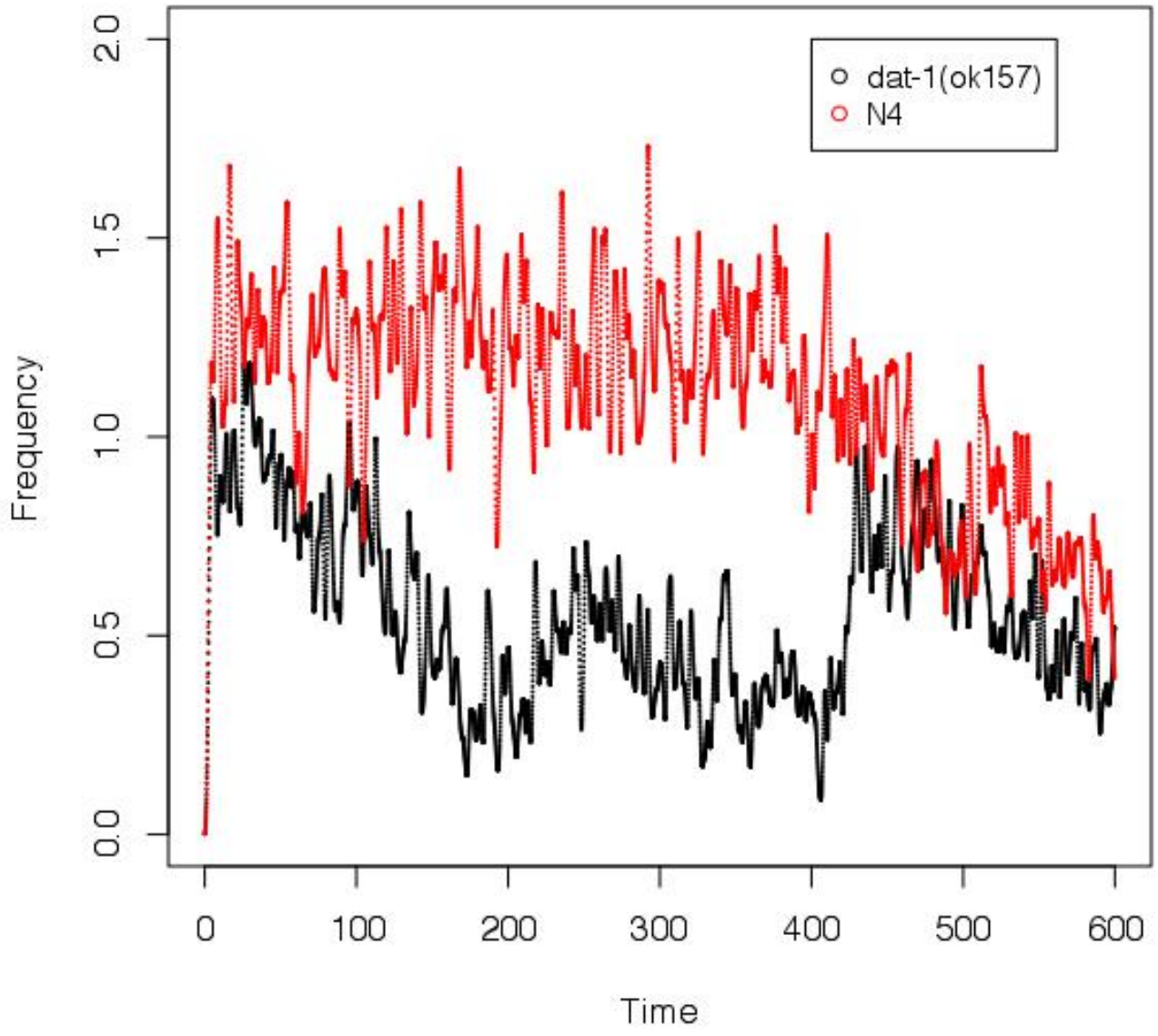

Figure 5: *Scatter plot of the average frequency plotted against time after outlier exclusion, but w/o smoothing*

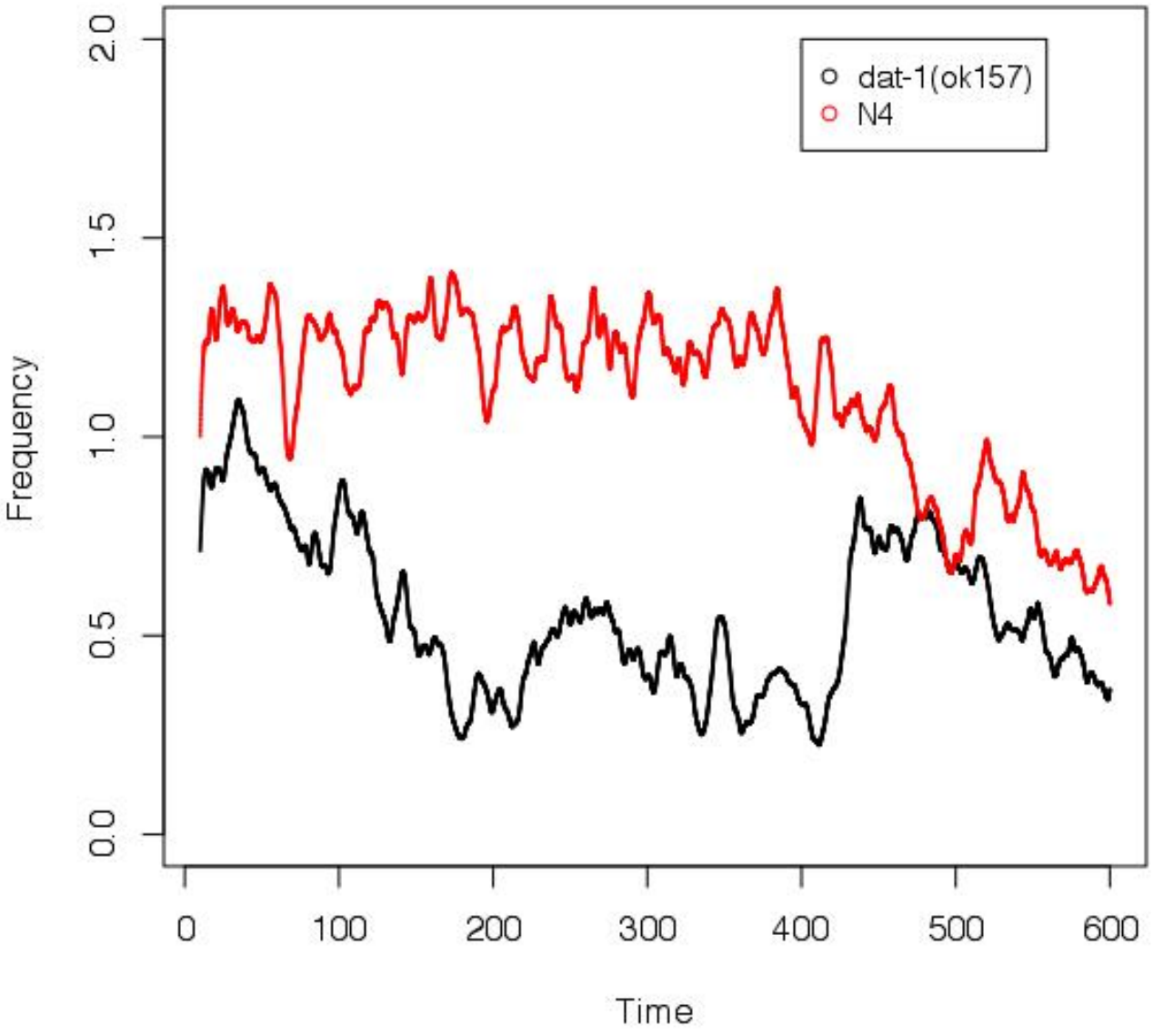

Figure 6: *Scatter plot of the average frequency plotted against time after outlier exclusion and smoothing*