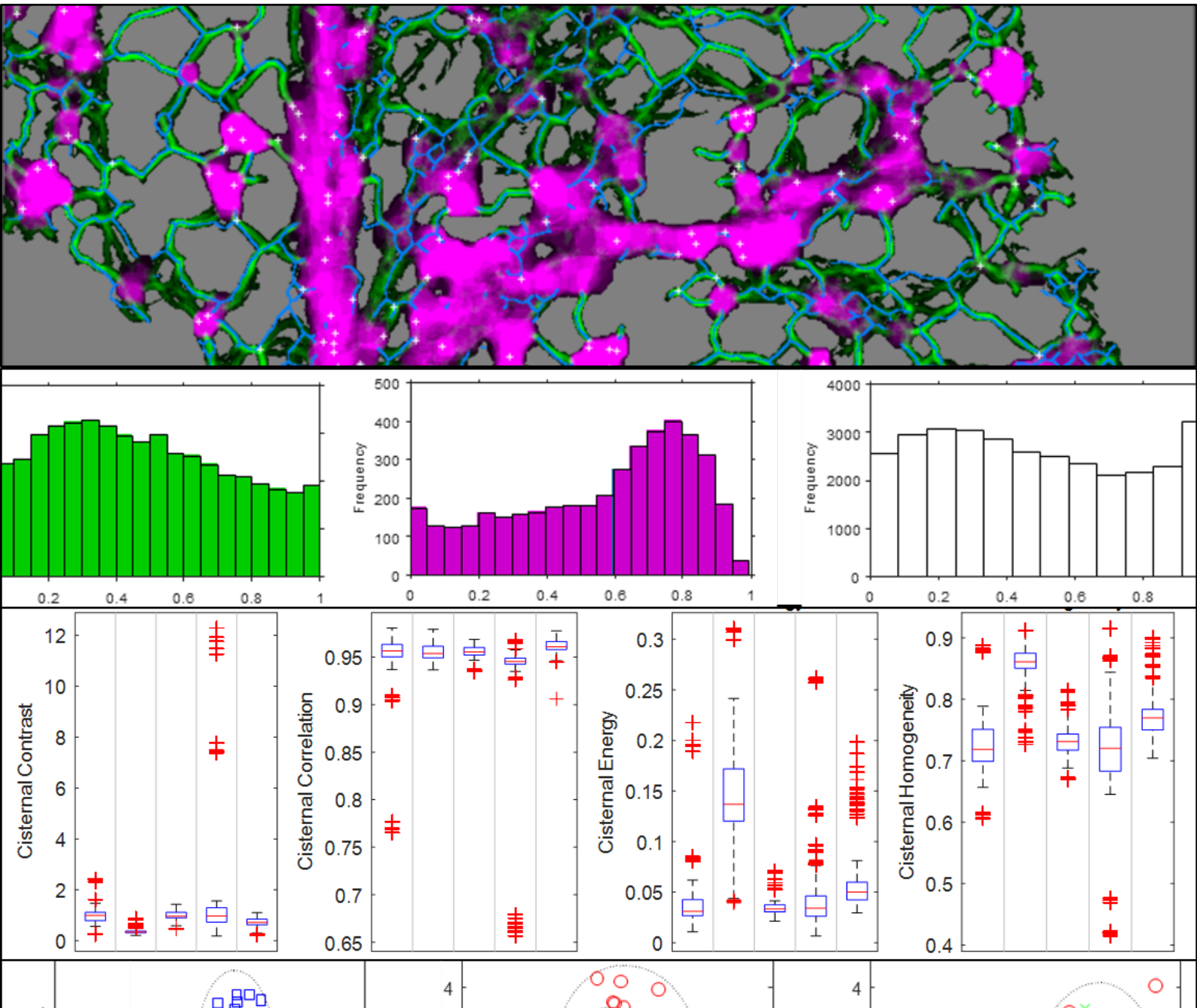
CHARLOTTE PAIN & MARK FRICKER

AnalyzER statistics Manual



Contents

Minimum system requirements 5

AnalyzER statistics interface 7

Parameter definitions 17

Bibliography 23

Minimum system requirements

For optimal performance of the AnalyzER statistics app, the minimum system requirements for analysis are as follows:

- A minimum screen resolution of 1600 x 900
- MATLAB 2017a or later with the Statistics and Machine Learning Toolbox
- The latest version of Java needs to be installed, and is available from:
  `https://java.com/en/download/`
- Output of images at full resolution uses *export_fig.m* originally written by Oliver Woodford (2008-2014) and now maintained by Yair Altman (2015-). When exporting to vector format (PDF or EPS) this function requires that ghostscript is installed on your system. Ghostscript can be downloaded from:
  `http://www.ghostscript.com`.
- When exporting images to eps and pdf formats, *export_fig* additionally requires pdftops, from the Xpdf suite of functions. This is included in the xpdf tools package and can be downloaded from:
  `https://www.xpdfreader.com/download.html`

To install the app:

1. Download the app to a local direction
2. Double click the installation file (characterised by the .mlappinstall suffix)
3. A dialog box will then open, select install
4. Once installed an icon will appear in the MATLAB toolbox strip. Double click this icon to open the AnalyzER statistical analysis app.

The function of this tutorial

This tutorial aims to provide an example of how to analyse a large ER morphology and dynamics datasets generated using AnalyzER. This example dataset was generated from a series of 5 frame timeseries confocal images collected of tobacco epidermal cells transiently expressing the ER lumenal marker GFP-HDEL with and without pharmacological (latrunculin B and Brefeldin A), stress (heat shock)treatments and protein over-expression (RFP-LNP2).

AnalyzER statistics interface

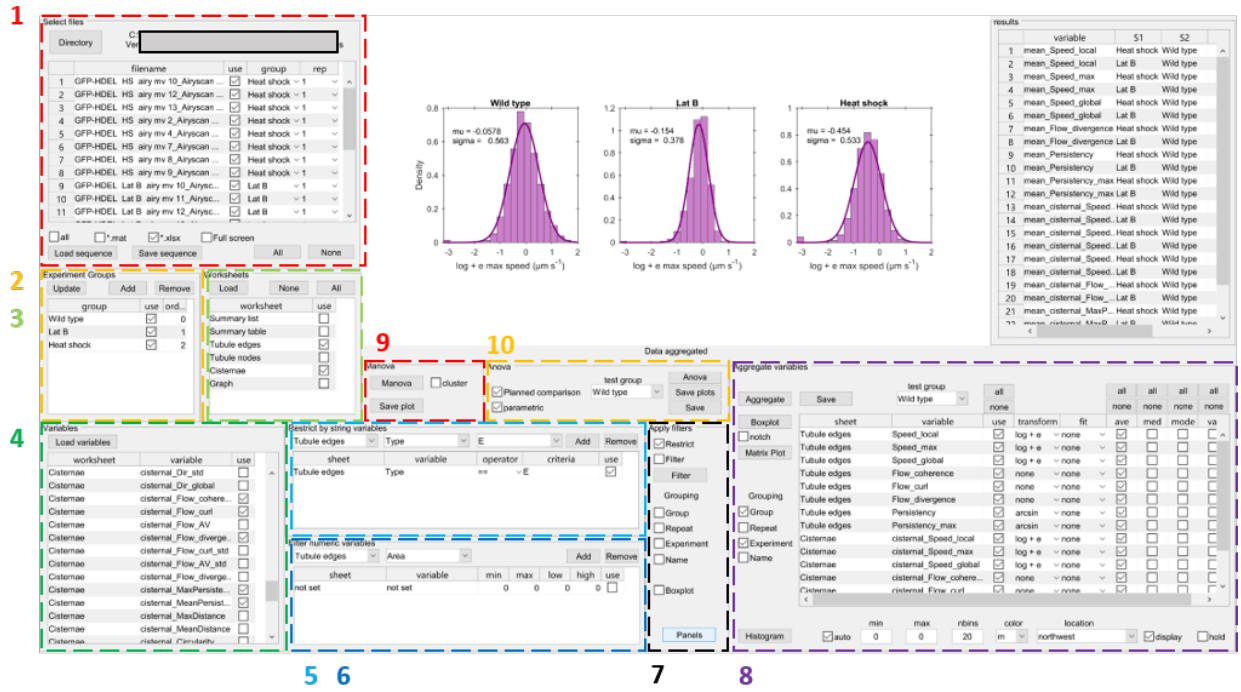


Figure 1: **The layout of the analyzer statistics package -** the results of an analysis of a truncated demo dataset.

The AnalyzER statistics package is divided into multiple panels, each of which must be run in sequence for accurate statistical analysis of AnalyzER outputs. The order is as follows:

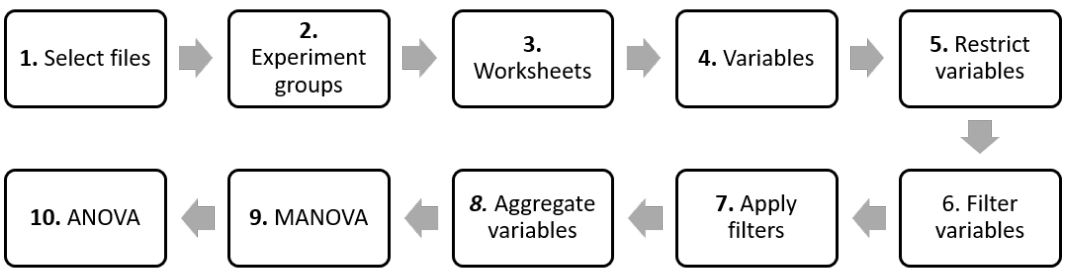


Figure 2: **The sequence of steps required for analysis.**

Using the AnalyzER associated statistics package

Datasets generated by AnalyzER are saved in a .xls file. This file contains 10 worksheets, including two summary worksheets and data on each ER morphological feature. The worksheets are named as follows:

Worksheet	Explanatory notes
Summary list	Mean value of all measurements made by AnalyzER displayed as a vertical table.
Summary table	Mean value of all measurements made by AnalyzER displayed as a horizontal table.
Tubule edges	The metrics of all identified tubules including intensity, length, area and tortuosity, listed by channel, section and frame.
Tubule nodes	The metrics of all identified tubule nodes including their dynamics, intensity and degree measurements of all tubules joining said node.
Tubule morphology	The mean morphology of tubules across channels, including the peak and trough values of intensity along tubules, including the co-variance between two channels.
Cisternae	The morphology and dynamics of cisternae, including area, circularity and persistency.
Cisternal perimeter	Measurements along the perimeter of cisternae.
Cisternal profile	Measurements of the average intensity distribution along radial transects normal to the boundary of the cisterna both inwards and outwards.
Polygons	Polygonal region area and morphology.
Graph	Graph-theoretic metrics of the ER, grouped by frame.

If an analysis is omitted, for example polygonal regions are not analysed, this worksheet will not be created.

Panels 1 and 2: Select files and designate experiment groups

The Select files panel is used to navigate to and load the AnalyzER results files selected for analysis.

1. Select the relevant file type checkbox at the bottom of the panel.
2. Use the **Directory** button to navigate to and select the folder containing all the relevant datasets. AnalyzER saves data as excel files stored in the folder 'results' which is stored in a newly generated folder named 'processed results'.
3. Move to panel 2 and use the **Add** and **Remove** buttons to designate the number of experimental treatments being analysed.
4. Each experiment has to be allocated as an experimental group. These are defined in the **experiment groups** panel.
5. Fill in the experiment group names and designate any specific ordering of the groups that may be desirable for creating boxplots later in the analysis.

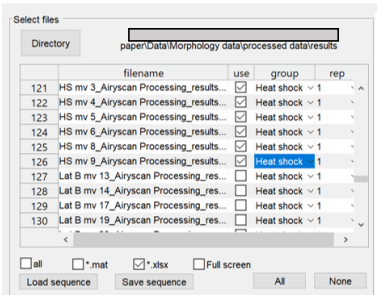


Figure 3: **Panel 1-** select the excel files to be used for analysis.

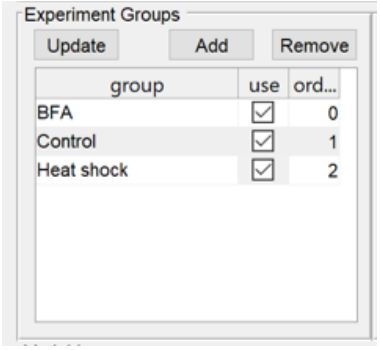


Figure 4: **Panel 2-** designate the experiment/treatment group labels.

6. Select the **Update** button to update the options available in the group column of the file input in panel 1.

7. In panel 1 it should now be possible to designate a specific experiment group to each excel spreadsheet using the dropdown options in the group column.

8. Designate all experimental groups prior to the next stage of analysis. There is also the option to allocate experiments to specific repetitions which can be used later in the analysis to group the data.

9. Use the **Save sequence** button to store these experimental group designations in the same directory, identifiable by the filename *sequence.mat*. This can then be re-loaded at any time by using the **Load sequence** button.

10. Select which data files are required for analysis using the use checkbox in panel 1. The **All** or **None** buttons select either all datafiles or deselects them respectively.

Panel 3: Selecting the worksheets of interest

Panel 3 is used to designate the worksheets of interest for later analysis. This process of selecting only relevant worksheets greatly speeds up the processes of loading data for analysis. For a reminder of the data outputs produced by AnalyzER see chapter 4.

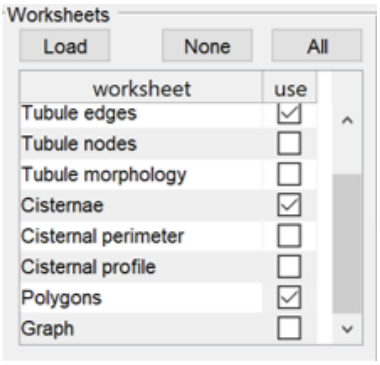


Figure 5: **Panel 3-** select the worksheets of interest.

1. Use the **Load** button to load all the available worksheets that are present in all the experimental results worksheets. If a set of results, e.g. polygons, is not available in all workbooks it will not be included.

2. Select the relevant worksheets using the checkboxes in the **use** column for further analysis.

3. The **None** or **All** buttons can deselect or select all worksheets respectively.

Panel 4: Selecting variables for analysis

Panel 4 is used to select the variables from the selected worksheets required for later analysis. A reminder of how these values is calculated is included in chapter 4.

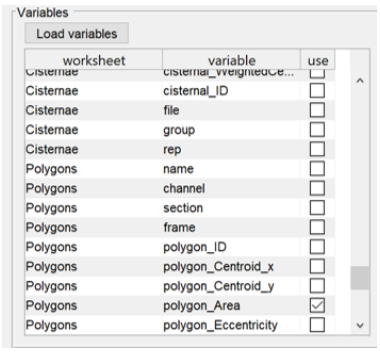


Figure 6: **Panel 4-** select variables of interest.

1. Use the **Load variables** button to load the variables from the worksheets selected in panel.

2. Select the variables of interest using the checkbox options in the **use** column.

3. This will load in all the variables for the selected worksheets from all the experiments into the program. This step will require some time.b Once the variabkes are loaded, they are all available for independent analysis and can be included ir excluded using the use column check box without having to load the data again.

Panel 5: Restricting variables

In some cases it may be appropriate to work on a subset of the variables selected from the categories. The **Restrict by string variables** panel is used ti select specific categories for analysis. **Restricting by string variables** is used to ensure that only variables within a certain category are taken forward for analysis. For example tubule edges includes graph edges that are associated with tubules (E) and edges that are associated with features such as cisternae (F). For the majority of analyses it is appropriate to restrict analysis to only edges that are associated with ER tubules (E).

In order to restrict by string variables, for example restricting data to tubule edges alone as opposed to edges associated with cisternae:

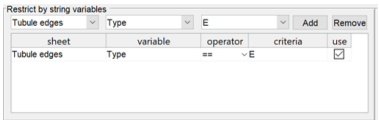


Figure 7: **Panel 5-** restrict by chosen criteria for later analysis.

1. Use the leftmost dropdown menu to select the worksheet to which the restriction should be applied. The next menu is automatically updated with the set of variables in chosen worksheet.

2. The second dropdown menu allows selection of the variable selection criteria including:

   - **Type** - referring to a specific variables within the worksheet on which restrictions can be applied
   - **Group** - referring the the experimental groups set in panel 2
   - **Name** - to restrict on the basis of datafile name

3. For this example of restricting by tubule edge classification, select **Type** from the central dropdown menu

4. The third dropdown menu contains the available values on which the data can be restricted, in this case select the **E** classification

5. Use the **Add** button to apply this restriction, this process can then be repeated to impose additional restrictions. The **remove** button can then be used to remove additional filters.

6. The operator dropdown menu in the top is used to select how the restriction operation is applied, by either ensuring that restriction is on the basis of, for example, inclusion or exclusion of the chosen criteria.

Panel 6: Filtering variables

The **Filter numeric variables** panel is used to select a specific subset of numeric variables for analysis that are filtered between minimum and maximum values. The process of filtering numeric variables is as follows:

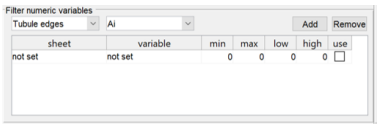


Figure 8: **Panel 6-** filtering numeric variables for later analysis.

1. Use the first of the dropdown lists to select the sheet that contains the variables to filter.

2. Select the variable to filter on from the second dropdown list.

3. Use the **Add** button to add the filter determined by the dropdown list.

4. Similarly the **Remove** button can be used to remove a given filter.

5. The minimum and maximum values of the variables currently in the selected dataset are displayed in the **min** and **max** columns. The highest and lowest value of the variables can then be set by typing values into the **low** and **high** columns.

6. Use the **use** tickbox to ensure that the selected filters are applied at later stages of the experiment.

Panel 7: Applying filters

Once the filters and restrictions have been defined in panel 5, they are selected to be applied in panel 6 using the filter button.

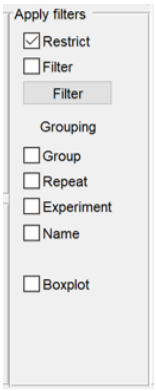


Figure 9: **Panel 7-** apply the restrictions and filters designated in the previous panel

1. Apply the filters and/or restrictions set in the previous panel by selecting the **Restrict** and **Filter** checkboxes and selecting the **Filter** button. This takes the selected data from panel 1, applies the filters and loads the data into the aggregate panel.

2. At this stage it is possible to display the data as boxplots data for visual inspection. Boxplots can be separated and grouped by: repetition, experiment or name.

3. Check boxplot to automatically produce boxplots of the data.

Panel 8: Aggregating, normalising and exporting data

Panel 8 is the key panel for display, transformation of filtered data and basic analysis of data. At this stage it is possible to export information in multiple forms including as box plots, as matrix plots and as histograms. Basic editing features are available as part of the export process. Below are methods for data visualization, transformations and export.

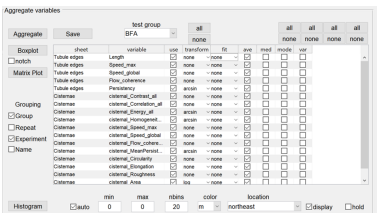
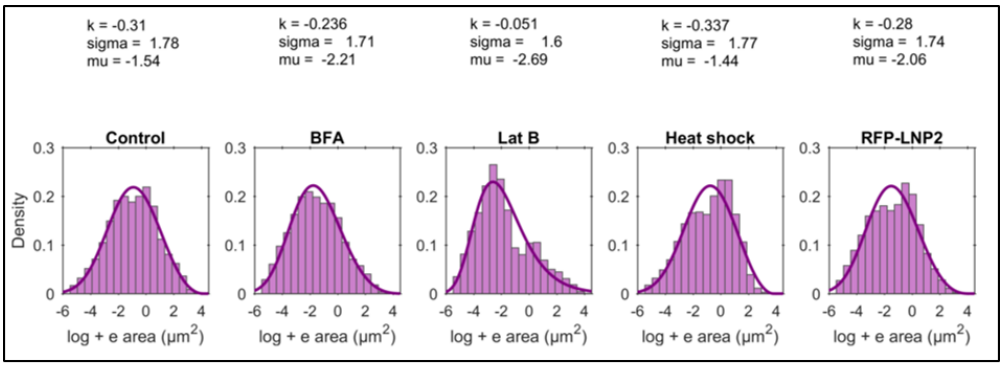


Figure 10: **Panel 8-** aggregate the chosen variables by experimental parameters and apply necessary transformations

Exploring results for a single variable

To create a histogram of the data sorted by group with a selected line fitting, select an option from the dropdown fit column. This produces the output shown in figure 11 see for more details. Selecting test from the **fit** column will apply the best fit line to histograms of the selected variable.



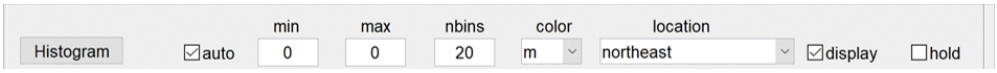


Figure 11: **Histogram display from panel 8** - (top) the best fit line applied to histograms of transformed data, grouped by designated experimental group. (bottom) the histogram display options at the bottom of panel 8.

The display of histograms can be modified using the series of display options at the bottom of panel 8. Modifications can be made as follows:

- The **auto** checkbox is used to apply the automatic display settings (shown in fig. 11).
- The **min** and **max** options set the minimum and maximum values of the histogram x axis.
- **nbins** sets the number of bins displayed with the histogram.
- The **color** dropdown menu changes the colour of the histogram displayed.
- **Location** sets the location of the relevant summary statistics of the line fit to the distribution.
- **Display** ensures that the distribution fit to histogram is displayed, alongside the parameters used to fit this line.
- **Hold** can be used to hold the currently displayed image, allowing the user to overlay an additional histogram over the top. This can be used to display, for example, mean speed and maximum speed on the same plot.

Transforming data

In order to transform the data for normality, use the dropdown options on the **transform** section to select the transformation. The test option is used to display both histograms of the selected variable

with different transformation (see fig. 12) and the Q-Q plot of the data. Displays only demonstrate the data for a test experimental group - which is selected at the top of panel 8.

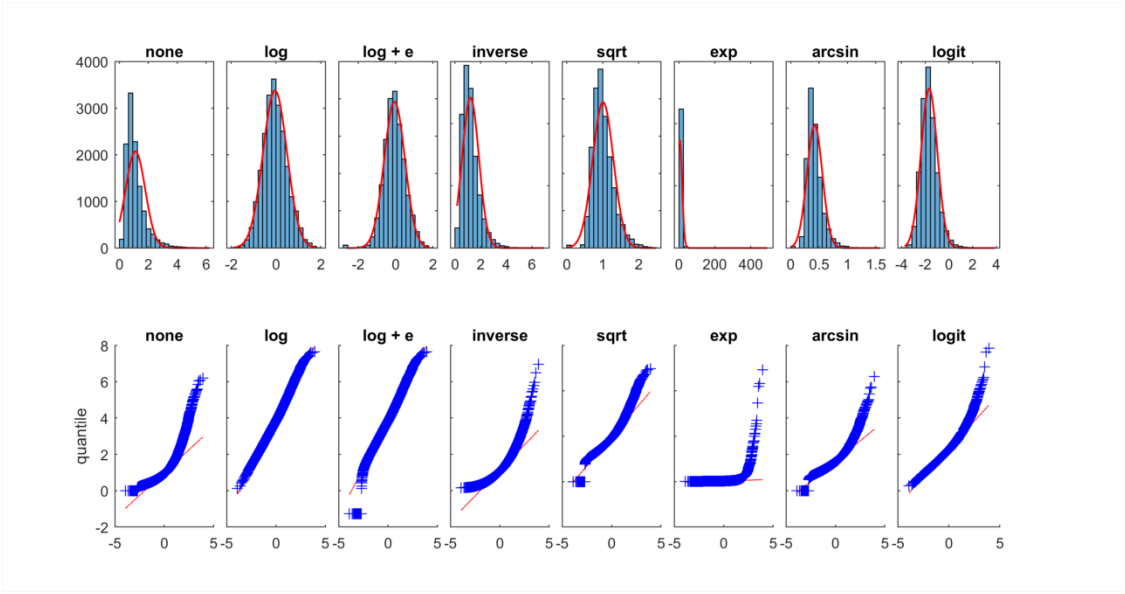


Figure 12: **Testing normality using panel 8**- the output of a selected test variable normality test, displaying data histograms and Q-Q plots after various transformations.

The transformations are performed as follows:

- **log** excludes all zeros from the analysis and applies a log transformation to the remaining values.
- **log + e** add half the minimum value to all zero entries to compensate for zero values that are below the detection limit
- **arscin** applies an arcsine transformation
- **logit** re-scales the data between 0-1 before taking a logit transformation

Once the most appropriate data transform has been identified, it is applied using the dropdown menus in the transform column on panel 8.

Visualizing multiple variables

Multiple variables can be visualized as either boxplots or as matrix plots. Below details the steps required to produce two plots, a matrix plot and boxplots. See figure 13 for more details:

1. Using the **Grouping** checkboxes, select which groupings are appropriate for the display of the relevant data. For example grouping by **Group** is recommended for boxplots as it produces a single boxplot for each treatment group as opposed to a plot per dataset.
2. The **Matrix Plot** button can be used to generate a plot of all the selected data.

3. The **Boxplot** button will produce boxplots of all the selected data, grouped by the checkboxes selected in the **Grouping** checkbox list. Selecting the **notch** checkbox will add a notch (a narrowing of the boxplot) around the median.
4. The **Aggregate** button groups data before any analysis, this button should be used after any changes are made to the selected data.
5. The boxplot can be saved using the **Save plot** button in panel 10.

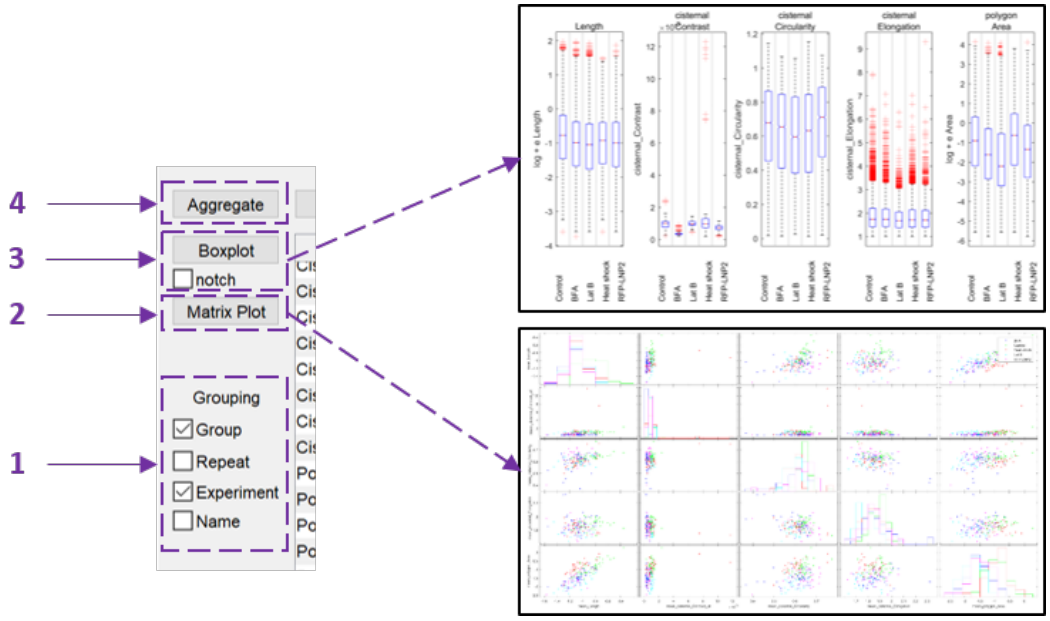


Figure 13: **Plots produced by panel 8** - left details a section of panel 8, and right details the two plot types that can be generated using this section.

Creating a summary of data

Create a summary of the aggregated data by selecting the statitical descriptors for export, including the average, median, mode and variance. These options are selected using the various statistical summary checkboxes in section 1 of panel 8, a processes which can be sped up by using the All or None buttons to select and deselect all boxes respectively. Use the Save button to export the summary of data.

Selecting data for further analysis

Once the appropriate data transformations have been selected, the use tickbox can be used to select the variables to be taken forward for further statistical analysis by MANOVA and ANOVA.

Panel 9: Performing a MANOVA

An initial MANOVA analysis can be performed to assess where the treatment groups differ significantly whilst controlling for type 1 errors. A MANOVA analysis is performed as below:

Figure 14: **Panel 9-** perform a MANOVA and save the plot

1. Re-select the **Aggregate** button in panel 7 to ensure that all the groupings and selected variables are updated and brought forward for analysis

2. Select MANOVA to perform a MANOVA analysis. The results are displayed in the main MATLAB window, including the results of serveral tests of MANOVA significance: Wilks' lambda, Pillai's trace, Hotelling-Lawley trace and Roy's largest root. Displayed in the central panel are pairwise scatterplots of the first 4 canonical variables from the MANOVA analysis grouped by treatment (If less are displayed this is due to fewer than 4 significant canonical variables being identified during the analysis). The dotted line around each group represents the 95% confidence limit.

3. If the cluster tickbox was selected, a dendogram plot of group means is also produced.

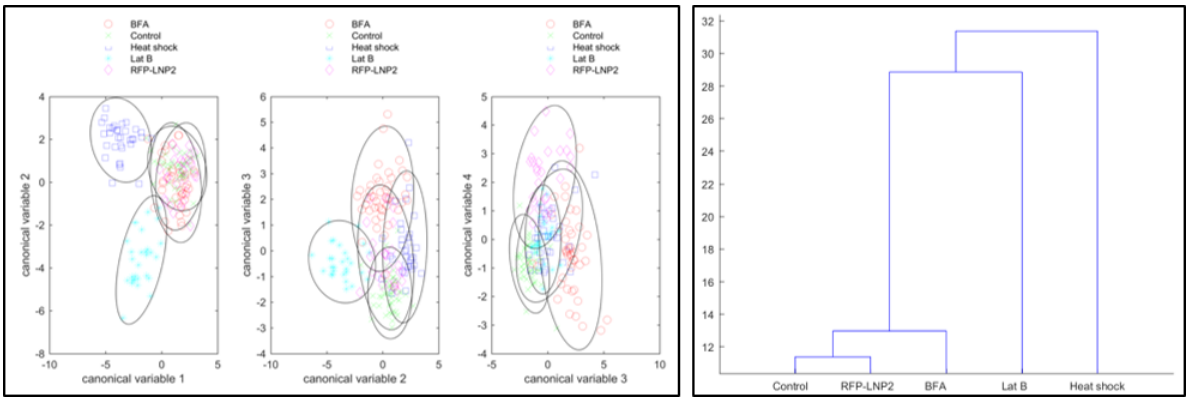


Figure 15: **MANOVA diplsay**- (left) the scatter plots of the first 4 canonical variables identified, the dotted line around each group represents the 95% confidence limit. (right) A dendogram plot of group means.

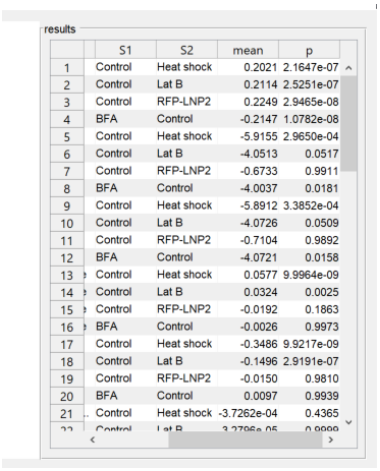
Panel 10: performing an ANOVA

Figure 16: **Panel 10-** perform multiple ANOVAs

Multiple ANOVAs can be performed after an initial MANOVA to test for significant separation of experimental groups.

An ANOVA is performed as follows:

- Selection of planned comparison will perform ANOVAs comparing all experimental groups against a control group selected in the test group dropdown box.
- If transformations have been used to normalise data the parametric tickbox must be selected. If the data is not normal a non-parametric test will be performed.
- Use the ANOVA button to perform the analysis and results are displayed in the far right results panel (see fig. 17).
- Use the Save button to save the outputs displayed in the results panel, including means comparison tests.



	S1	S2	mean	p
1	Control	Heat shock	0.2021	2.1647e-07
2	Control	Lat B	0.2114	2.5251e-07
3	Control	RFP-LNP2	0.2249	2.9465e-08
4	BFA	Control	-0.2147	1.0782e-08
5	Control	Heat shock	-5.9155	2.9650e-04
6	Control	Lat B	-4.0513	0.0517
7	Control	RFP-LNP2	-0.6733	0.9911
8	BFA	Control	-4.0037	0.0181
9	Control	Heat shock	-5.8912	3.3852e-04
10	Control	Lat B	-4.0726	0.0509
11	Control	RFP-LNP2	-0.7104	0.9892
12	BFA	Control	-4.0721	0.0158
13	Control	Heat shock	0.0577	9.9964e-09
14	Control	Lat B	0.0324	0.0025
15	Control	RFP-LNP2	-0.0192	0.1863
16	BFA	Control	-0.0026	0.9973
17	Control	Heat shock	-0.3486	9.9217e-09
18	Control	Lat B	-0.1496	2.9191e-07
19	Control	RFP-LNP2	-0.0150	0.9810
20	BFA	Control	0.0097	0.9939
21	Control	Heat shock	-3.7262e-04	0.4365

Figure 17: **Panel 9-** perform multiple ANOVAs

Parameter definitions

Tubule edge parameters

Metric	**Explanatory notes**
Original, cv	The mean intensity of pixels in the edge from the original image, along with the coefficient of variation (cv)
Intensity, cv	The mean intensity and cv after background subtraction
Width, Width cv	The estimated width (2r) and cv (µm)
Width center	The estimated width excluding overlap regions at the node
Length, length center	The total length and length excluding the overlap at nodes (l, µm)
Area, Volume	The cross-sectional area ($a = \pi r^2$, µm^2) and volume ($v = a \times l$, µm^3)
Resistance	The predicted resistance to flow assuming Poiseuille flow (l/r^4, µm^{-3})
Number	The number of pixels in the edge
Tortuosity	The Euclidean distance between the nodes divided by the total length of the edge
Speed - local	The scalar sum of the speeds calculated for each pixel (µm s^{-1})
Speed - max	The maximum speed for any pixel (µm s^{-1})
Speed - global	The vector sum of speeds for each pixel (µm s^{-1})
Dir - local, std	The angular mean and standard deviation of the flow direction at each pixel (°)
Dir - global	The direction of the vector sum (°)
Flow coherence	The ratio of the local speed (scalar sum) to the global speed (vector sum)
Flow curl, std	The local angular rotation of the vector field (°)
Flow AV, std	The local angular velocity of the vector field (° s^{-1})
Flow divergence, std	The divergence of the vector flow field (°)
Persistency, cv	The mean period of time that each pixel forms part of a tubule (s)
Distance	The geodesic distance of each pixel from a manually defined reference point (µm)
Region	The integrated sum of intensities in the region closest to each edge
BC	The edge betweenness centrality

Tubule morphology parameters

Metric	Explanatory notes
peak number	The total number of peaks measured excluding nodes. Values are given for each channel analysed
peak density	The number of peaks per unit length (μm^{-1})
length	The total length of tubules analysed (µm)
peak value	The mean intensity at the peak (I_P)
peak width	The estimated width of the peak (µm)
peak separation	The mean separation of peaks, which is only calculated if there is more than one peak per tubule (µm)
peak covar	The mean covariance of intensities from two channels measured at the peak ($(I_{P1} - \overline{I_{P1}}) \times (I_{P2} - \overline{I_{P2}})$). Positive values indicate both channels show similar behaviour, whilst negative values indicate they move in opposite directions
peak ratio	The mean ratio of intensities at the peak from two channels (I_{P1}/I_{P2}) trough number The number of troughs (measured as peaks in the inverted intensity profile)
trough density	The number of troughs per unit length of tubule (µm)
trough value	The mean intensity in the trough (I_T)
trough width	The mean width of the trough (µm)
trough separation	The mean separation of troughs if there is more than one trough per tubule (µm)
trough covar	The mean covariance of intensities from two channels measured in the trough ($(I_{T1} - \overline{I_{T1}}) \times (I_{T2} - \overline{I_{T2}})$)
trough ratio	The mean ratio of intensities from two channels measured in the trough (I_{T1}/I_{T2})

Node parameters

Metric	Explanatory notes
node degree	The number of tubules connected to each node, excluding cisternal nodes
total width	The sum of the tubule widths incident at each node
mean width	The average of the tubule widths incident at each node
intensity	The original intensity at the node
persistency	Node persistency
speed	Node speed
direction	Node direction
distance	The distance from the reference point
Maj, Mid and Min	The widths of the three main tubules incident at the node. Only one value is given for a terminal node, whilst few node have a degree greater than three
Min-Maj, Mid-Maj and Min-Mid	The ratio of the tubule widths for the three main incident tubules
Omaj, Omid and Omin	The orientation of the three main incident tubules, determined as a linear segment between the node and the midpoint of the tubule
Omin-Omaj, Omid-Omaj and Omin-Omid	Branch angles between the three main incident tubules, determined from a linear segment from the node to the midpoint of each tubule

Graph parameters

Metric	Explanatory notes
G	The number of sub-graphs in the network. This is typically one as only the largest connected component is selected
nodes	The number of nodes (excluding cisternal nodes)
links	The number of edges (excluding cisternal edges)
total length	The total length of the network tubules (mm)
mean length	The average length of the tubules (µm)
median length	The median length of the tubules (µm)
mean width	The average width of the tubules (µm)
median width	The median width of the tubules (µm)
volume	The total volume of the network (mm^3)
k	The average node degree
G efficiency	The global efficiency of the network
R efficiency	The root efficiency of the network calculated to the first cisternal node
cyclomatic no.	The cyclomatic number
alpha	The alpha coefficient or meshedness
beta	The beta coefficient
gamma	The gamma coefficient
diameter	The longest shortest path through the network

Cisternae morphology parameters

Metric	Explanatory notes
ID	The node identity displayed on the image overlay
area	The area of each cisterna
major axis and minor axis	length (in pixels) of the major and minor axis of the ellipse that has the same normalized second central moments as the cisterna
orientation	The angle between the x-axis and the major axis
solidity	The proportion of the pixels in the convex hull that are also in the cisterna
perimeter	The perimeter of the cisterna
elongation	The ratio of the major axis to the the minor axis
roughness	The ratio of the perimeter2 to the area
circularity	The ratio of the radius determined from the area to the radius determined from the perimeter

Intensity-based cisternae metrics

Metric	Explanatory notes
mean Intensity, max Intensity and min Intensity	metrics calculated for each cisterna from the background-subtracted image
node degree	The number of connecting tubules incident on the cisterna
node strength	The sum of the widths of the tubules connecting to the cisterna (µm)
average node strength	The node strength divided by the node degree, (µm)
max and mean persistency	The maximum and average persistency of pixels in each cisterna (s)
speed - local, speed - max and speed - global	The scalar sum of local speeds at every pixel, the maximum speed across all pixels, and the global (vector sum) speed ($\mu m\, s^{-1}$)
Dir - local, Dir - std, and Dir - global	The mean angular direction of movement for all pixels in the cisterna along with the angular standard deviation, and the mean global direction from the vector sum (°)
Flow coherence	The ratio of the local speed (scalar sum) to the global speed (vector sum)
Flow curl, std	The local angular rotation of the vector field within the cisternae (°)
Flow AV, std	The local angular velocity of the vector field ($^\circ\, s^{-1}$)
Flow divergence, std	The divergence of the vector flow field (°)
Persistency, cv	The mean period of time that each pixel forms part of a tubule (s)
max distance and mean distance	The maximum and average distance of any pixel to the edge of the cisterna (µm)
variance	The square of the difference in intensity of each pixel from the mean within the cisterna

Cisternal texture metrics

Metric	Explanatory notes
Contrast	A measure of the intensity contrast between a pixel and its neighbour. Values range from 0 to $(nbins - 1)^2$. An idealised cisternal sheet would have a contrast of zero.
Correlation	A measure of how correlated a pixel is to its neighbour. Values range from -1 (un-correlated) to 1 (fully correlated). An idealised cisternal sheet would have a value of 1.
Energy	Gives the sum of squared elements in the GLCM. Values range from 0 to 1. An idealised cisternal sheet would have an energy of 1.
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the diagonal. Values range from 0 to 1. An idealised cisternal sheet would have a value of 1.

Morphological metrics of polygonal regions

Metric	Explanatory notes
ID	The identity of the polygonal region
area	The area of each polygonal region with the ER tubules thinned to a single-pixel wide skeleton
major axis and minor axis	length (in pixels) of the major and minor axis of the ellipse that has the same normalized second central moments as the region
orientation	The angle between the x-axis and the major axis
solidity	The proportion of the pixels in the convex hull that are also in the region
perimeter	The perimeter of the region
elongation	The ratio of the major axis to the the minor axis
roughness	The ratio of the perimeter^2 to the area
circularity	The ratio of the radius determined from the area to the radius determined from the perimeter
max distance	The furthest distance within the region to the ER network
mean distance	The average distance within the region to the ER network
exclusive area	The area excluding the width of the ER tubules

Bibliography