statSuma *v*.1.3 Manual

Citation:

Leigh R.J., Murphy R.A., and Walsh F. (2021) Paper title, Journal metrics

Description:

statSuma is a user-friendly script for selecting (and performing) the most appropriate statistical tests in microbiome studies.

Table of contents:

Before you start:

Data should be **tab delimited** format (.tsv) with the first column containing the taxon IDs (or names) and the first row containing the groups each sample belongs to. The first row should contain the identifier word "**Group**". An example of this data is shown below (Table 1):

Table 1: Example dataset

Group	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
TaxonA	117	118	104	108	108	1117	577	33	217	200	227	203	218	570	1446	144
TaxonB	221	208	206	227	214	945	1730	147	223	230	214	230	216	1901	1205	55
TaxonC	40	32	30	36	45	481	297	13	43	43	46	49	46	412	291	21
TaxonD	64	70	68	62	71	361	410	18	63	71	64	69	75	419	565	25
TaxonE	116	120	100	115	108	1103	845	58	106	100	120	120	111	1110	692	70
TaxonF	25	21	16	30	17	347	189	7	209	217	212	214	208	645	1752	83

Where to start:

This script is presented as an "interactive python notebook" (ipynb) file and can be ran using Google Colaboratory (Colab) (https://colab.research.google.com/). Once Colab is open in your browser, select "**upload notebook**" in the "**File**" tab in the top, right-hand corner of the screen (**Figure 1**), and upload the associated ipynb file (statSuma_vX.X.ipynb) available at https://github.com/RobLeighBioinformatics/statSuma

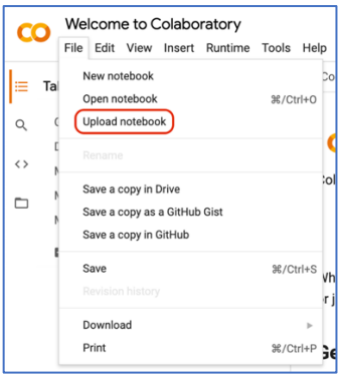


Figure 1: How to upload the ipynb file

Setting up statSuma

The critical α thresholds that statSuma uses to determine whether a distribution is Gaussian or whether two distributions are equivariant can be configured by double clicking on the "change critical alpha" cell in statSuma (Figure 2).

Figure 2: The "change critical alpha" cell in statSuma

Each critical α is set to 0.05 by default but these can be changed to any float between 0 and 1. This value is set **without quotations** (Figure 3).

```
#@title **Click here to set critical alpha va

pairwise_equivariance_critical_alpha = 0.5
pairwise_gaussian_critical_alpha = 0.5
listwise_gaussian_critical_alpha = 0.5
listwise_equivariance_critical_alpha = 0.5
```

Figure 3: Critical α settings

Data scaling ensures all samples have an equivalent sum. By default, statSuma scales data to mitigate errors due to data with low coverage/depth. This can be switched off by double clicking on the "scale data" cell (Figure 4) and setting "**scale_data**" to **"no"** (**with quotations**).

Figure 4: The scale data cell in statSuma

Running statSuma:

When ready, statSuma can be ran using a using the "**Run all**" option in the "**Runtime**" tab at the upper, right-hand side of the screen (**Figure 5**). When ran, statSuma will display its recommendations beneath the "**Click here to conduct pairwise tests**" and "**Click here to conduct listwise tests**" cells respectively.

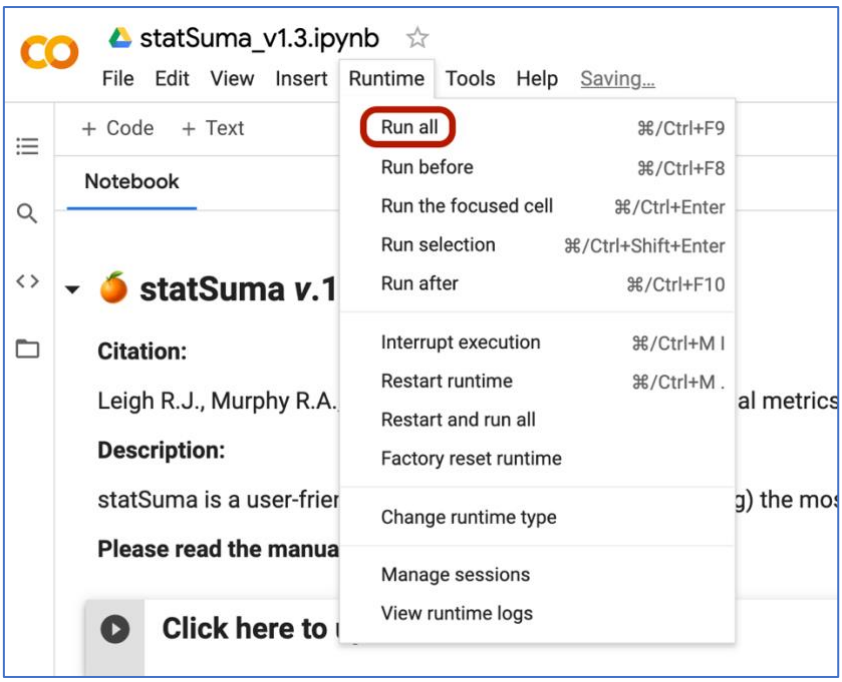
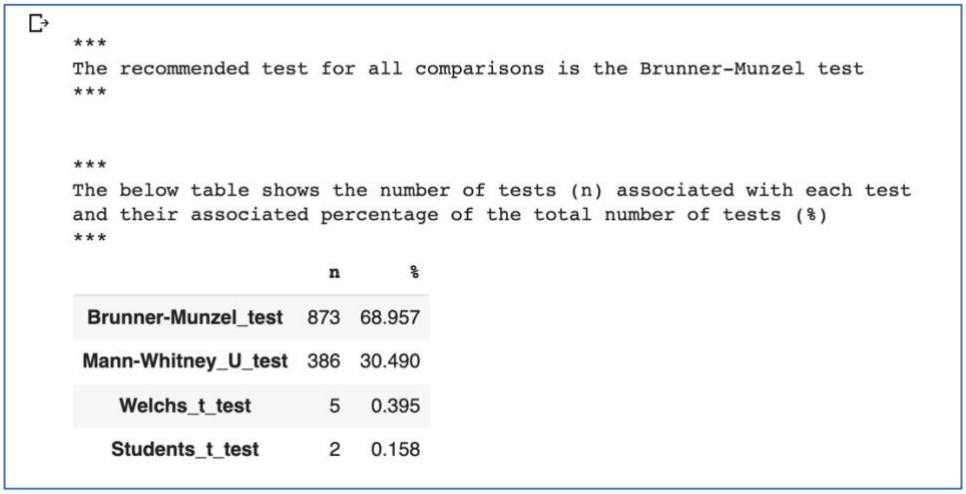


Figure 5: Location of the "Run all" command

Once the "Run all" command has been selected. The user is required to upload a file in the presented window of the "**Click here to upload a dataset file**" cell (Figure 6).

Figure 6: Location of the file upload function

Alternatively, once files have been uploaded, each cell can be ran independently. As alluded to above, the "**Click here to conduct pairwise tests**" and "**Click here to conduct listwise tests**", perform all comparative statistical test. As mentioned above, **the recommended statistical test is printed below these cells** (Figure 7).



```
***
The recommended test for all comparisons is the Brunner-Munzel test
***


***
The below table shows the number of tests (n) associated with each test
and their associated percentage of the total number of tests (%)
***
```

	n	%
Brunner-Munzel_test	873	68.957
Mann-Whitney_U_test	386	30.490
Welchs_t_test	5	0.395
Students_t_test	2	0.158

Figure 7: Recommended test box

The remaining 3 cells produce a series of plots to visually verify the results of the Levene's test and Shapiro-Wilk test. The "**Click here to plot variance distributions**" cell produces a plot for each pair of standardised distributions for a given taxon between two geographical (anatomical) sites. The "**Click here to plot data standardised distributions against a Gaussian distribution**" and "**Click here to plot QQ-plots**" cells produce a standardised distribution for each taxon at each site alongside a Gaussian distribution as a histogram or QQ-plot respectively.

Output files:

Upon completion, statSuma produces eleven files, five pairwise statistics files (denoted by the test used to generate the result (*eg.* BrunnerMunzel) with "PairwiseResults.tsv" (*eg.*

BrunnerMunzelPairwiseResults.tsv), two listwise statistics files, two "Recommended comparisons files" (listwise and pairwise) and three pdf files containing the Gaussian distribution comparisons, QQ-plots and variance comparisons respectively. These files can be accessed and downloaded using the **folder icon** on the left side of the screen (Figure 8).

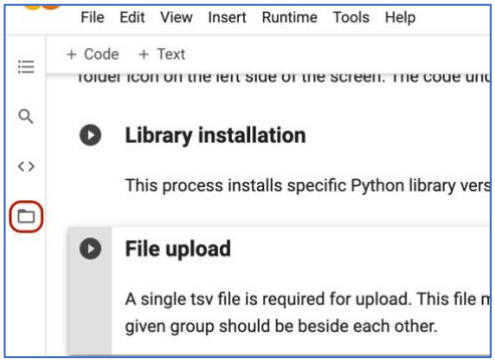


Figure 8: Location of the results files

Pairwise results files contain 22 columns, the first three denote the taxon and two groups/sites being compared (group A and group B), the following 14 columns describe the mean, standard deviation, median, variance, minimum and maximum for groups A and B respectively. The final four columns are "comparative statistics" and consist of the statistic computed by the associated test, naïve *P*-value, the Bonferroni-Dunn corrected *P*-value (P_{BD}) and the "Difference" between the two groups (Group B with respect to Group A). The difference can be one of "increase", "decrease" or "no change" and is calculated using the mean for *t*-tests and using the median for all other tests.

Listwise results files consist of five columns, the test used, the taxon sampled, the computed statistic, the naïve *P*-value and the P_{BD}.

Pairwise test recommendation files consist of eleven columns: the taxon sampled, the two groups/sites being sampled, the sizes of group A and group B, Levene's statistic (used to compute equivariance), the Shapiro-Wilk statistic for group A and group B, the Levene and Shapiro-Wilk *P*-values, and the recommended test.

Listwise test recommendation files consist of six columns: the taxon sampled, Levene's statistic, Levene's *P*-value, the minimum Shapiro-Wilk *P*-value and associated statistic, and the recommended test. Levene's tests can be visually examined using the output of the "**Click here to plot variance distributions plots**" (*eg.* Figure 9).

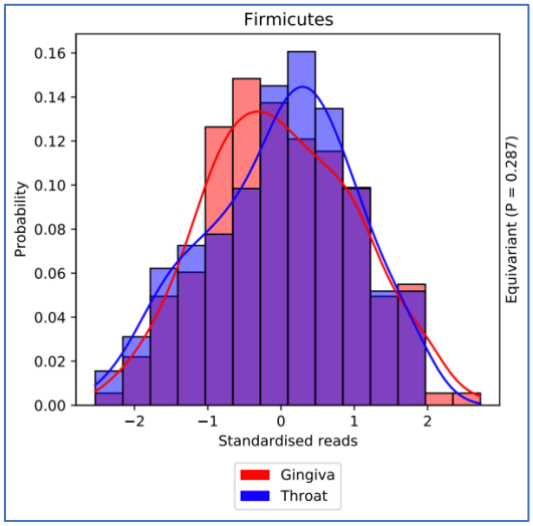


Figure 9: Example equivariance distribution plot

Likewise, Shapiro-Wilk tests can be visually inspected using the output of the "**Click here to plot data standardised distributions against a Gaussian distribution**" (Figure 10) and "**Click here to plot QQ-plots**" cells (Figure 11)

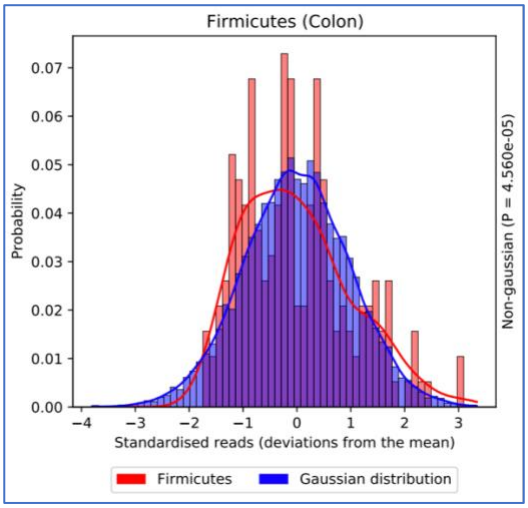


Figure 10: Example Gaussian distribution comparison (histogram format)

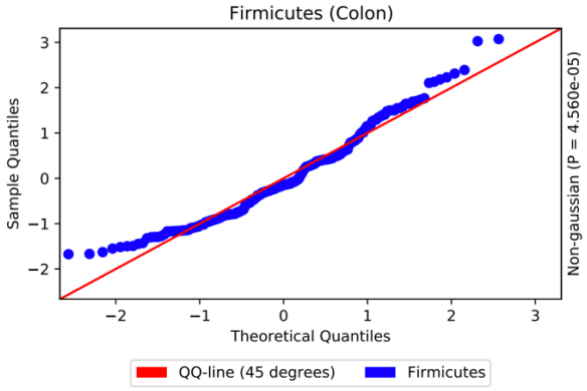


Figure 11: Example Gaussian distribution comparison (QQ-plot format)