

16S Statistical Analysis Tool



Prepared by:
STREMBLE Ventures Ltd
www.stremble.com



Contents

1	Description.....	3
2	Input.....	4
2.1	Analysis Name	4
2.2	Samples' Info Google Drive share link.....	4
2.2.1	Samples' Info File Description	4
2.2.2	Creating the Google Drive Share Link	5
2.3	Control Group	7
2.4	Significance Level.....	7
2.5	Aggregated Counts Files	7
2.6	Save Results To.....	8
3	Output	8
3.1	Report.....	8
3.2	Analysis files	8
3.3	Figures and Tables	9
4	Support.....	9

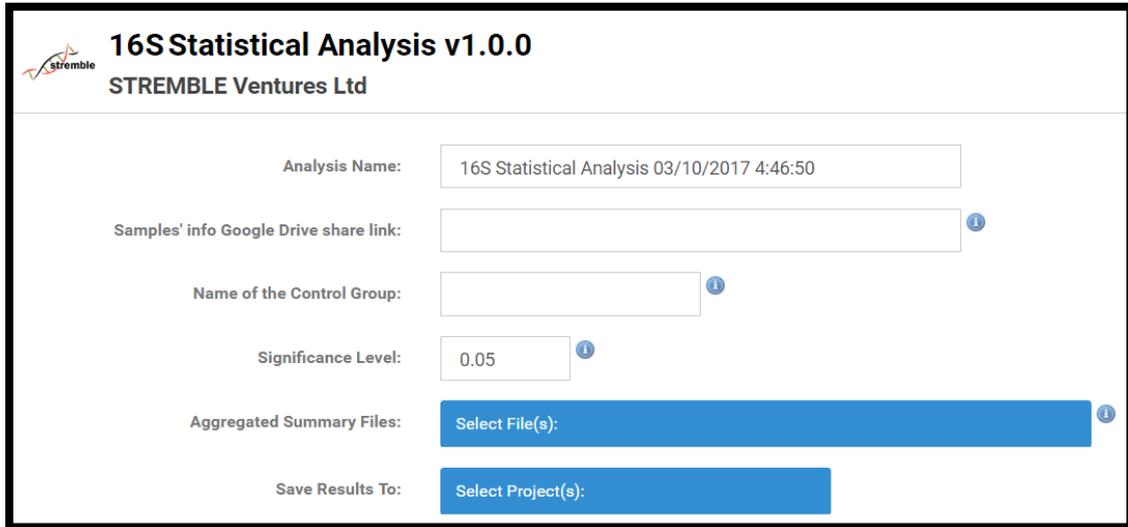
1 Description

The statistical analysis of 16S metagenomics experiments is challenging, especially when they include multiple visits and groups. The complexity is further enhanced when correlating the samples analysed with Illumina's BaseSpace "16S Metagenomics" app with other collected data. This application is able to correlate the Group, Visit and Source of each sample, with the output of the "16S Metagenomics" app. The "16S Statistical Analysis" app, provides the correlated data in CSV format (compatible with R, SAS, Excel, etc.) along with the following analyses:

1. Quality analysis on the number of classified reads per sample
2. Normality tests of the data
3. Alpha diversity analysis between visits and between groups
4. Statistical analysis on each organism's representation:
 - 4.1. Per group between visits
 - 4.2. Per visit between groups
 - 4.3. Per visit between case control groups
 - 4.4. Per visit change between case control groups

2 Input

In the following figure, the main user interface of the app in BaseSpace is illustrated.



16S Statistical Analysis v1.0.0
STREMBLE Ventures Ltd

Analysis Name: 16S Statistical Analysis 03/10/2017 4:46:50

Samples' info Google Drive share link:

Name of the Control Group:

Significance Level: 0.05

Aggregated Summary Files:

Save Results To:

Each input parameter will be explained in the following subsections.

2.1 Analysis Name

This input parameter defines the name that will be used by Illumina's BaseSpace to store the analysis that will be performed. By default, an automatic analysis name is created, which is the app name along with the current date and time.

2.2 Samples' Info Google Drive share link

2.2.1 Samples' Info File Description

The Sample's Info file, is a supplementary Comma Separated Value (CSV) file, which is used for linking each sample with additional information. An example of such a file is shown in the following table and can be downloaded from <http://www.stremble.com/downloads/samplesInfo.csv>.

SampleId	Id	Group	Visit
s1	1	G1	1
s2	1	G1	3
s4	2	G2	1
s5	2	G2	2
s6	2	G2	3
s7	3	G1	1
s8	3	G1	2
s9	3	G1	3

s11	4	G3	2
s12	4	G3	3
s13	5	G3	1

The first row has the header of the table and must not be changed, whereas the rest of the rows contain the actual data. The file contains the following four columns:

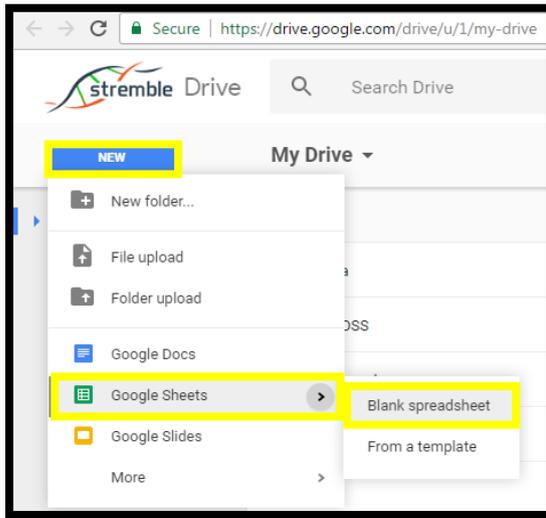
1. **SampleId:** This column has the id of the samples that we want to include in the analysis and *can contain alphanumeric values* (e.g. s1, 1, a_3-1 etc). This has to be identical to the SampleID used in BaseSpace sequencing runs as it will be used as the key to merge the two databases.
2. **Id:** This column has the id of the source of the sample. In the case of clinical experiments the Id of the subject from which the sample was taken. In the case of environmental data this could be the Id of a specific location sampled at multiple time intervals. This allows identifying the changes in the samples of subjects/sources between visits. *The id can contain alphanumeric values.*
3. **Group:** This column contains the Group that the sample belongs to. At least two groups are needed to perform the analysis. The group *can contain alphanumeric values.*
4. **Visit:** This column defines the visit number in which the sample was collected. In the example shown in the table above, there are 3 visits. This column *can contain only consecutive numeric values starting from 1*. For example if a study had 5 visits in total but samples were only collected on visits 3 and 4, then visit 3 should be denoted as visit 1 and visit 4 should be denoted as visit 2 in the samples info file. As shown in the example table above, it is ok if subjects do not have samples in all of the visits (e.g. Subject with id 1 does not have a sample for visit 2). *Visit 1 is treated as the baseline in some of the analyses.*

2.2.2 [Creating the Google Drive Share Link](#)

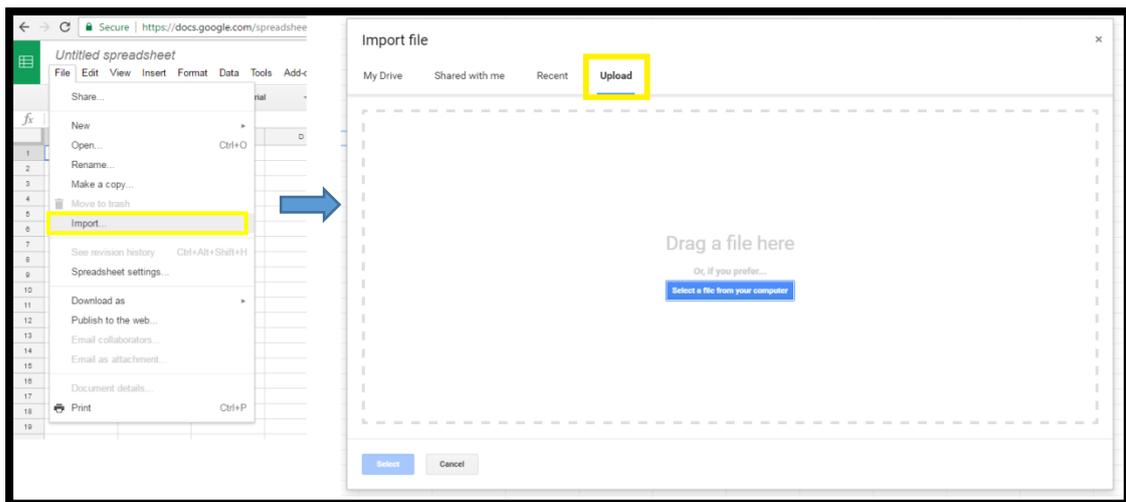
BaseSpace does not support importing Comma Separated Value files (CSV) in its platform. As a work around, we are using Google Drive share links to include such a file in this analysis.

To create the share link you need to follow these steps:

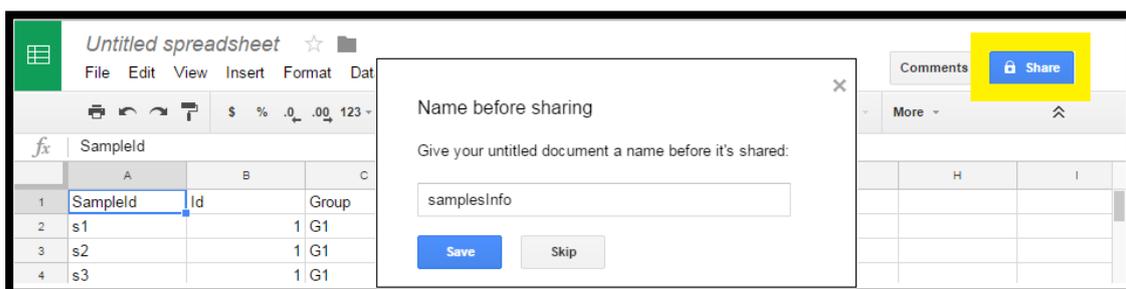
1. Go to your Google Drive and select New -> Google Sheets -> Blank spreadsheet



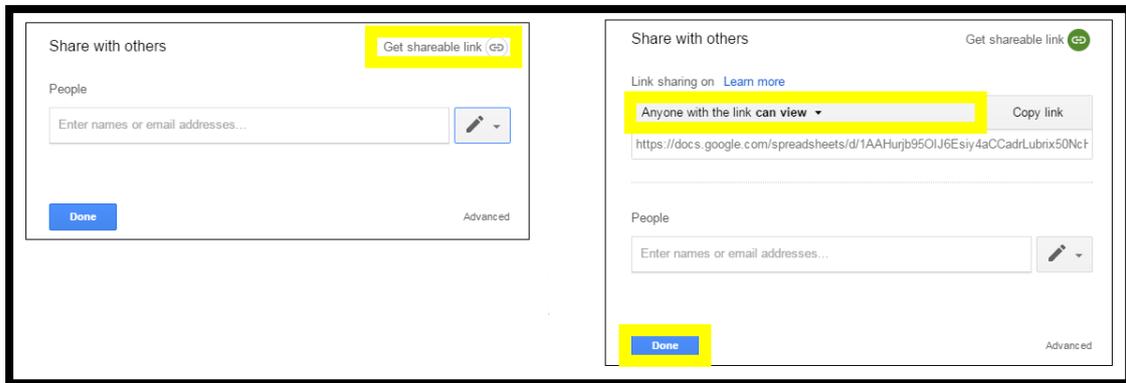
2. Import your samples info file from File->import and then select the Upload tab. Then follow the instructions to select the file you want to import.



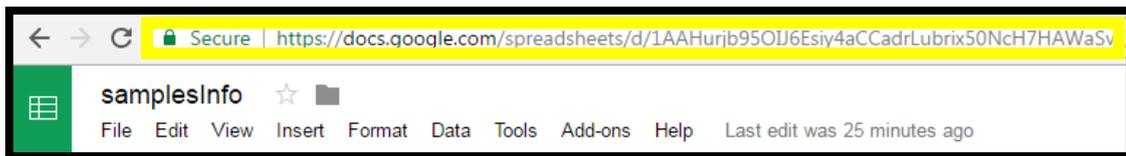
3. Select the Share button on the top right and give a name to your document if prompted.



4. Select the “Get shareable link” and select the “Anyone with the link can view” option and press “Done”.



5. Copy the Link from your browser and paste it in the 16S Statistical Analysis App.



2.3 Control Group

This input parameter defines the name of the control (or baseline) group. For example in a study where group 1 takes a specific treatment, whereas group 2 does not take any treatment, group 2 would be considered as the control group. This parameter can be alphanumeric.

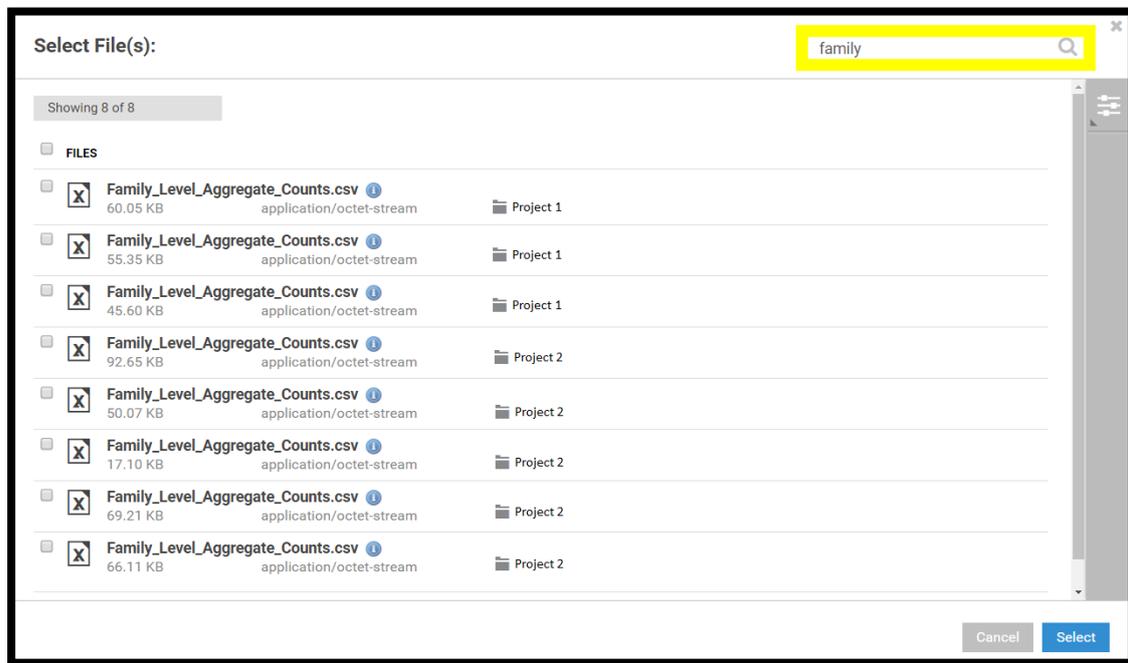
2.4 Significance Level

This input parameter defines the significance level that will be used in the statistical tests (default is 0.05).

2.5 Aggregated Counts Files

The app takes as input the *Aggregated Counts* files from Illumina's BaseSpace “16S Metagenomics” app. Our app supports the inclusion of multiple Aggregated Counts files. Additionally if your Aggregated Counts files include extra samples that are not in your Samples' Info files (section 2.2.1), then our app will automatically ignore those extra samples from the analysis. To select the Family level aggregated counts of the “16S Metagenomics” app, you can write “Family” in the search bar and select the

files from the projects of interest. You can only select Aggregated Counts of the same level (i.e. all of Genus level).



2.6 Save Results To

In this parameter you select the project in which you want to save the analysis results.

3 Output

Sample results can be downloaded from the following link <http://www.stremble.com/downloads/sampleResults.zip>. The report is in the report.html file.

3.1 Report

The report includes figures and tables of the statistically significant results of all analyses. It is in html format and can be viewed within your web browser.

3.2 Analysis files

The merged data files of the OTUs and the samples' information are provided as output in CSV format.

3.3 Figures and Tables

The complete set of figures and tables of the analysis are included in PDF and in CSV format. These files also include the non-statistically significant results.

4 Support

If you require additional information, please contact us at support@stremble.com