RSCABS:
An R Package for Performing the Rao-Scott Adjusted Cochran-Armitage Trend Test by Slices

Joe Swintek
Badger Technical Services

April 26, 2017

Introduction

RSCABS[3] (Rao-Scott adjusted Cochran-Armitage trend test By Slices) is a modification to the Rao-Scott[5] adjusted Cochran-Armitage trend test[1, 2] that allows for testing at each individual severity score often seen in histopathological data. The test was originally developed and implemented in SAS™ by John Green([1]) as part of the Medaka Extended One Generation Reproduction Test[4] (MEOGRT) with the purpose of testing the effects of endocrine disruptors on histopathological endpoints. The RSCABS package allows for easy use of the RSCABS analysis with the capability of using both command line and graphical user interface (GUI) driven operations.

The RSCABS analysis is specificity designed to analyze histopathological results from standard toxicology experiments, for example the MEOGRT. These experiments typically have a number of organisms (e.g. medaka) contained in various amounts in different holding apparatuses (e.g. fish tanks). Multiples of each holding apparatuses are exposed to a either an experimental control (plain water) or one of several different concentrations of a chemical. At the end of the experiment several histopathological endpoints are evaluated on every organism. A severity score is assigned to every histopathological endpoint, which is typically an integer ranging from 0 (no effect) to 5 (an incredibly severe effect). A greater severity score indicates a more severe effect but the differences in severity scores are not consistent. For example the difference in severity of a score of 2 to a score of 1 is not the same as the change in severity moving from a score of 5 to of 4. Thus, even though severity scores have an order to them, they must be treated as categorical variables.

To develop an analysis of severity score data, several steps must be completed. The first step is to develop a basic test which tests a set of organisms for an increase in the presences (score > 0) or absence (score $= 0$) of an effect with an increase in the dose concentration of the treatments. The Cochran-Armitage (CA) trend [1,2] test was developed to test for this increase. However, it is common for group of organisms to be contained within the same holding apparatus. This could lead to organisms within the same apparatus having results that more closely resemble each other, then organisms in other apparatuses. The Rao-Scott (RS) adjustment controls for this by calculating an adjusting to the CA test statistic from correlation of organisms within each apparatuses. The by slices (BS) portion of the test allows for testing at each severity score instead of just presences or absence. By slices works by splitting the the severity scores associated with an endpoint into two groups based on the severity score being tested. One group contains all severity score less then the target severity score and the other group contains all severity scores equal to or greater then the target severity score. The RSCA test statistic is calculated based on these two groups instead of just a presences (score ≥ 1) and absence (score < 0). For example testing at a severity score of 2 would involve splitting the data into a group of severity scores < 2 and a group of all severity scores ≥ 2. RSCABS is a step down anylsis so, if the test statistic is calculated to be significant (p-value ≥ 0.05) then highest treatment level is removed from the analysis and the RSCA test statistic is recalculated. The process is repeated until the test statistic is not significant or there are no treatment levels left. This step-down procedure is repeated for each unique score assigned to an endpoint. Further details an examples of RSCABS can be found in [3].

RSCABS GUI

Starting Histopath

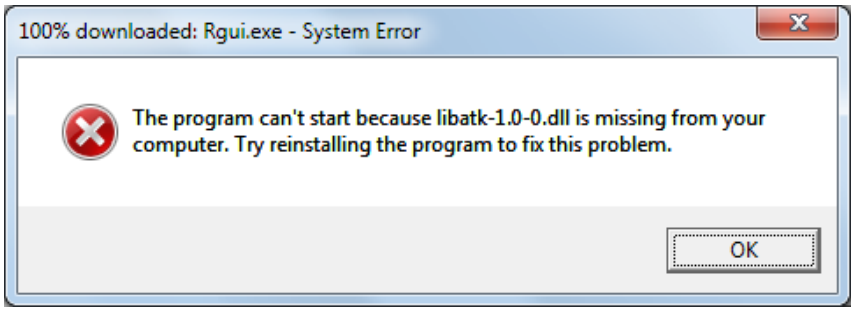
For ease of operation the RSCABS package has two ways of performing the RSCABS analysis, a command line function runRSCABS or a GUI front-end called Histopath. To call the GUI simply type the following into the console:

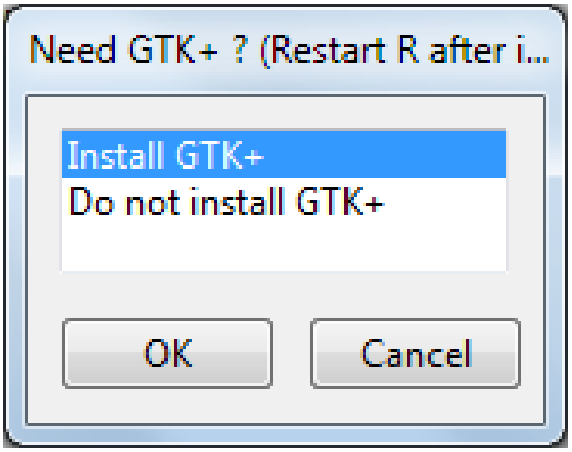
[1] DuPont Applied Statistics Group

R Input

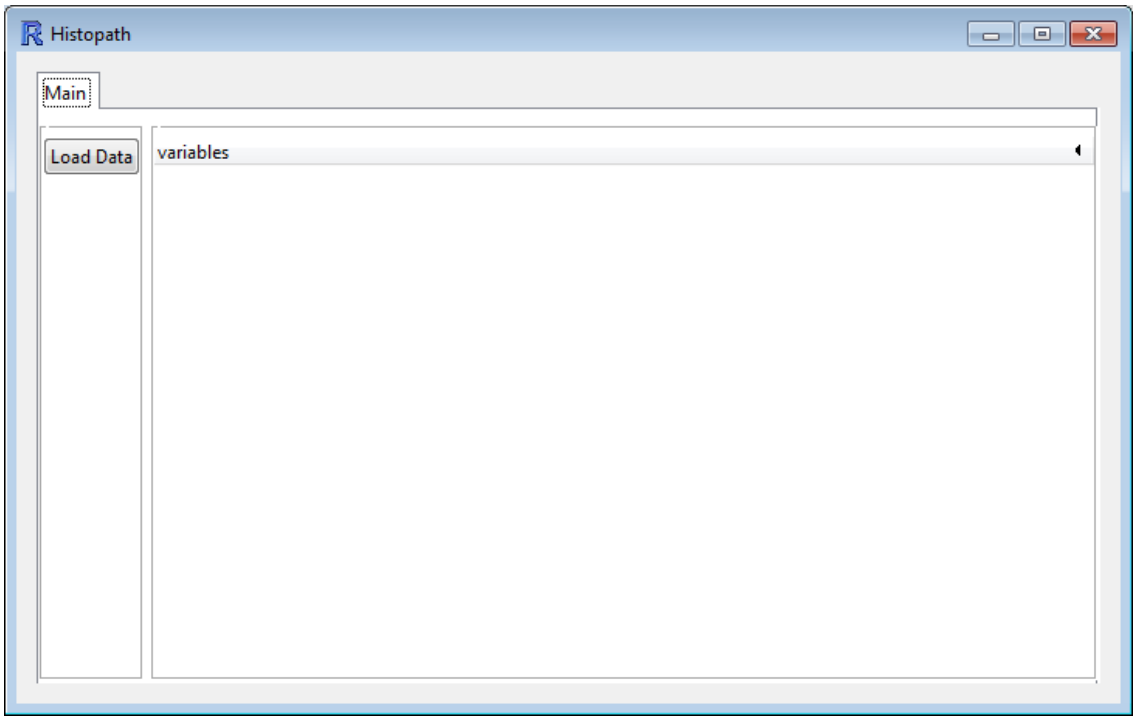
```
> install.packages(RSCABS)        #Install RSCABS from CRAN
> library(RSCABS)                       #Load the RSCABS library
> Histopath()                               #Calls the GUI for RSCABS
```

RSCABS is depended on the RGtk2 package which uses gtk+ and can be found at http://www.gtk.org/. If gtk+ is not installed, using "Histopath()" will cause R to produce an error message and then prompt for the installation of gtk+.



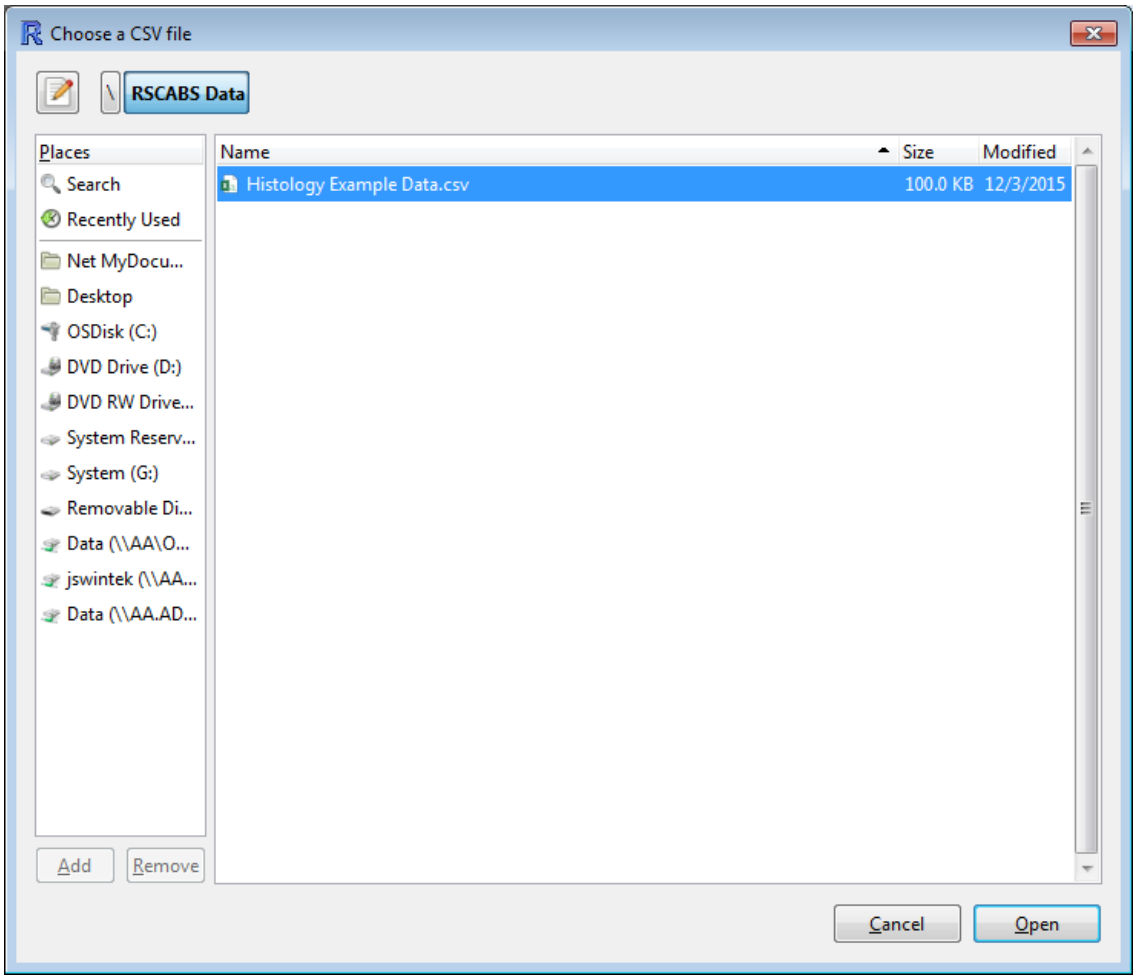


The instalation of gtk+ can be accomplished within R by selecting the "Install GTK+" option. After gtk+ is finished installing, R will need to be restarted before continuing. After restarting R using the "Histopath()" command should produce this window:

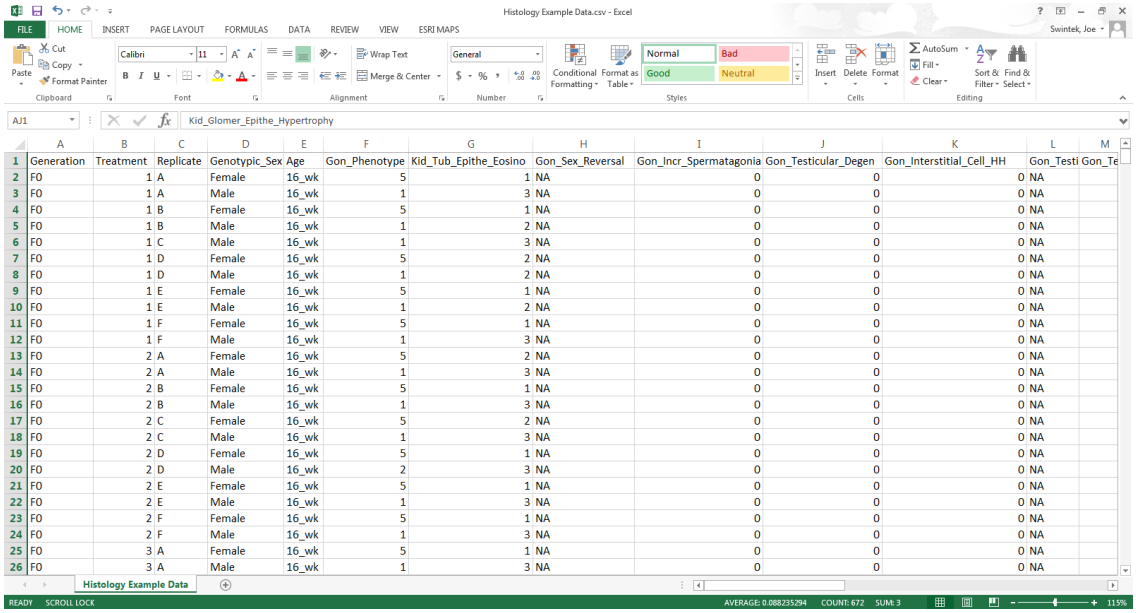


Loading Data

After Histopath has been successfully called, data can be loaded into the program by clicking on the [Load Data] button. This will open a dialogue box where the data file can be selected. Due note, this browser window cannot navigate through short cuts and will give an error if tried.



Below is a screen shot of the example histology data in comma separated value (csv) provided with in the RSCABS package.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Generation	Treatment	Replicate	Genotypic_Sex	Age	Gon_Phenotype	Kid_Tub_Epithe_Eosino	Gon_Sex_Reversal	Gon_Incr_Spermatagonia	Gon_Testicular_Degen	Gon_Interstitial_Cell_HH	Gon_Testi	Gon_Te
2	F0	1	A	Female	16_wk	5	1	NA	0	0	0	NA	
3	F0	1	A	Male	16_wk	1	3	NA	0	0	0	NA	
4	F0	1	B	Female	16_wk	5	1	NA	0	0	0	NA	
5	F0	1	B	Male	16_wk	1	2	NA	0	0	0	NA	
6	F0	1	C	Male	16_wk	1	3	NA	0	0	0	NA	
7	F0	1	D	Female	16_wk	5	2	NA	0	0	0	NA	
8	F0	1	D	Male	16_wk	1	2	NA	0	0	0	NA	
9	F0	1	E	Female	16_wk	5	1	NA	0	0	0	NA	
10	F0	1	E	Male	16_wk	1	2	NA	0	0	0	NA	
11	F0	1	F	Female	16_wk	5	1	NA	0	0	0	NA	
12	F0	1	F	Male	16_wk	1	3	NA	0	0	0	NA	
13	F0	2	A	Female	16_wk	5	2	NA	0	0	0	NA	
14	F0	2	A	Male	16_wk	1	3	NA	0	0	0	NA	
15	F0	2	B	Female	16_wk	5	1	NA	0	0	0	NA	
16	F0	2	B	Male	16_wk	1	3	NA	0	0	0	NA	
17	F0	2	C	Female	16_wk	5	2	NA	0	0	0	NA	
18	F0	2	C	Male	16_wk	1	3	NA	0	0	0	NA	
19	F0	2	D	Female	16_wk	5	1	NA	0	0	0	NA	
20	F0	2	D	Male	16_wk	2	3	NA	0	0	0	NA	
21	F0	2	E	Female	16_wk	5	1	NA	0	0	0	NA	
22	F0	2	E	Male	16_wk	1	3	NA	0	0	0	NA	
23	F0	2	F	Female	16_wk	5	1	NA	0	0	0	NA	
24	F0	2	F	Male	16_wk	1	3	NA	0	0	0	NA	
25	F0	3	A	Female	16_wk	5	1	NA	0	0	0	NA	
26	F0	3	A	Male	16_wk	1	3	NA	0	0	0	NA	

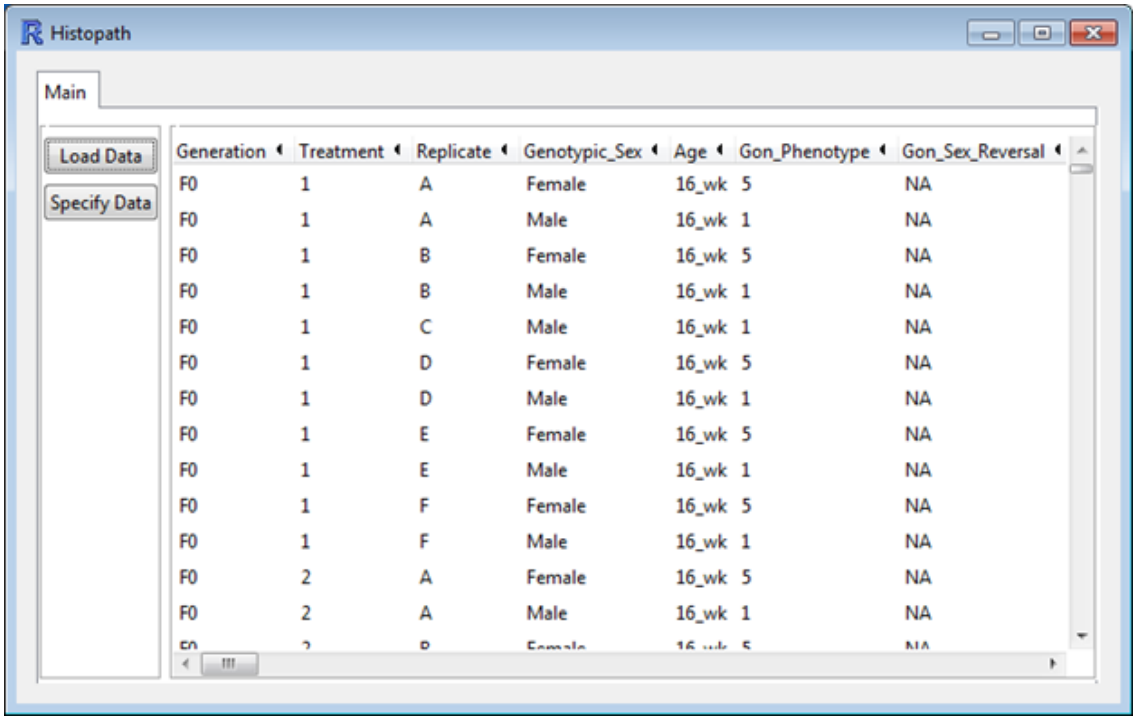
Data sets must be in a csv format. Each column indicates contains information used to identify a specimen or the severity scores of an endpoint while each row is a different specimen, which in this example is a fish. The exception is the first row which is the header row and contains the names of the fields. As with any csv file imported into R, missing data are indicated by either NA or a blank cell. Character entries (names etc...) may contain spaces, underscores ("_"), or dots (".") to separate words, however R will convert all spaces to dots upon importing the file.

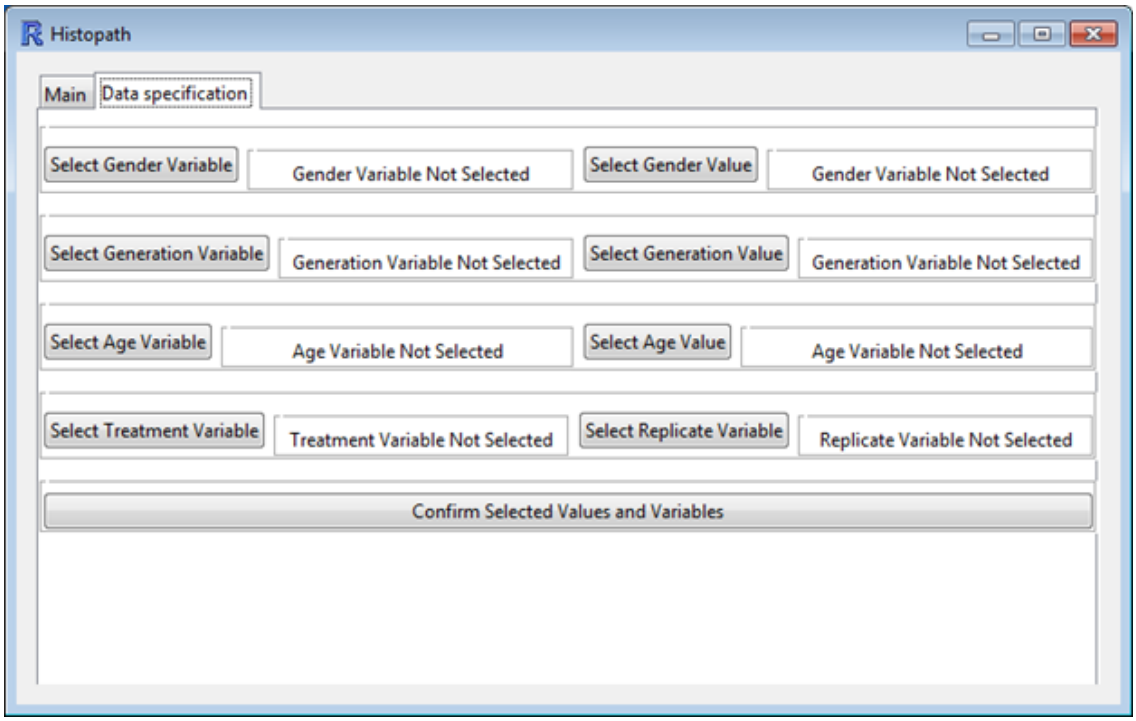
Histopath separates the data into three types of fields; identification, pathology endpoints, and ignored fields. The identification fields include a field for each of: **gender**, **generation**, **age**, **treatment**, and **replicate**. Of the identification fields **treatment** must be included in the data set while **gender**, **generation**, **age**, and **replicate** are not mandatory. However, if the replicate field is not included in the data set, Histopath will assume each specimen is independent and not apply the Rao-Scott adjustment. This may increase the rate of false positives if this independence assumption does not hold. If an identification field is used then every entry that field must have a value otherwise the row associated with that field will be removed when RSCABS is ran. The **treatment** field must only contain integers, with 0 indicating controls and each larger number indicating a larger dose. The other identification fields do not have this restriction and can contain any combination of number and letters.

Every field that is not an identification field is potentially a pathology endpoint. These fields may contain any entry, however, any entry that is not a **0** or a **positive number** is treated as missing data and is not included in the analysis. Due to how RSCABS groups severity scores non-integer numbers are treated as the next smallest integer, e.g. both 1.1 and 1.9 are treated as 1. Columns that are not identification fields, do not contain any number larger then **0**, or contain numbers larger then **20** will be ignored by Histopath.

Specifying Identification Fields

After the data set is loaded into Histopath, a [Specify Data] button will appear. Clicking the button will create a Data Specification tab.

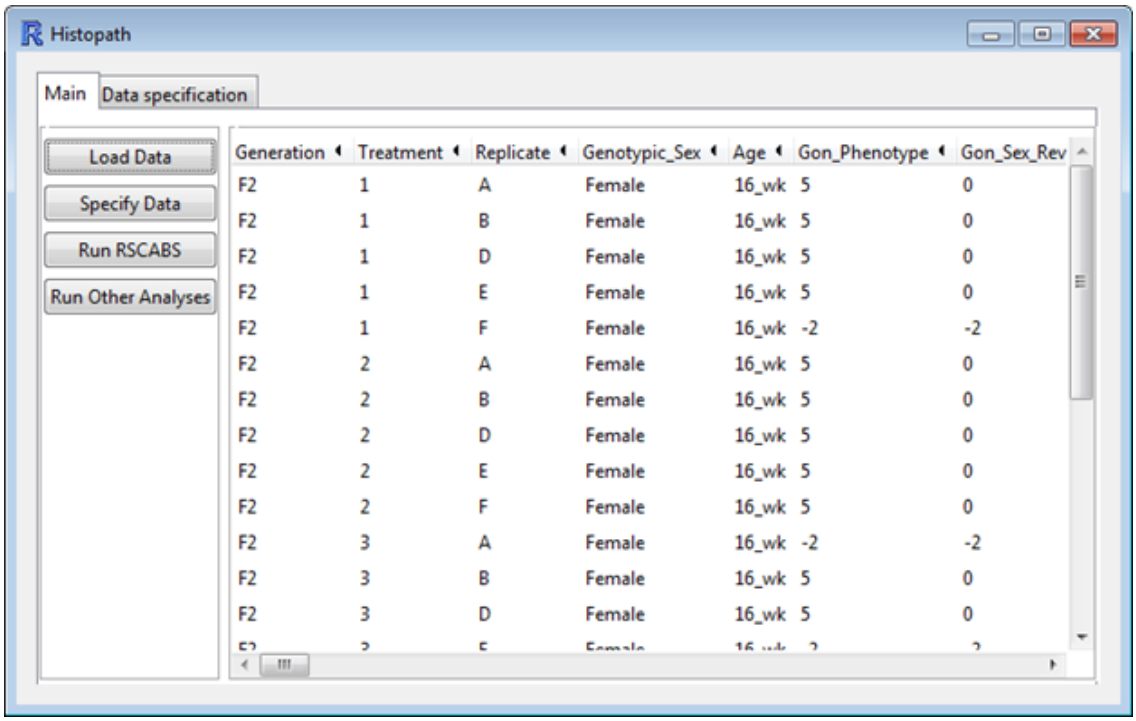




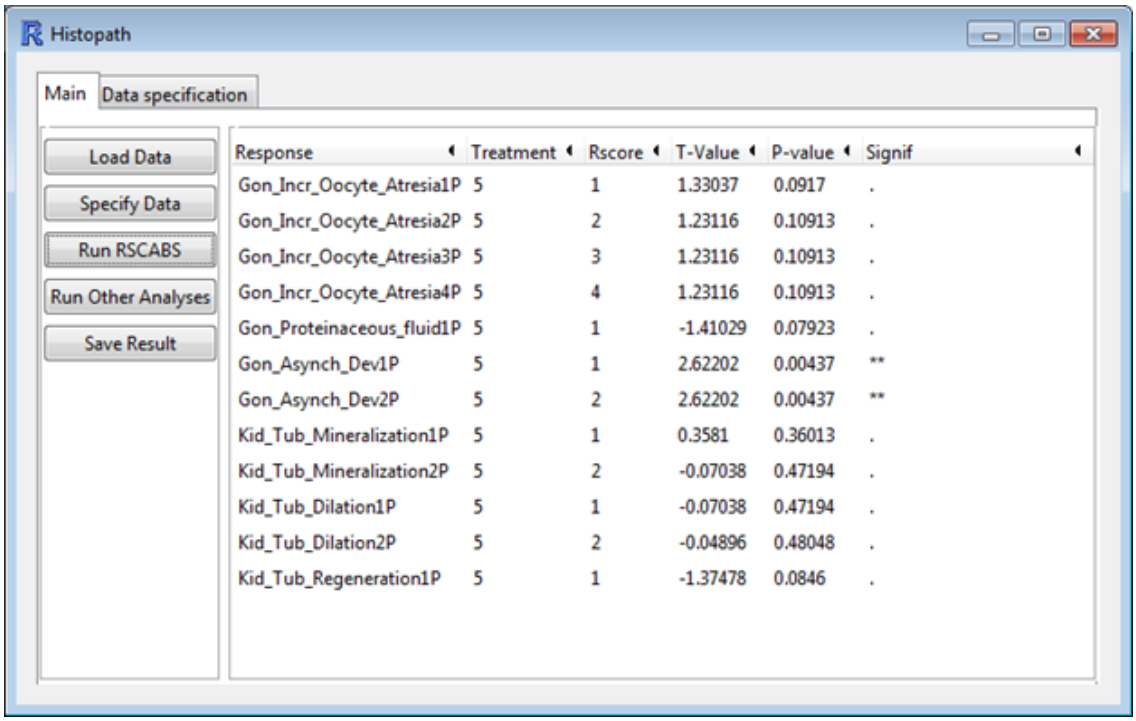
The form above is where all the identifiers for the data set are specified. The only entry in the form that must be specified is the **treatment** variable. However, if the **replicate** variable is not specified, Histopath will default to using SCABS (Standard Cochran-Armitage trend test By Slices) which is RSCABS without the RS correction. A warning; if a replicate structure was used in the experiment but is not specified in the analysis, pseudo-replication will occur which may lead to an increase in the number of falsely positive results. After all entry forms are filled out, click on the [Confirm Selected Values and Variables] button to set the selected variables into Histopath. After the selection is set, you can navigate back to the main tab to perform the RSCABS analysis. Note, that at any time you may navigate back to the Data specification tab to change a selection, just re-click the [Confirm Selected Values and Variables] button after a new selection is made to accept the change. Clicking on [Confirm Selected Values and Variables] is what causes Histopath to recognise the change in the variables.

Running RSCABS

After the data has been specified and the [Confirm Selected Values and Variables] has been clicked, the Histopath main tab should be navigated back to.



Two buttons; [Run RSCABS] and [Run Other Analyses] will have appeared on the Histopath main tab. Clicking on the [Run RSCABS] will perform RSCABS (or SCABS if a replicate variable is not defined), on the data.



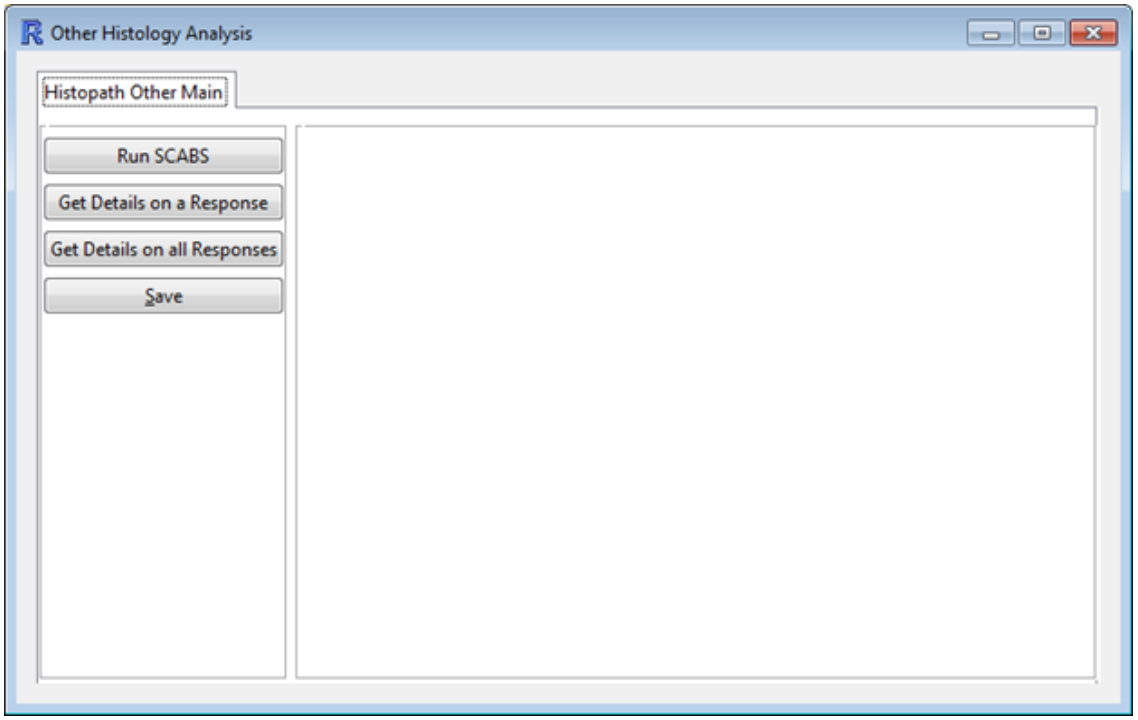
After the analysis on the data is ran, you may save the results by using the [Save Result] button, which will create a dialogue box that will prompt the saving of the results from the

RSCABS analysis as a csv file. Clicking the [Run Other Analyses] button will create a new window with options to perform SCABS, or get further details on a response.

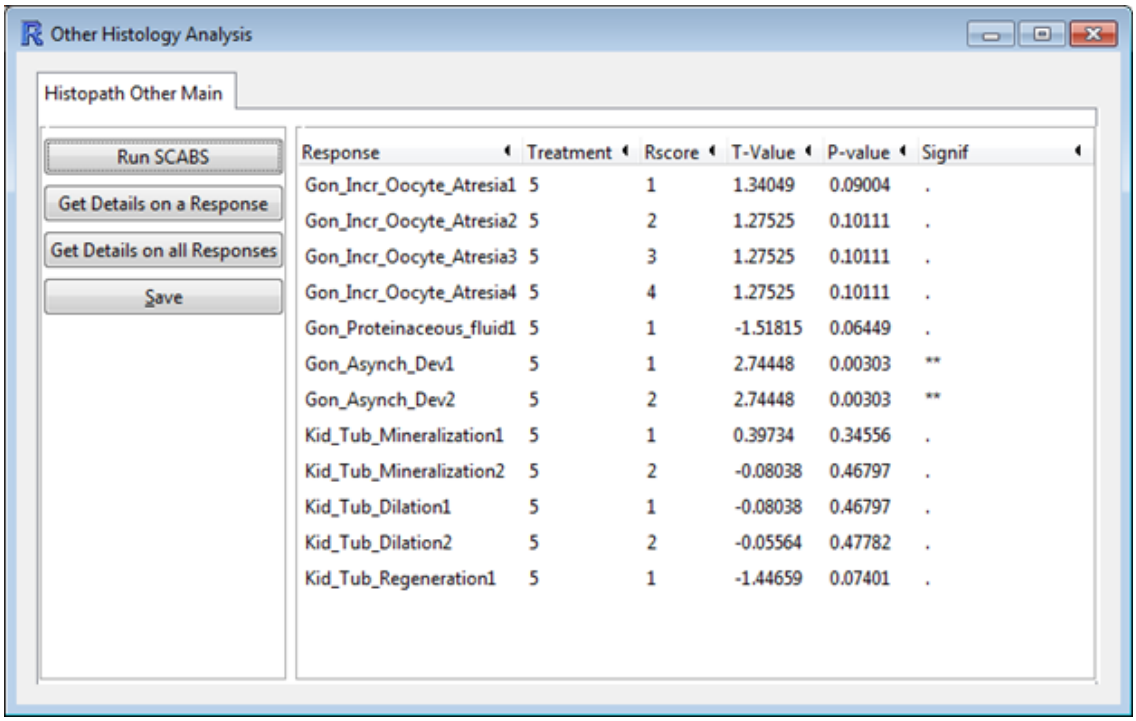
The results themselves appear in the box on the right hand side of the window. The **Response** is the endpoint that is being tested, **Treatment** is the treatment level, **R-Score** is the severity score, **Statistic** is the test statistic corresponding to that row's endpoint, treatment level, and R-Score, with **P-Value** as the corresponding p-value. **Signif** is the significance flag where "." is a p-value > 0.05, "*" is a $0.01 <$ p-value ≤ 0.05, "**" for $0.001 <$ p-value ≤ 0.01, and "***" for p-value ≤ 0.001.

Running Other Analyses

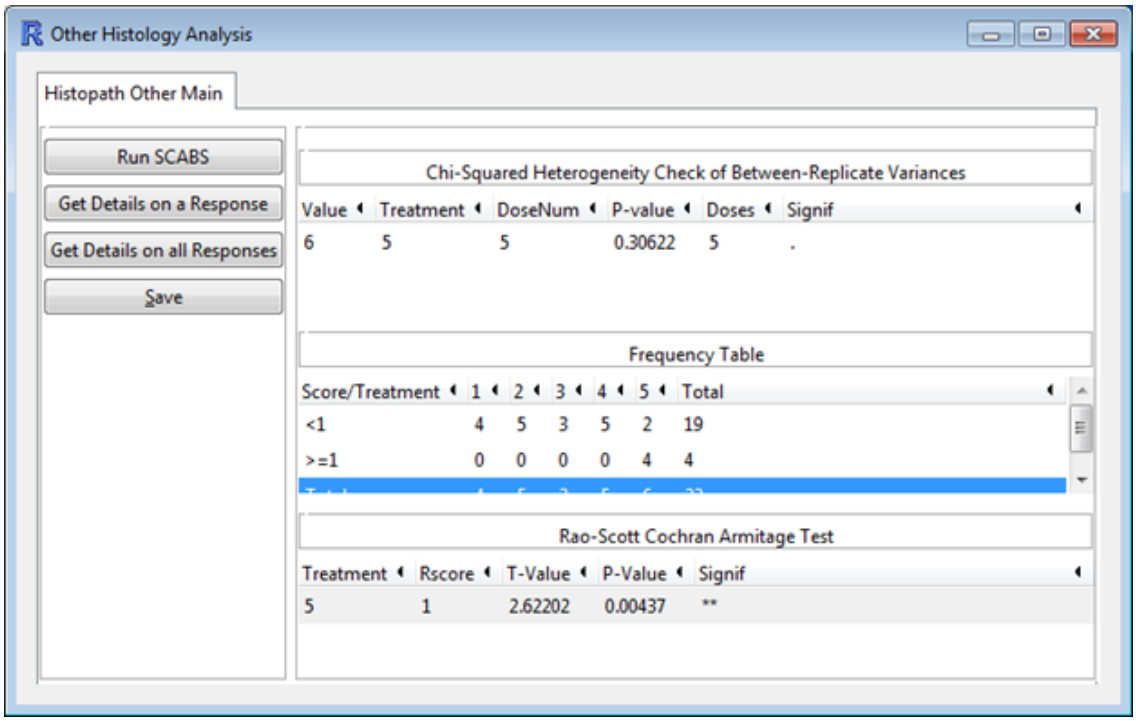
More details on each histopathological endpoint can be attained through the use of the [Run Other Analyses] button. This will bring up the Other Histology Analysis window.



On the Other Histology Analysis window the [Run SCABS] button will run a SCABS analysis on the data set. After the analysis is ran a table will appear with the results of the analysis. This table contains the same information as the results table from the RSCABS analysis.



The [Get Details on a Response] button will supply three tables for the selected response; a table for the Chi-squared (χ^2) test for heterogeneity of between-replicate variances. A frequency table, which contains the total observations for each combination of treatment (shown in the columns) and slice of score (shown in the rows). There will also be an additional table showing the results of RSCABS for that treatment level. If there are several unique severity scores for an endpoint, results for the lowest severity score will be displayed in the main tab while results for additional severity scores will be added to the window in extra tabs. The [Save] button will save the current result displayed in the window, whether it is results from SCAB or the more detailed results.



Finally, the [Get Details on all Responses] button will produce the three tables generated by [Get Details on a Response] for all responses. It does this by creating a new folder and populating that folder with HTML files containing the information.

RSCABS Command Line

As an alternative to using the GUI, RSCABS can also be run through command line. This is done through the runRSCABS function.

R Input

```
>runRSCABS(Data, Treatment, Replicate, Effects ,test.type)
```

The **Data** variable is a data set in the same format needed for the GUI. **Treatment**, is the name of the treatment variable, **Replicate** is the name of the replicate variable, and **Effects** is a vector of the endpoints exampleHistData.Sub tested. If **Effects** is not specified then the analysis will default to all columns that has at least one integer greater than 0 and no integers less then 20. The **test.type** input can be either "RS" or "CA" for either the RSCABS analysis or the SCABS analysis respectively. The **Replicate** input does not need to be specified, however if it is not specified **test.type** will default to "CA". An example of using runRSCABS is below. The code will produce the same results table as the example analysis in the **Running RSCABS** section.

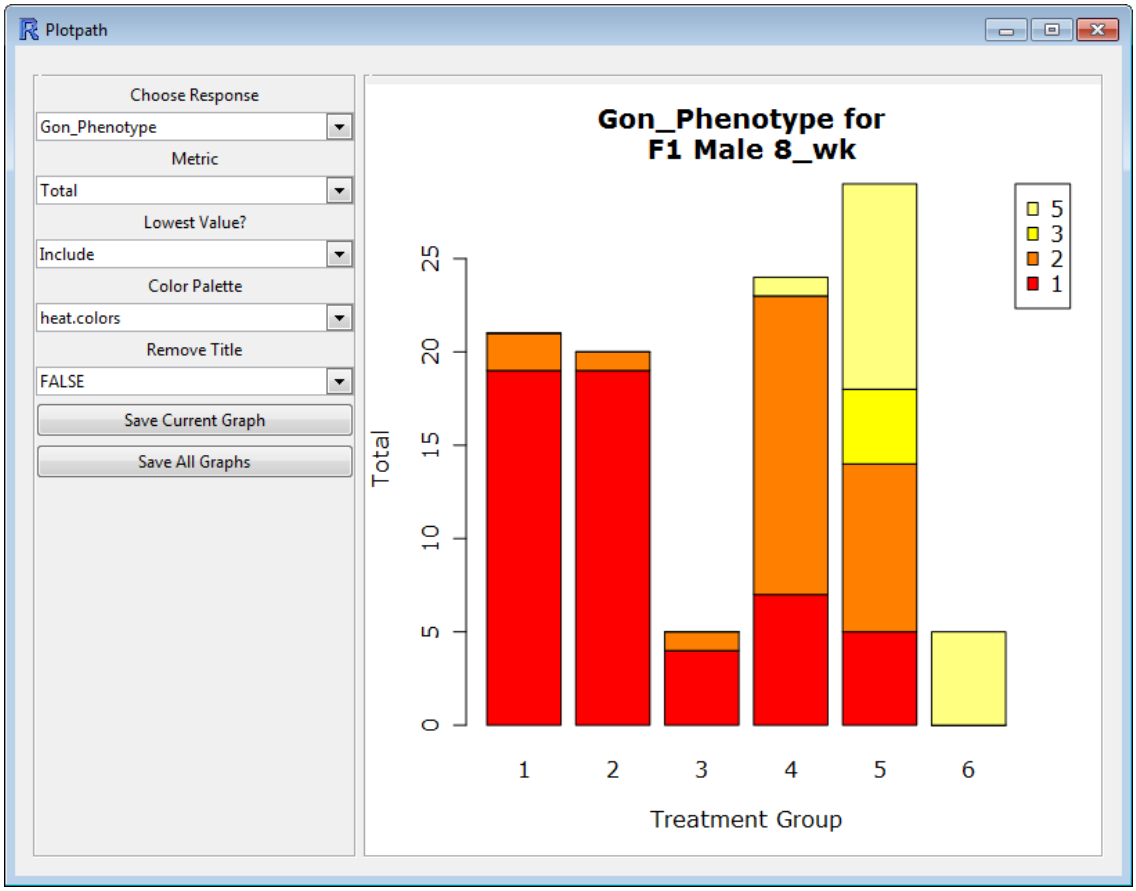
R Input

```
#Take the subset corresponding to F0-females of 16 weeks of age
  data(exampleHistData)
  subIndex<-which(exampleHistData$Generation=='F2' &
        exampleHistData$Genotypic_Sex=='Female' &
        exampleHistData$Age=='16_wk' )
  exampleHistData.Sub<-exampleHistData[subIndex, ]
#Run RSCABS
  exampleResults<-runRSCABS(exampleHistData.Sub,'Treatment',
        'Replicate',test.type='RS')
```

Plotting

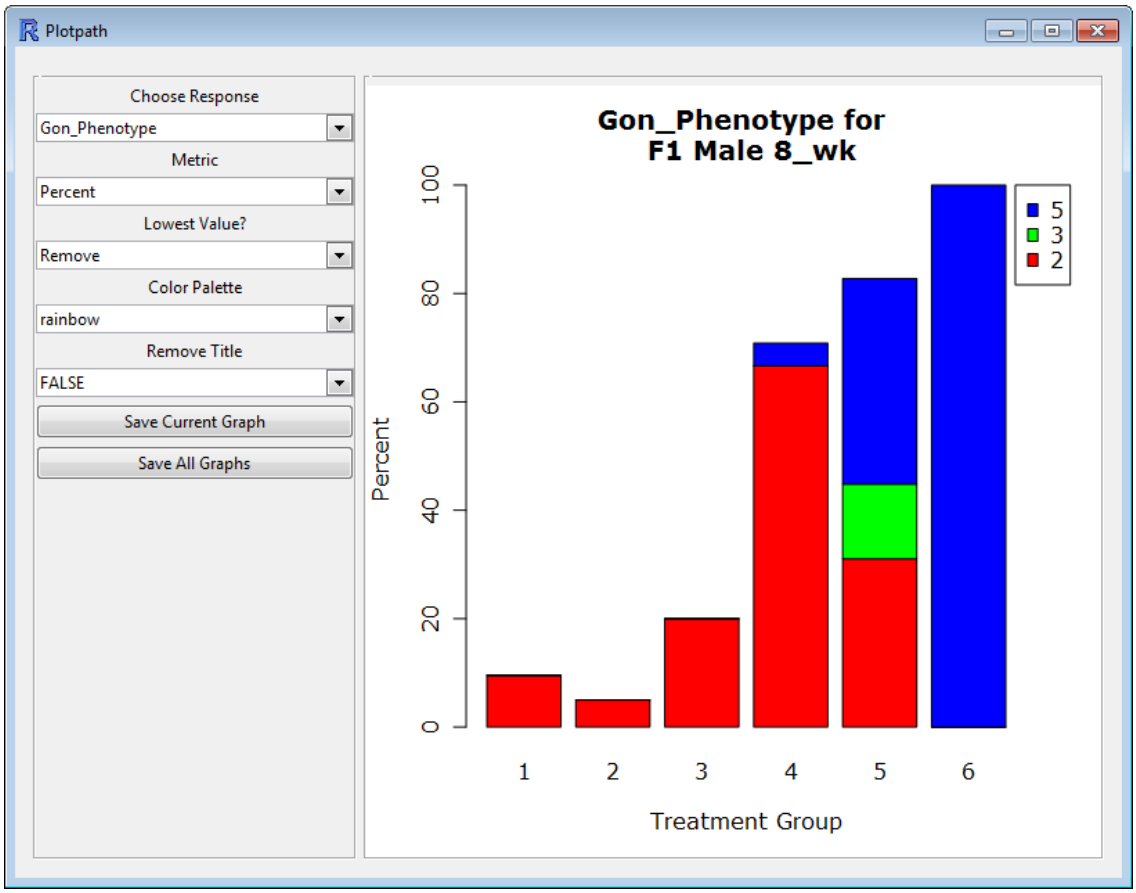
Plotting by GUI

The plotting window should appear after the [Run RSCABS] button from the Histopath main window is pressed.

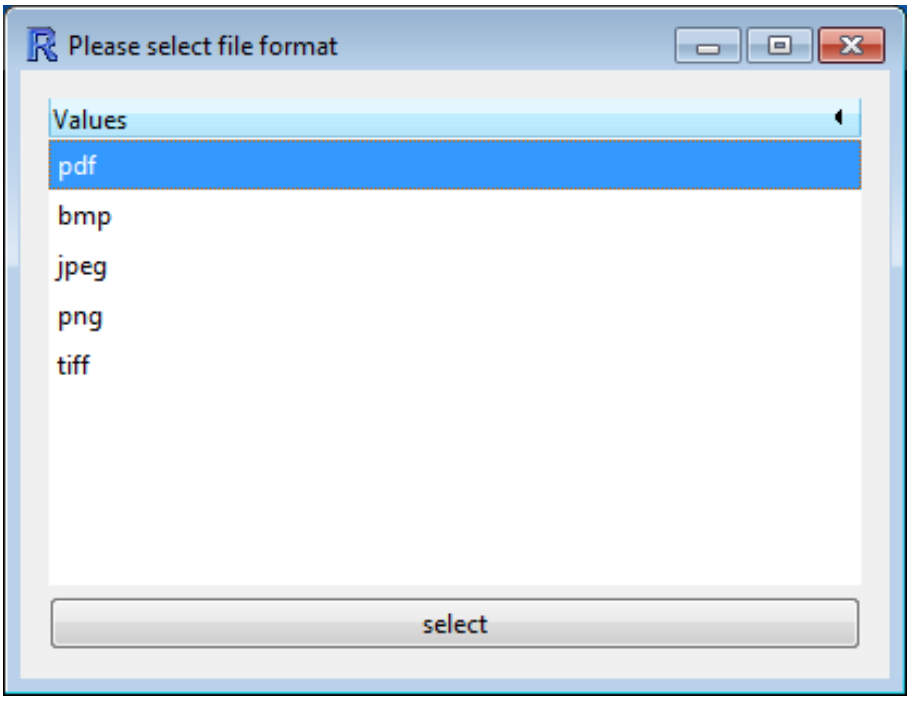


The plotting window uses stacked bar plots for visual representation of histopathological data. The left side of the window contains the plotting controls. Starting from the top of the left side; the **Choose Response** category allows for selections of the endpoint to graph. The list of possible endpoints is generated from the endpoint fields in the data set. Next, **Metric** controls the

y-axis as it allows for graphing either total counts or for percent of the total observations for each severity score and treatment combination. The entry "**Lowest Value?**" allows for the removal of the lowest value, which is in most cases 0 and indicates that the endpoint was not observed for an organism. Selecting "Percent" as the metric and choosing to "Remove" the lowest value will allow for easer observation of treatment effects when they are present. The **Color Pallet** option allows for the use of any of the standard R color pallets (including grey scale) and **Remove Title** will remove the title from the graph.



The **Save Current Graph** will save the current graph being displayed in any of the default file formats available to R. Selecting "tiff" will generate high quality tiff images at 600 dpi. The **Save All Graphs** button will generate graphs using the current setting for all endpoints containing at least one severity score greater then zero.



Plotting by Command Line

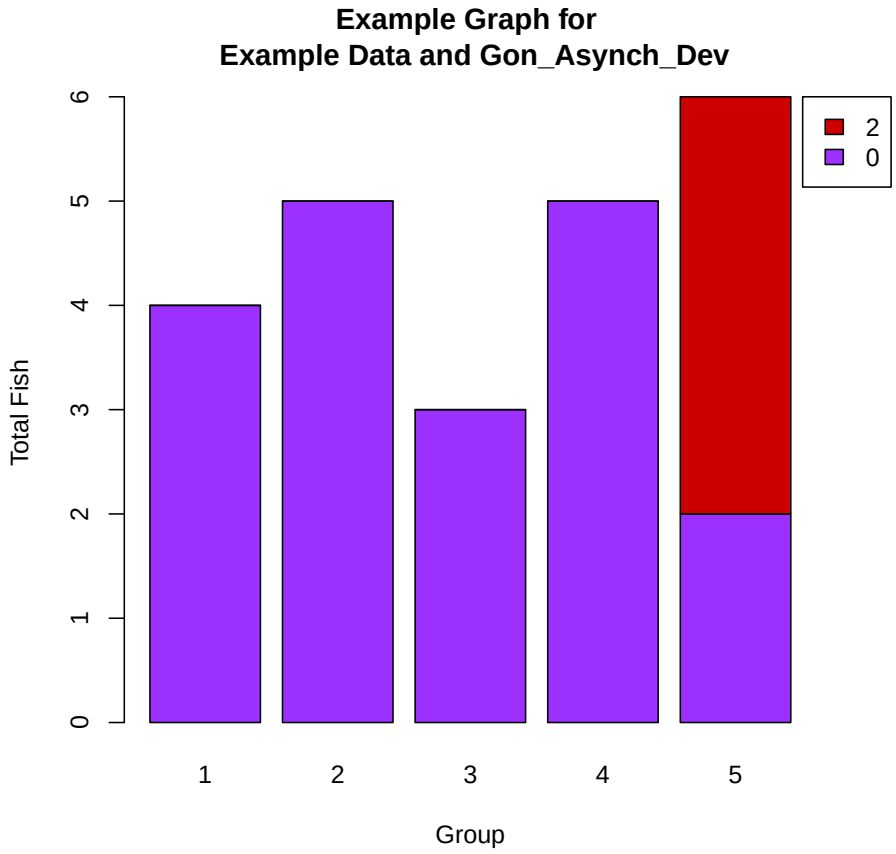
When greater control over the plot is desired, the plotting function can be called through the command line using the plotRSCABS function as shown in the example below.

R Input

```
#Sub-set the data
require(RSCABS)
data(exampleHistData)
subIndex<-which(exampleHistData$Generation=='F2' &
  exampleHistData$Genotypic_Sex=='Female' &
  exampleHistData$Age=='16_wk')
exampleHistData.Sub<-exampleHistData[subIndex, ]

xlab<-'Group'
ylab<-'Total Fish'
main<-'Example Graph for \n Example Data and Gon_Asynch_Dev'
col<-c('purple1','red3')

plotRSCABS(Data=exampleHistData.Sub, Effect="Gon_Asynch_Dev",
        Treatment="Treatment", Metric="Total", Lowest = "Include",
        PlotParms =PlotParms, Format = NULL, File = NULL,
        xlab=xlab,main=main,ylab=ylab,col=col)
```



The arguments of the function are; **Data**, which is a standard data set used by RSCABS, **Effect** which is the name of the endpoint plotted, and **Treatment** which is the name of the treatment variable. **Metric** controls for plotting either total counts ("Total") or percent of total counts ("Percent"), while **Lowest** allows for the removal ("Remove") or inclusion ("Include") of non-effected responses. **Format** is the name of any file format R can save graphs in, including high resolution graphs which are saved using the 'tiff' format. **File** is the name of the file the graph is saved to. Lastly, the **plotRSCABS** function can also include any argument used by the **barplot** function.

Acknowledgments

I would like to acknowledge Rodney Johnson and Kevin Flynn for their guidance through this project, Kevin Flynn as a beta tester, and Tim Dawson as a reviewer of the documentation. The R version of RSCABS was built for and paid by the USEPA under Contract EPD—13—052.

References

1. Armitage, P. 1955. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11**(3): 417—451

2. Cochran, WG 1954. Some methods for strengthening the common χ-squared tests. *Biometrics* **10**(4): 417—451

3. Green, John W. and Springer, Timothy A. and Saulnier, Amy N. and Swintek, Joe. 2014 Statistical analysis of histopathological endpoints. *Environmental Toxicology and Chemistry*, **33**(5): 1108—1116

4. OECD Guidelines for the Testing of Chemicals, Section 2. Test No. 240. DOI10.1787/20745761

5. Rao, J. N. K. and Scott, A. J. A. 1992. Simple Method for the Analysis of Clustered Data. *Biometrics*, **48**: 577—586.