StatFingerprints

Version 2.0

Processing and statistical analysis of molecular fingerprint profiles

May 26, 2010

Rory J. Michelland and Laurent Cauquil
StatFingerprints@gmail.com

Welcome to StatFingerprints

This program is a free package for the R statistical program with a user-friendly graphical user interface (GUI). It is intended to help microbial ecologists to analyse fingerprint profiles. It provides procedures both to process fingerprint profiles and to perform numerous univariate and multivariate statistical analyses on them. StatFingerprints is also able to plot fingerprint profiles and several graphical results in 2 or 3 dimensions. It supports import and export of all ASCII files, a format easily writable or readable with text editors, plus the ability to convert FSA files from an ABI Prism sequencer into ASCII files. For advanced use, all procedures can be executed from the R prompt.

The processing part contains procedures:

- to align fingerprint profiles
- to delete background under peaks and carry out parameterisation
- to homogenise the baseline between fingerprint profiles
- to normalise fingerprint profiles with 3 algorithms
- to transform into presence/absence the raw fingerprint profiles with parameterization possibilities for peak detection
- estimate 7 ecological diversity indices from fingerprint profiles and parameterize peak detection
- to correct defective peaks

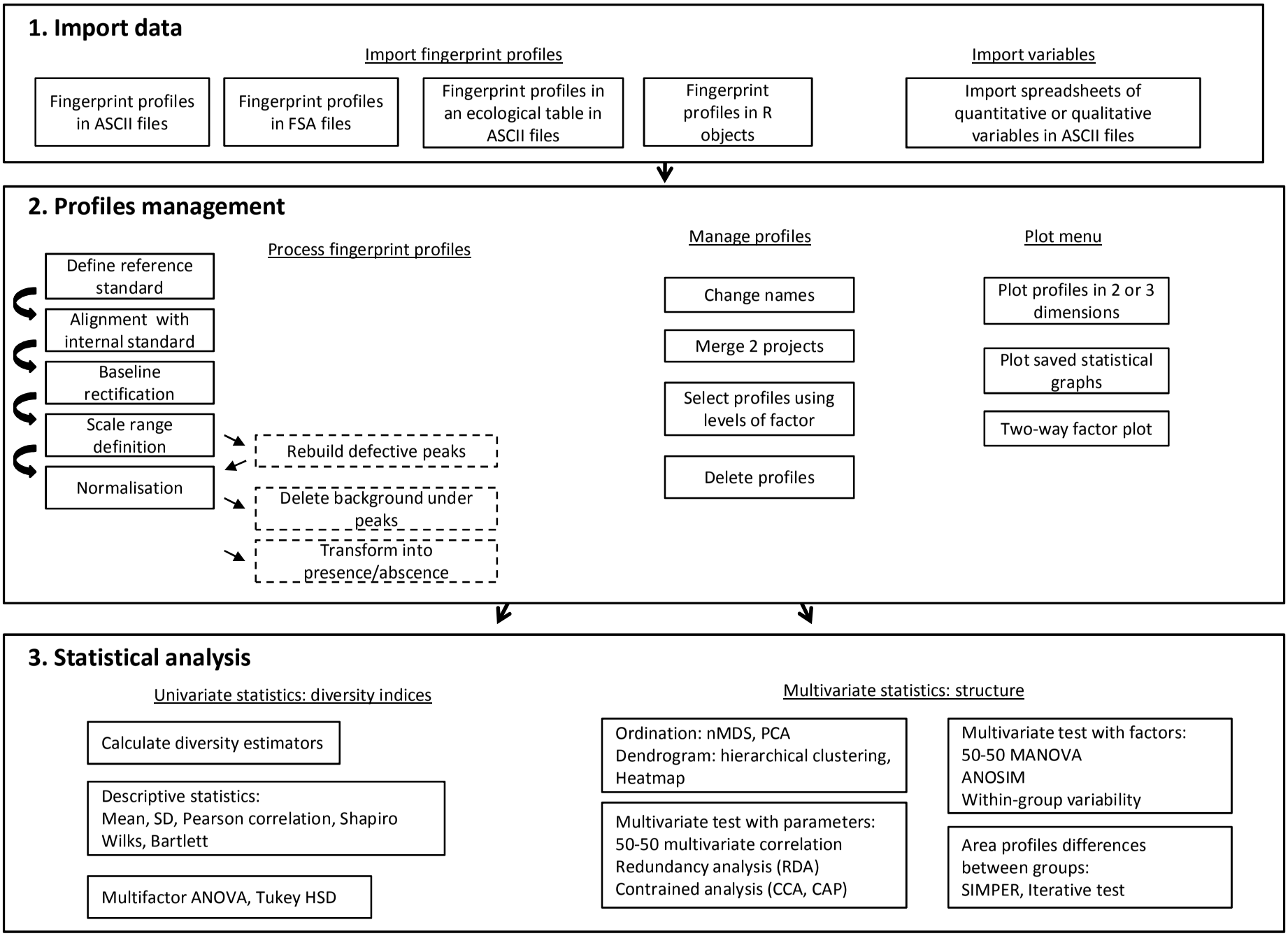
The univariate statistical part contains the following possibilities:

- Shapiro Wilks test
- Bartlett test
- Multifactor ANOVA
- Pearson correlation

The multivariate statistical part contains the following possibilities:

- Random start non-metric MultiDimensional Scaling (nMDS) with 2- or 3- dimensional dynamic plotting and the ability to add qualitative and quantitative variables
- Principal Components Analysis (PCA) with 2- or 3- dimensional dynamic plotting and the ability to add qualitative and quantitative variables
- Hierarchical clustering with 13 similarity measures and 7 plotting algorithms
- Heatmap with 13 similarity measures
- 50-50 Multivariate ANalysis Of VAriance (50-50 MANOVA)
- Analysis Of SIMilarity with the global and pairwise algorithms

- Analysis of within-group variability
- SIMilarity PERcentages Procedure (SIMPER)
- Iterative tests (T test, Mann Whitney and Fisher's exact test) to define areas presenting differences along fingerprint profiles
- 50-50 multivariate correlation
- Redundancy analysis (RDA)
- Constrained correspondence analysis (CCA)
- Constrained analysis of proximities (CAP)


1. Import data
Import fingerprint profiles
Fingerprint profiles in ASCII files
Fingerprint profiles in FSA files
Fingerprint profiles in an ecological table in ASCII files
Fingerprint profiles in R objects
Import variables
Import spreadsheets of quantitative or qualitative variables in ASCII files
2. Profiles management
Process fingerprint profiles
Define reference standard
Alignment with internal standard
Baseline rectification
Scale range definition
Normalisation
Rebuild defective peaks
Delete background under peaks
Transform into presence/abscence
Manage profiles
Change names
Merge 2 projects
Select profiles using levels of factor
Delete profiles
Plot menu
Plot profiles in 2 or 3 dimensions
Plot saved statistical graphs
Two-way factor plot
3. Statistical analysis
Univariate statistics: diversity indices
Calculate diversity estimators
Descriptive statistics:
Mean, SD, Pearson correlation, Shapiro Wilks, Bartlett
Multifactor ANOVA, Tukey HSD
Multivariate statistics: structure
Ordination: nMDS, PCA
Dendrogram: hierarchical clustering,
Heatmap
Multivariate test with parameters:
50-50 multivariate correlation
Redundancy analysis (RDA)
Contrained analysis (CCA, CAP)
Multivariate test with factors:
50-50 MANOVA
ANOSIM
Within-group variability
Area profiles differences between groups:
SIMPER, Iterative test

Contents

List of Figures

1 Introduction to StatFingerprints

1.1 Installation and running

R works on a wide variety of UNIX, Windows and MacOS platforms but due to the GUI and some graphical function the package StatFingerprints may not function correctly under other system than Windows. Before installing StatFingerprints you first need to install R [1]. Sources, binaries and documentation for R can be obtained via CRAN, the Comprehensive R Archive Network.

To install the StatFingerprints program, write at the R prompt:

```
> install.packages("StatFingerprints",dependencies=T)
```

Select the CRAN mirror and the package. Then the package StatFingerprints and its dependencies will be downloaded and installed. Once you start R, you don't need to re-install the StatFingerprints package but you need to load it by writing at the prompt:

```
> library(StatFingerprints)
```

Once loaded, you can write at the prompt first one of the following line command to run StatFingerprints:

```
> StatFingerprints()
```

or

```
> sf()
```

1.2 Presentation of the main window

The main window of the StatFingerprints program is divided in different parts (Fig. 1).

In the menu bar on the top, all procedures of the program can be launched using one of the 7 menus. In the main part of the window the following information can be found:

- The name and the path of the current project
- The current work directory of R
- The data loaded: profiles, qualitative and quantitative variables
- The status of each step of the fingerprint profile processing menu can be checked
- The diversity indices

The Change current directory button allows to change the current directory of R. Each Edit button allows to view data in an edit window and eventually change names. The Export matrix button allows to export any of matrices already computed during the processing of the fingerprint profile.

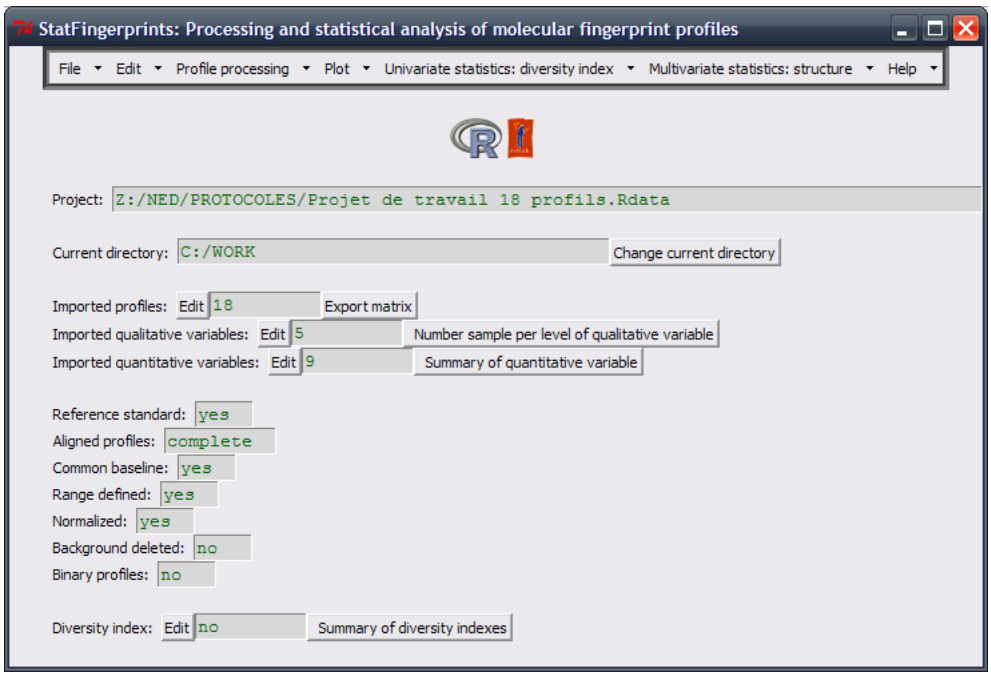


Fig 1: Main window

The NUMBER OF SAMPLES PER LEVEL OF QUALITATIVE VARIABLE button gives information on the R console about the number of observation in the levels of each factor which will be taking into account for further hypothesis-driven statistical testing. The SUMMARY OF QUALITATIVE VARIABLES button gives a summary of the qualitative variables in the R console. The SUMMARY OF DIVERSITY INDICES button gives a summary of the diversity indices in the R console.

2 File menu

2.1 Convert FSA files and import

Raw fingerprint profiles from an ABI Prism 310 or 3100 sequencer (Applied Biosystems) are available as FSA files which can be automatically converted to ASCII files and then loaded into STATFINGERPRINTS. The conversion step of this procedure needs the free program DataFileConverter (Applied Biosystems) available here

When running this procedure, the first step is to specify the folder where DataFilesConverter has been stored and then the folder where the FSA files are stored. During this step FSA files are converted into ASCII and stored in a folder named txt located in the FSA file folder. This step can take several minutes depending on the number of fingerprint profiles and your computer system. In the second step, channels containing the community profiles and the internal standards must be specified. This last step can easily be done using the

HOW TO CHOOSE COMMUNITY AND INTERNAL STANDARD LOCATION button which allows visualising each channel of the first profile (Fig. 2).

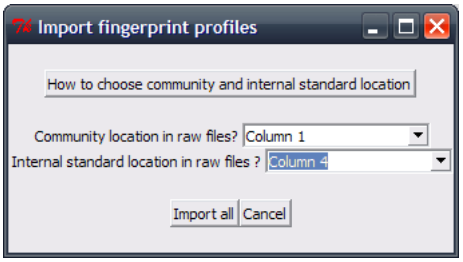


Fig 2: Import fingerprint profiles

2.2 Import fingerprint profiles in ASCII files

Fingerprint profiles in ASCII format can be imported with the StatFingerprints package as it can set the following parameters: field separator, decimal separator, and occurence of header. The easiest solution is to export files in CSV format using a text editor or spreadsheet program.

The first step is to select your ASCII file (at least two) and the second is to state the field separator, the decimal separator, the occurence of headers and the columns containings the community and internal standard profiles. Specifying the columns containing the community and the internal profiles can easily be done using the How to choose community and internal standard location button.

2.3 Import ecological table (ASCII)

Fingerprint profiles can also be imported using an ecological table as an ASCII file. An ecological table contains each microbial distribution by row or column. Internal standards for each community are not included in the table, so fingerprint profiles cannot be aligned.

To import an ecological table, first choose your file location and then complete the structure of the table (fingerprint profiles are in rows or columns), the field separator and the occurrence of headers in columns and rows (Fig. 3).

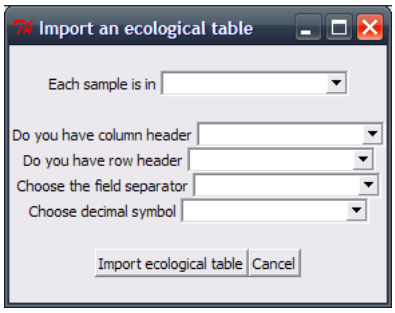


Fig 3: Import an ecological table

2.4 Import R objects

With this function you can directly import into a project the matrix with your ecological communities and the factor and parameter when they are already transform in R objects. R objects must be loaded in the R environment before using this function.

2.5 Imported variables

Quantitative or qualitative tables in ASCII files are imported with this procedure. Be careful to separate your quantitative and qualitative data in two files as importing files containing both quantitative and qualitative variables is not supported. Special characters and special formats (bold, italic, underlined etc.) are also not supported. The Missing values must be indicated as *NA* for not available. To import a table of variables, first select your file location and then specify the field separator, the decimal separator, the occurrence of header and the type of the variables: qualitative (factor) or quantitative (parameter).

2.6 Load project and save project as

We advise you to save your data regularly. All data and objects created are stored in an Rdata file (see 8.1). The LOAD procedure allows a saved project to be loaded. The SAVE PROJECT AS procedure is the classical procedure to save a project in a specified directory.

3 Edit menu

3.1 Change names and profiles

Names of fingerprint profiles can easily be changed using this procedure. First double click the fingerprint profile, next write the new name and hit enter to validate. The name of the fingerprint profile will be immediately updated.

3.2 Merge two projects

This procedure allows two projects to be merged. Be aware that it is only possible to merge projects with fingerprint profiles with the same length or that are aligned in the same way (ie. same internal standard). This procedure requires two consecutive steps (Fig 4.):

- import the two projects. The LOAD TWO PROJECTS button produces two consecutive exploratory windows to select them.
- merge the two projects.

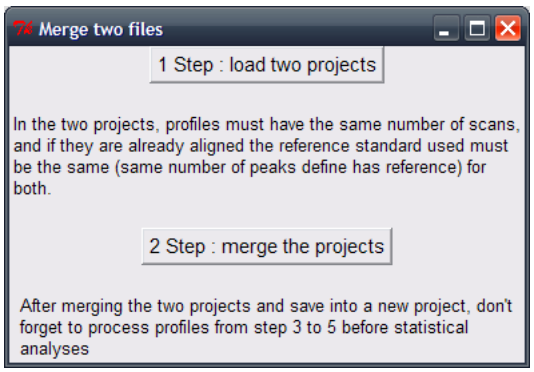


Fig 4: Merge two files

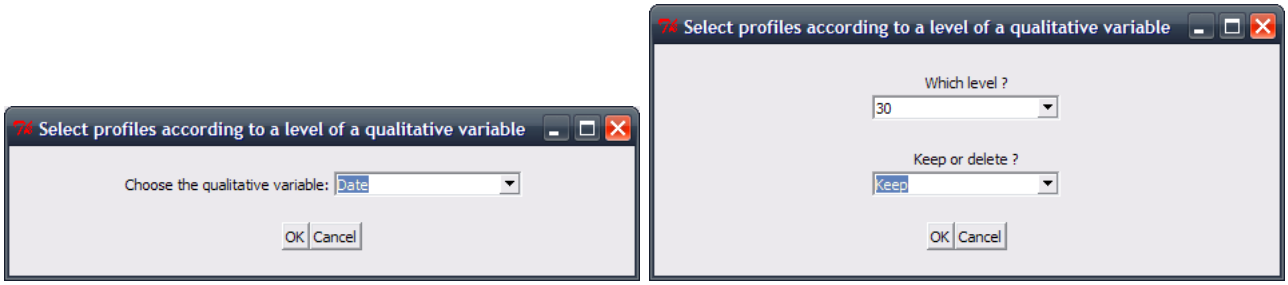
After merging, processing profiles must be done again from step 3 of the PROFILE PROCESSING menu.

3.3 Delete profiles within the project

This procedure deletes one or several fingerprints profiles in your project. The corresponding qualitative and quantitative variables are also updated. This procedure should be use with care as the deleted profiles will be completely erased from the project.

3.4 Select profiles using levels of factor

This procedure allows deleting a group of profiles selected by the level of a factor (Fig. 5).



(a) Select a factor (b) Select a level

Fig 5: Select profiles using levels of factor

4 Profile processing menu

Before statistical analysis, fingerprint profiles have to be aligned. Then several other treatments can be applied to the aligned fingerprints profiles to make them comparable.

4.1 Define peaks using your own reference standard

A reference standard is required to align fingerprint profiles. Internal standards of each fingerprint profile were be aligned on this reference standard. This procedure consists of 2 steps: defining and then validating peaks of the reference standard that will be use in the alignment process.

Defining peaks:

First, the x-position at the maximum height of each peak in your reference standard must be entered in the dialog box (each abscissa must be separated by a comma). If necessary, you can use the HELP TO DEFINE PEAKS OF THE REFERENCE STANDARD button. This button allow you to select peaks from your first sample:

- Zoom the area containing peaks to select by clicking on the top left corner and on the bottom right corner of the area (Fig. 6).
- In the next window, left click precisely on the the x-position at the maximum height of each peak you want to keep as a reference. Right click to stop (Fig. 7).

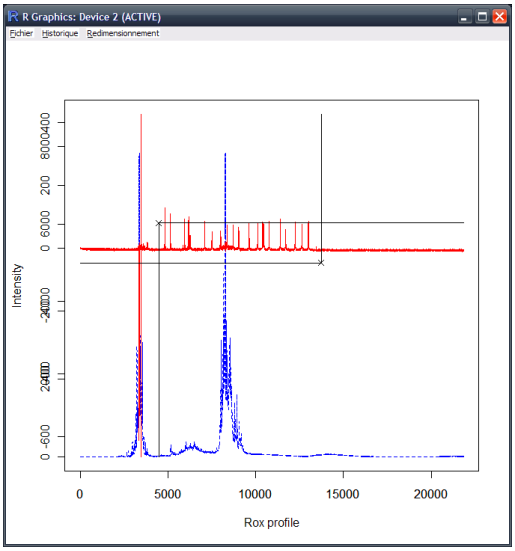


Fig 6: Zoom area

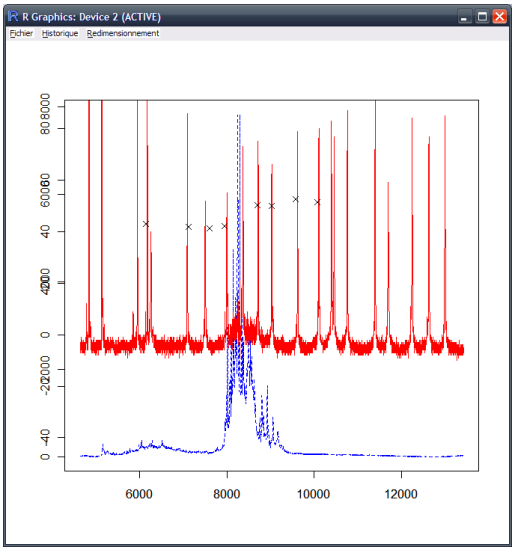


Fig 7: Select peaks

The x-positions of each selected peak are automatically printed in the dialog box (Fig. 8).

Validating:

As As molecular fingerprints frequently only occupy part of the range of the reference standards, the alignment is not improved by basing it on the entire set of reference standards, thus alignment on the whole peaks of the standard is useless. Select the area where alignment should be done. In the window, peaks are symbolize by vertical red lines, left click to the left of the first peak and to the right of the last peak of the area. The peaks used for the alignement will turn to green lines (Fig.9).

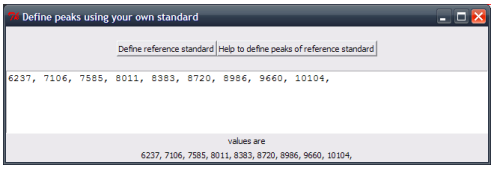


Fig 8: Peaks selected

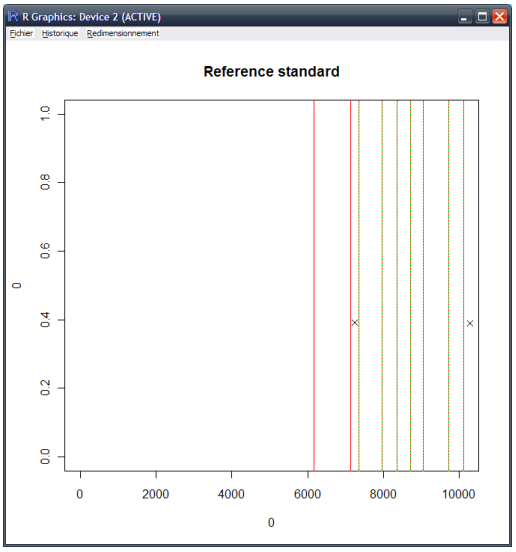


Fig 9: Select range peak

4.2 Use peaks of ROX defined in the file Rox.ref

Once x-values for reference peaks are validated you can store them in a txt file named "Rox.ref" and located in the StatFingerprints package directory. This function will load the peak positions automatically and avoids the definition and validation of them for each project.

4.3 Align profiles one by one

Peaks of the reference must be defined before using this procedure (see 4.1). This procedure uses an algorithm of cubic spline interpolation of fingerprint profiles which consists of *"an exact cubic spline through the four scans at each end of the scans"*[4]. Basically, the algorithm aligns peaks from the internal standard of the processed fingerprint profile with those of the reference standard and then it applies the same computed transformations to the community signal of the processed fingerprint profile (Fig. 10). Therefore, alignement is only a matter of computing each fingerprint profile and must be applied in the same way for the entire set of fingerprint profiles.

Alignment consists in 5 consecutive steps:

- Select the fingerprint profile to align
- In the open graph, zoom in the area corresponding to the peaks previously defined in the reference standard. To do this, use two left clicks as before on either side of the area to be marked (Fig. 11(a)).
- In the following R graph, select the area containing the same peaks as those defined in the internal reference standard by clicking before the first peak and after the last peak corresponding to those defined in the reference standard. The y-value of the first click must be less than the maximum y-value of all peaks (Fig. 11(b)).
- False peaks due to artefacts may occur along the internal standard. This step allow to delete them by clicking before and after each false peak. Once all

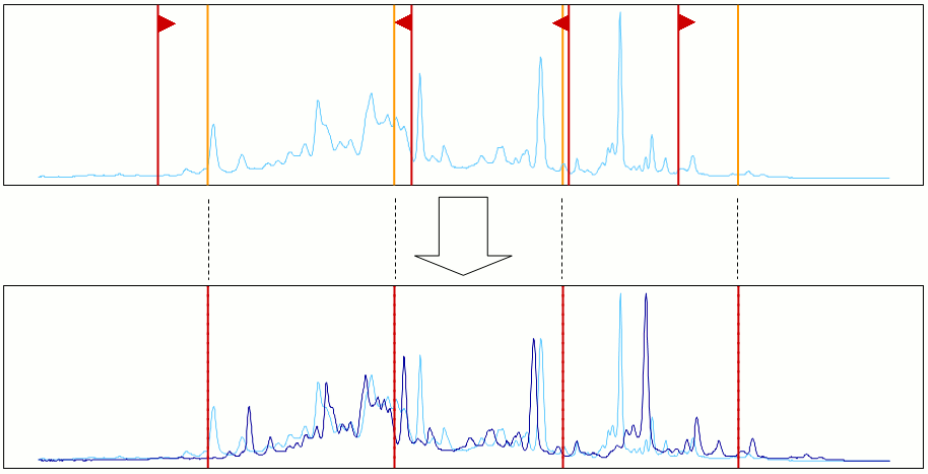
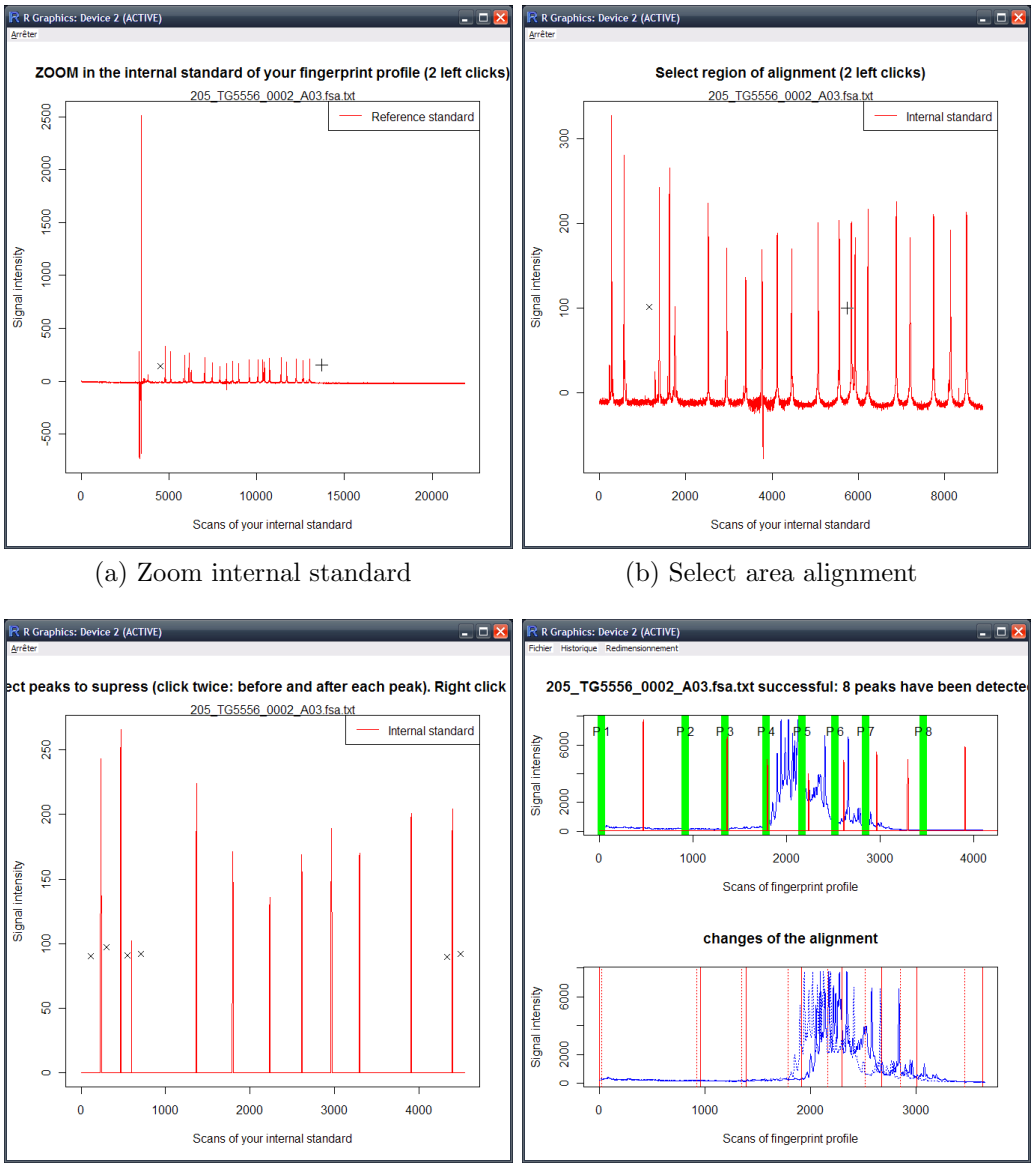


Fig 10: Scheme of the alignment. Peaks of the reference standard are in orange

false peaks have been selected, go to the next step using a right click and the stop option. If no false peak is present, go to the next step by left clickings twice in an area containing no peak, and exit using a right click and the stop option (Fig. 11(c)).

- Finally, the open R graphic window represents the result of the alignement. The number of detected peaks on your fingerprint profile must correspond to the number of peaks of the reference standard in order to be aligned (Fig. 11(d)). If the two numbers of peaks differ the alignment is not performed and a warning box appears. Once correctly aligned *"Align"* will be added to the name of the fingerprint profile.



(a) Zoom internal standard

(b) Select area alignment

(c) Select peaks to suppress and validate

(d) Final alignment

Fig 11: Alignment process

4.4 Check quality of the alignment

This procedure plots all the aligned fingerprint profiles in three dimensions. The alignment can be visually checked (Fig. 12).

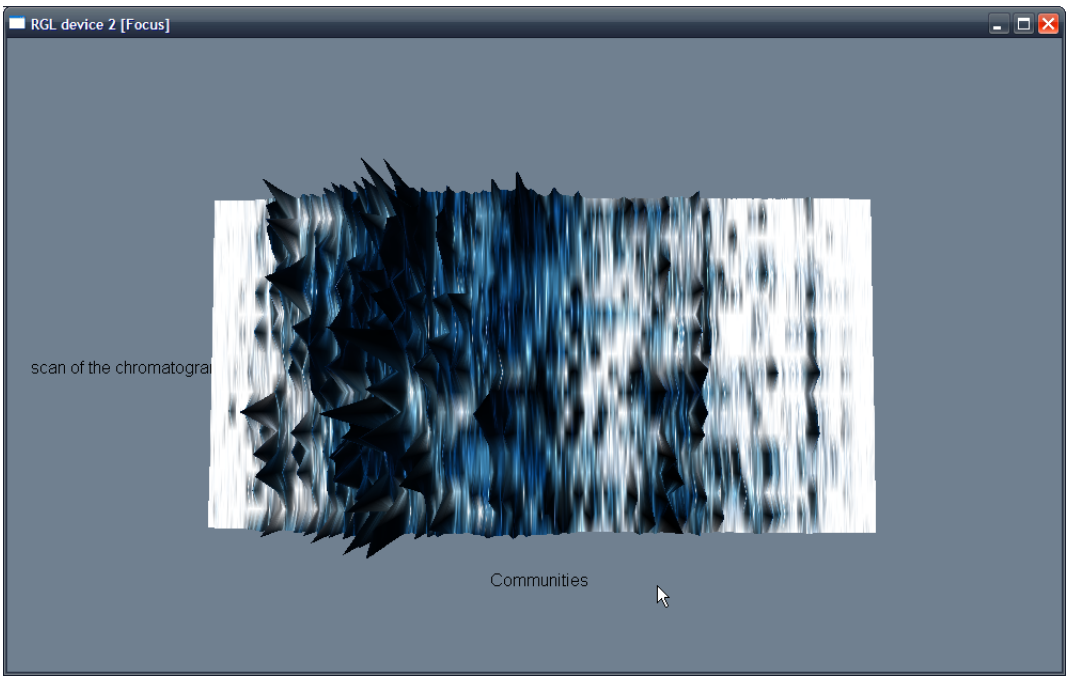


Fig 12: Check alignment

4.5 Define a common baseline for all fingerprints profiles

As illustrated in Fig. 13 (top), fingerprint profiles often have their baselines non-aligned (not the same y-value) or some fingerprint profiles are not perfectly parallel to the x-axis. To align as well as establish a horizontal baseline for all fingerprint profiles, two left clicks are needed, the first just before the beginning of the signal and the second just after it (Fig. 13).

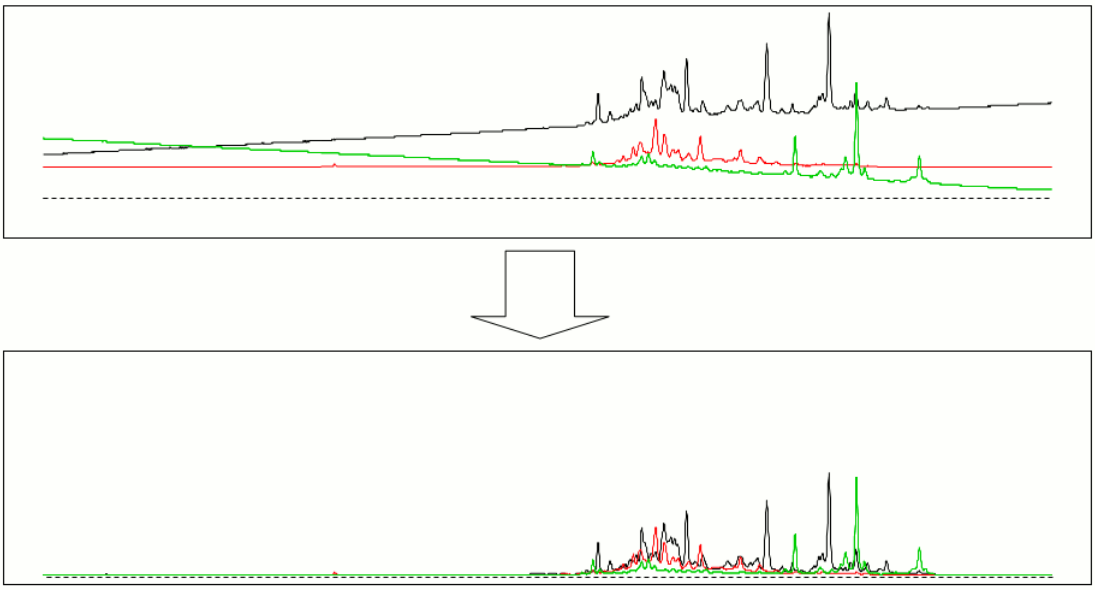


Fig 13: Baseline method

4.6 Define the range of the fingerprint profiles

This procedure allows only the area corresponding to the microbial community to be kept. On the R graphic, left click just before and after the area of interest of all profiles (Fig. 14).

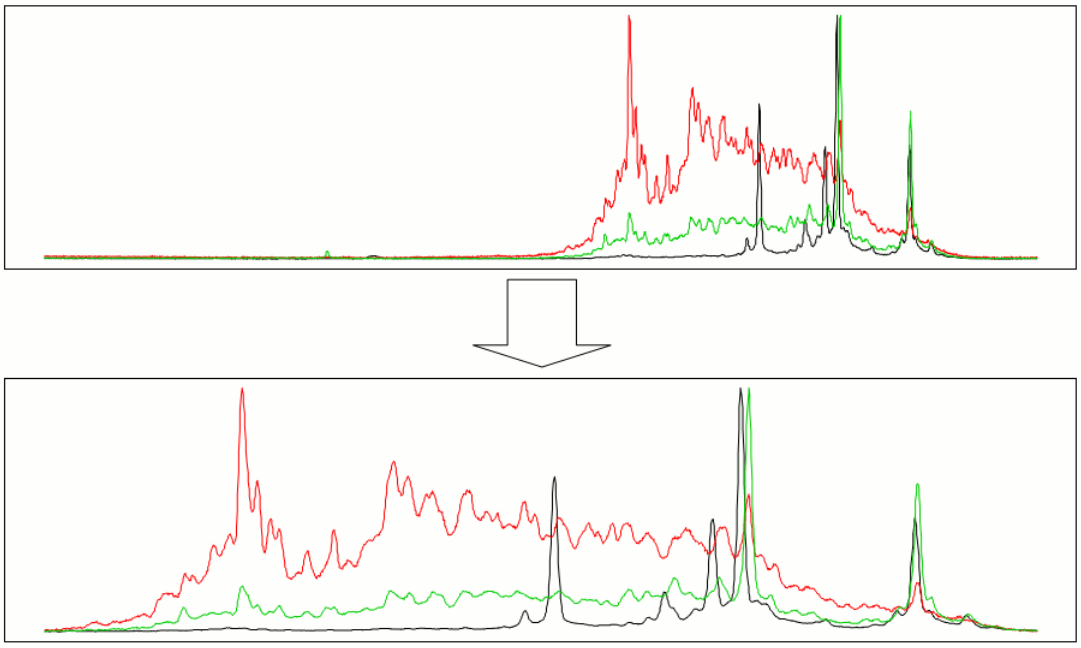


Fig 14: Define the range of the fingerprint profiles

4.7 Rebuild peaks of defective fingerprint profiles (optional)

Care must be taken when using this procedure and generating clean raw data by re-running the molecular analysis should be preferred. A profile can be defective if it contains one or more saturated peaks, it can be checked using the function "Plot profiles in 2 dimensions" (see 5.1). In this type of profile, peaks are truncated because of the signal saturation. Data sets with numerous fingerprint profiles containing one or more defective peaks indicate raw data of poor quality. This could be caused by too small amounts of DNA provided to the sequencer, quality of the gel in the sequencer, adjustements of the sequencer etc.). However, a saturated peak can occur in a few of fingerprint profiles due to, for example, low diversity with one dominant operational taxonomic unit (OTU). In this particular case, the peak can be rebuilt using this procedure. Generally, the shape of each peak in the fingerprint profiles can be described using the same mathematical function. Consequentely peaks can be corrected accurately by this function which calculates the equation of a reference peaks and applies it to the defective peak (Fig. 15).

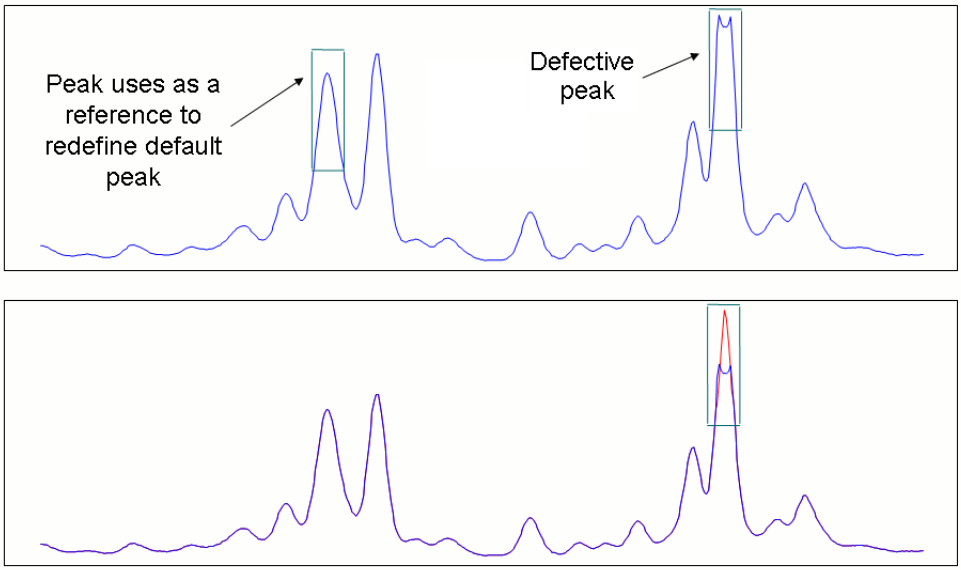
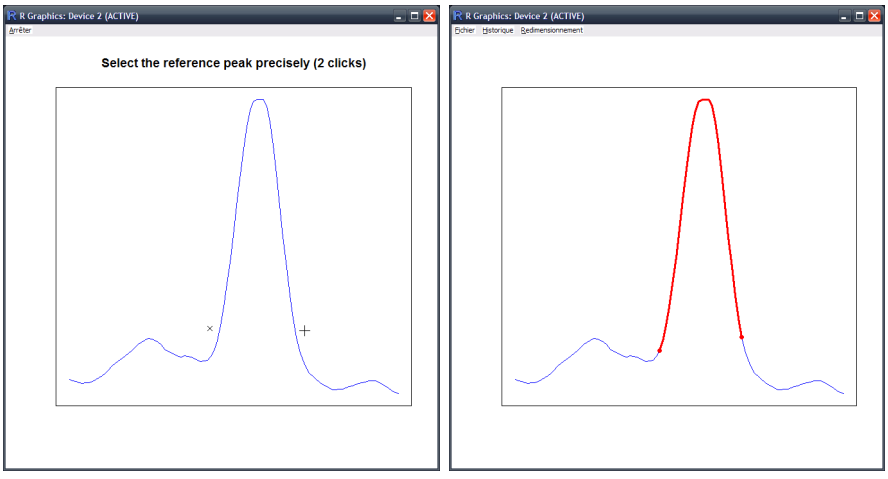


Fig 15: Rebuild peaks of defective fingerprint profiles

This procedure consists of six steps:

- Select the fingerprints profile with defective peaks and press the COMPUTE PEAK MODIFICATION button.
- Zoom in (using two left peaks) an area presenting a peak similar to the defective one. This peak will represent the reference peak (Fig. 16(a,b)).
- Select precisely the start and the end of the reference peak using two left clicks.



(a) Select a reference peak (b) Resulting peak

Fig 16: Selecting a reference peak

The equation describing the reference peak will be defined as a model to rebuild the defective peak.

- Zoom in the area of the defective peak using two left clicks.
- Select precisely the start and the end of the defective peak using two left clicks.

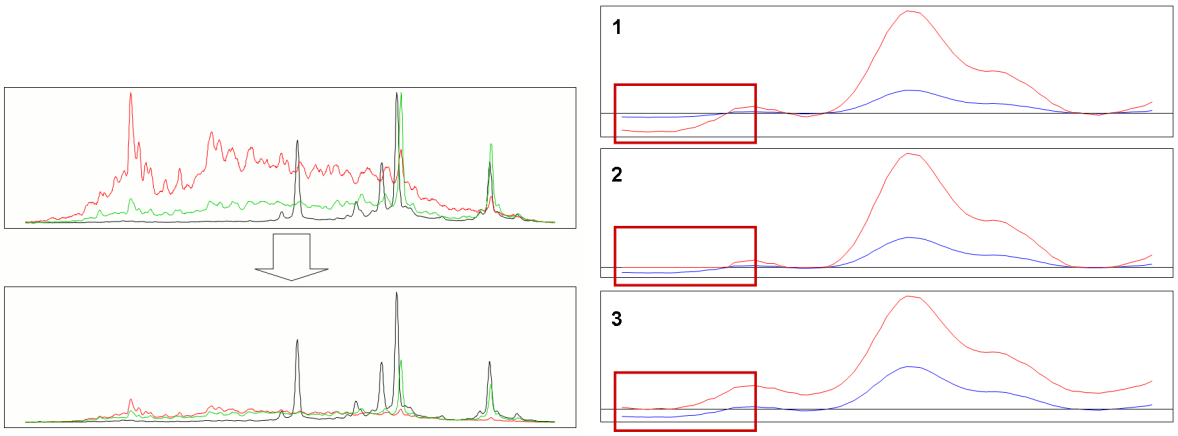
- Check the quality of the new peak compared with the old one. If you are satisfied, save it by pressing the SAVE THE MODIFIED PEAK button; otherwise restart the procedure.

4.8 Normalise the area under the profiles

As illustrated in Fig. 17 (a, top), the area under each fingerprint profile is usually not the same. To efficiently compare fingerprint profiles, it is strongely advised to normalise them so that the new area under the curve is equal to one (Fig. 17(a)).

Three different algorithms are provided for normalisation (Fig. 17(b)):

- Normalise area under curve ignoring negative values.
- Convert all negative values into 0 and then proceed to normalise.
- Take the minimum y-value of the curve and substract the absolute of this value from all values whithin the curve (recommended).



(a) Normalisation methods (b) Normalisation

Fig 17: Normalise the area under the profiles

4.9 Delete background under the profiles

Optially, the background under all fingerprint profiles can be eliminated using the "rollball" algorithm. The algorithm removes the area under the trajectory of a virtual ball, rolling under the signal of the community of the fingerprint profile (Fig. 18). A threshold must be specified to determine the radius of the virtual ball to delete more or less background. The most suitable radius for your fingerprint profiles can easily be defined using the HELP TO DEFINE THE ROLLBALL button. This procedure plots your first fingerprint profile before and after deleting the background using the specified radius of the rollball.

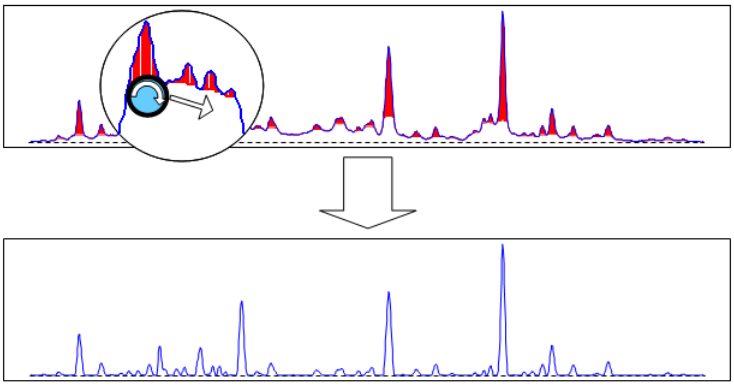


Fig 18: Delete background with the rollball method

4.10 Transform profiles into presence/absence profiles

This procedure transforms quantitative fingerprint profiles into binary fingerprint profiles. Binary fingerprint profiles are required when using binary proximity measures (Jaccard, Dice-Sørensen, Ochiai, Steinhaus). It means that each scan value of the fingerprint profiles is transformed into 1 to 0 according to whether it is located within a peak or not. The algorithm used to detect peaks can be fully parameterized to best fit each fingerprint profiles set. Parameterized features are: the radius of the rollball, the threshold, the width of the peak area, and the interval size (for details of the last 3 features, see 6.1) The value of these features can easily be defined using the HELP TO DEFINE CHARACTERISTICS OF PEAK DETECTION button. As a result (Fig. 19), the R graphic window displays a plot with each fingerprint profile drawn as a dotted line, the black horizontal segments representing detected peaks (value equal to 1). Once this procedure has been executed, the previous fingerprint profiles with the quantitative information can be easily recovered by executing the normalisation procedure (see 4.8).

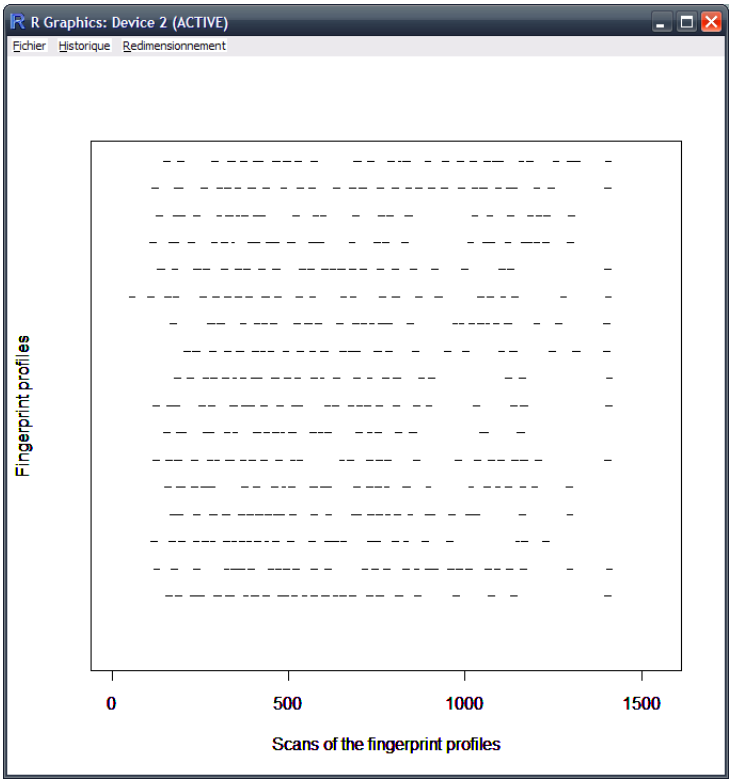


Fig 19: Transform profiles into presence/absence profiles

5 Plot menu

5.1 Plot profiles in 2 dimensions

This procedure plots the community signal of one or several fingerprint profiles. It automatically detects the last step of fingerprint profile processing (from import to the transformation into presence/absence fingerprint profiles) and thus always plots the selected fingerprint profiles in its latest state.

5.2 Plot profiles in 3 dimensions

This procedure plots the community signal of all the fingerprint profiles in 3 dimensions. It automatically detects the most recent step of fingerprint profile processing (from import to the transformation into presence/absence fingerprint profiles) and thus always plots fingerprint profiles in their most recent state. For the alignment, the procedure plots the aligned fingerprint profiles only if all fingerprint profiles are aligned. The plot can be rotated using left clicks and zoomed using right clicks. The plot can be saved as a file in the *.png format using the SAVE PICTURE menu.

5.3 Plot saved nMDS or PCA in 2 dimensions

This procedure allows results of ordinations already computed (nMDS or PCA) to be plotted in two dimensions. It can be used only after having performed and saved an ordination plot (see 7.1 or 7.2). The following features can be specified: selection of axes, labels, colour of points according to a qualitative variable, regression line

according to a quantitative variable. The firs step is to select a saved ordination and the second is to specify features of the plot. The plot can be saved as a file in the *.PNG using the SAVE PICTURE menu.

5.4 Plot saved nMDS or PCA in 3 dimensions

This procedure allows results of ordinations already computed (nMDS or PCA) to be plotted in three dimensions with dynamic control. It can be used only after having performed and saved an ordination plot (see 7.1 or 7.2). The following features can be specified: selection of axes, labels, colour of points according to a qualitative variable, regression line according to a quantitative variable. The firs step is to select a saved ordination and the second is to specify features of the plot.

5.5 Two-way factor plot

This procedure computes and plots basic descriptive statistics (mean and standard deviation) of either a diversity index or a quantitative variable according to the levels of one or two qualitative variables (Fig. 20). Four different kinds of plots are available: line, line and SD, boxplot, histogram.

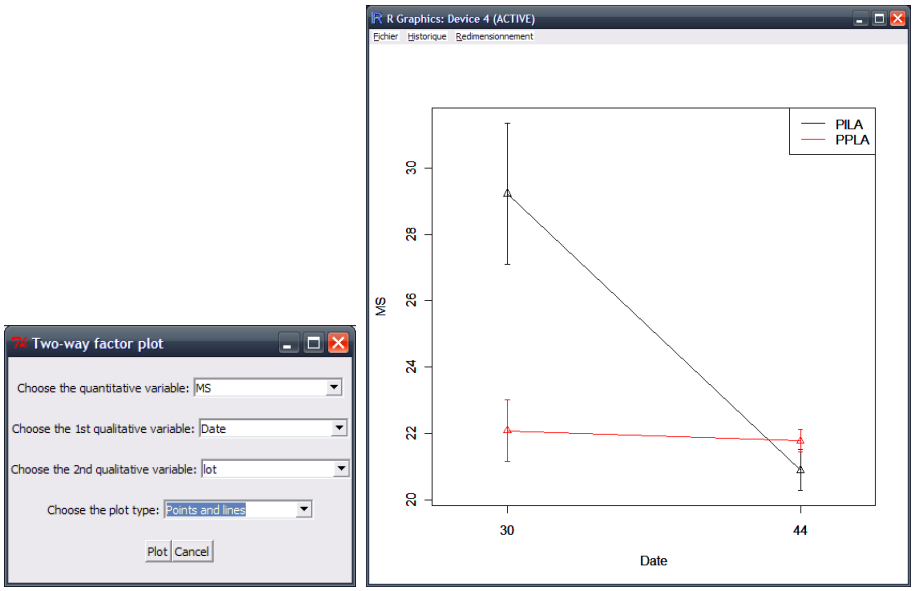


Fig 20: Two-way factor plot

6 Univariate statistics: diversity index menu

6.1 Compute diversity index

Estimating a diversity index consists of summarizing a complex community represented by a fingerprint profile as a single value. Various diversity indices are calculated by taking account either the number of peaks of the fingerprint profile or the number of peaks and their relative abundances (area or height under each peak).

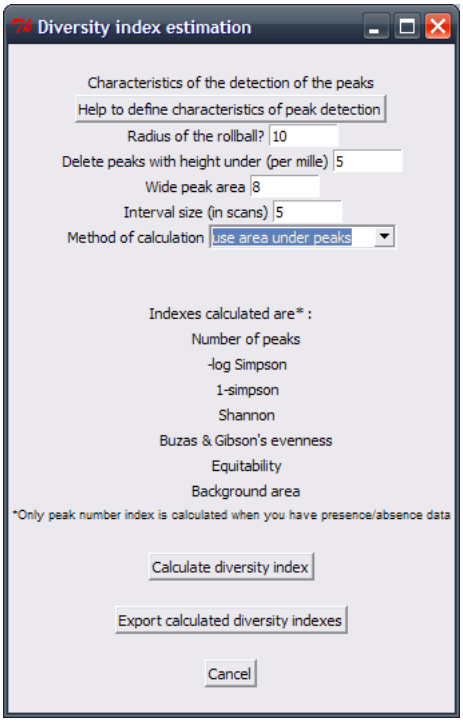


Fig 21: Diversity index estimation

The following diversity indices are available [8]:

- Peak number S (often named Richness)
- The minus logarithm of Simpson

$$D = -\log \sum \left(\frac{a_i}{\sum a_i}\right)^2$$

where a_i is the relative abundance of each peak. If normalization is performed, the minus logarithm of Simpson is calculated as

$$D = -\log \sum a_i^2$$

This index ranges from 0 (a single peak) to infinity(an infinite number of peaks of equal abundance).

- One minus Simpson

$$D = 1 - \sum \left(\frac{a_i}{\sum a_i}\right)^2$$

It ranges from 0 (a single peak) to 1 (an infinite number of peaks of equal abundance).

- The Shannon index (entropy)

$$H = -\sum a_i . \log a_i$$

where a_i is the relative abundance of each peak. It varies from 0 for communities with a single peak to high values for communities with many peaks, each with little abundance.

- Buzas and Gibson's evenness

$$\frac{\exp^{-\sum a_i . \log a_i}}{S}$$

where a_i is the relative abundance of each peak and S is the number of peaks.

- Equitability

$$\frac{-\sum a_i . \log a_i}{\log S}$$

- Background area gives the area under the curve for each fingerprint profile

Before computing diversity, the procedure needs to detect peaks and their sizes. The algorithm for peak detection can easily be parameterized to best fit your fingerprint profiles.

The following parameters can be specified:

- the radius of the rollball
- a threshold below which peaks are deleted
- the width of the detected peaks
- the interval size. This feature fixes the scanning range within which the maximum y-value is sought. The smaller the value, the more precise the result at the cost of computational time.
- the method of calculation. There are 2 methods for calculating abundance, one using the maximum height of peaks and the other using area under peaks (recommended).

The Help to define characteristics of peak detection button helps to choose the values of parameters. It plots the result of a single fingerprint profile with the given values of parameters. Once computed, all these indices are merged to the quantitative variables.

6.2 Descriptive statistics

This procedure computes basic descriptive statistics (mean and standard deviation) of either a diversity index (if they are computed) or a quantitative variable according to the levels of a qualitative variable. The normality of the distribution of a quantitative variable and the homogeneity of the variance (required assumtions for ANOVA) are also available.

6.3 Multifactor ANOVA

Multifactor ANOVA is a statistical procedure for testing the null hypothesis that a quantitative variable has the same mean cross each factors, and that there are no dependencies (interactions) between these factors. This procedure computes a type II ANOVA. The samples are assumed to be roughly normally distributed and have similar variances. If the sample sizes are equal, these two assumptions are not critical.

To build the ANOVA model, first select the quantitative variable, the two qualitative variables, specify their link (independently, interaction, independently + interaction), and press the ADD ELEMENT TO THE MODEL OF ANOVA button. The model is printed at the bottom of the Compute ANOVA window. Other quantitative variables can be added with the same procedure until the complete model is achieved. If the model has an error, it can be reset using the RESET THE MODEL OF ANOVA button. Once the model is complete, use the COMPUTE THE ANOVA button to compute the classical result of ANOVA and Root Mean Squared Error (RMSE). When the ANOVA result is significant, the Tuckey HSD retrospective test can be computed.

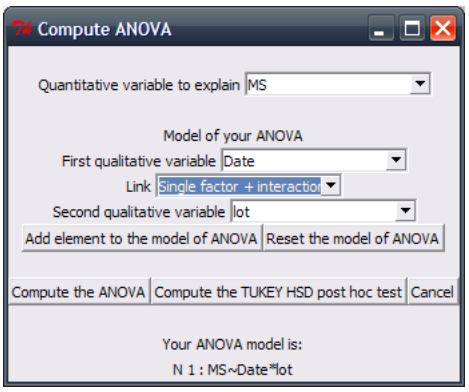


Fig 22: ANOVA function

6.4 Simple correlation

This procedure calculates correlation between two quantitative variables using Pearson's method. The result gives the Pearson R-squared, its p-value and the equation of the regression on the R console.

7 Multivariate statistics

7.1 Non-metric Multidimensional Scaling (nMDS)

Non-metric multidimensional scaling can be based on a proximity matrix computed with any of 13 supported proximity measures, as explained below. NMDS is a two-dimensional display where each fingerpint profile is represented by a single plot [3]. They are plotted so as to conform as well as possible with the rankings of the proximities between each pair of points. The degree to which the display matches the underlying distances is assessed by using Kruskal stress (in percent); the smaller the stress, the better the representation

The algorithm used in this program allows the proximity index, the number of dimensions and the number of random starts to be chosen. The greater the number of dimensions, the better is the result of nMDS and thus the lower the stress. As the random start procedure provides different plots at each computation, it is recommended to save the nMDS. To view the nMDS in 2 or 3 dimensions and to choose the axes to plot or other tools, the nMDS should have been saved and then loaded using plot procedures (see 5.1).

Four kinds of proximity measure are supported: [7] [10]

- Distances (Euclidean, Maximum, Manhattan, Canberra, Minkowski).
- Correlation with the Pearson coefficient
- Similarity with abundance (Bray Curtis, Chi-squared, Ruzicka, Roberts). These indices take into account the relative abundances, of each peak. This index ranges from 0 (no proximity) to 1 (the two fingerpint profile are identical).
- Similarity with presence/abscence (Jaccard, Dice-Søorensen, Ochiai, Steinhaus). These indices take into account only the presence (value 1) or the abscence (value 0) on each scan of the fingerprint profiles. This index ranges from 0 (no proximity) to 1 (the two fingerpint profiles are identical).

  For these indices, don't forget to transform your fingerprint profiles into presence/abscence fingerprint profiles.

7.2 Principal Components Analysis (PCA)

Principal Component Analysis (PCA) is a procedure for finding hypothetical variables (components) which account for as much of the variance in your multidimensional data as possible. These new variables are linear combinations of the original ones. In this program PCA can be centered and scaled or not. The PLOT PROPORTION OF THE PRINCIPAL COMPONENTS button which plots the respective contribution of the new components can be helpful to choose the plotting axis. As for nMDS, PCA can be saved and plotted in 2 or 3 dimensions with advanced features (see 5.1).

7.3 Compare PCA vs nMDS

This procedure compares the nMDS and the PCA ordinations with the Pearson correlation method using a Euclidean metric:

- first calculate the Euclidean distances between fingerprint profiles pairwise (initial distance matrix).
- next calculate the Euclidean distances between points pairwise, computed by the two ordination methods (the ordination distance matrix).
- compare the Pearson R-squared between the initial distance matrix and the two distance matrices for the ordinations. Note that the nMDS is calculated with a single random start.

7.4 Hierarchical clustering

The hierarchical clustering algorithm produces a dendrogram with clusters of fingerprint profiles according to their proximities. Seven different algorithms are available: ward, single (often named nearest neighbour), complete, average (often named Unweighted Pair-Group Average UPGMA), McQuity, median and centroid [5] [9]. Different proximity measures can be used to compare the fingerprint profiles (see 7.1). However, for Ward's method, Euclidean distance is inherent in the algorithm.

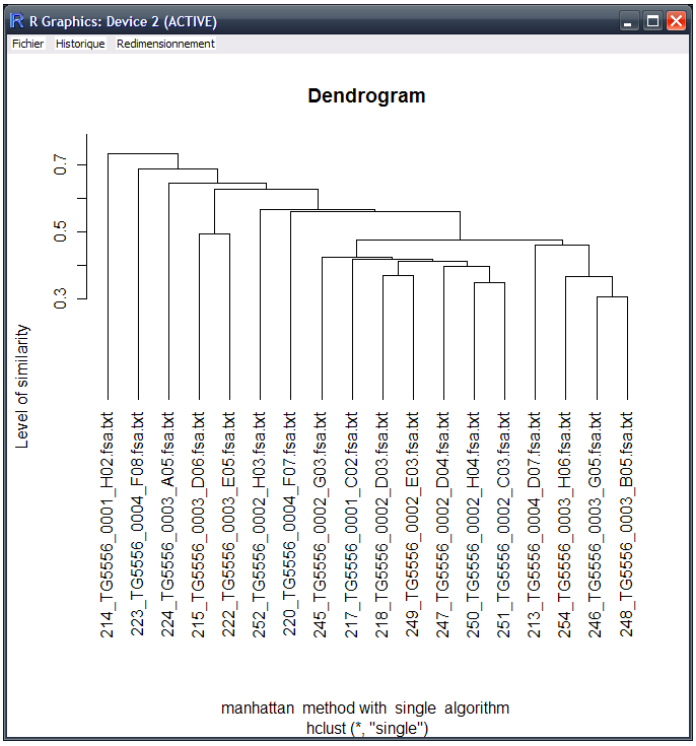


Fig 23: Dendrogram

7.5 Heatmap

A heatmap is a hierarchical clustering with summarized fingerprint profiles added (the signal intensity is proportionnal to a gradient of colours) to improve the visual interpretation of the plot. Fingerprint profiles are in rows. Seven and fourteen algorithms can be used for plotting the dendrogram and for calculating proximity measures respectively between fingerprint profiles (see 7.4 and 7.1).

7.6 Multivariate ANOVA

General linear modelling of fixed-effect models with multiple responses is performed. The procedure calculates 50-50 MANOVA p-values, ordinary univariate p-values and adjusted p-values using rotation testing [6]. We advise using at least 1000 rotations to produce an accurate p-value. For help to generate models (see 6.3). Statistics are displayed on the R console.

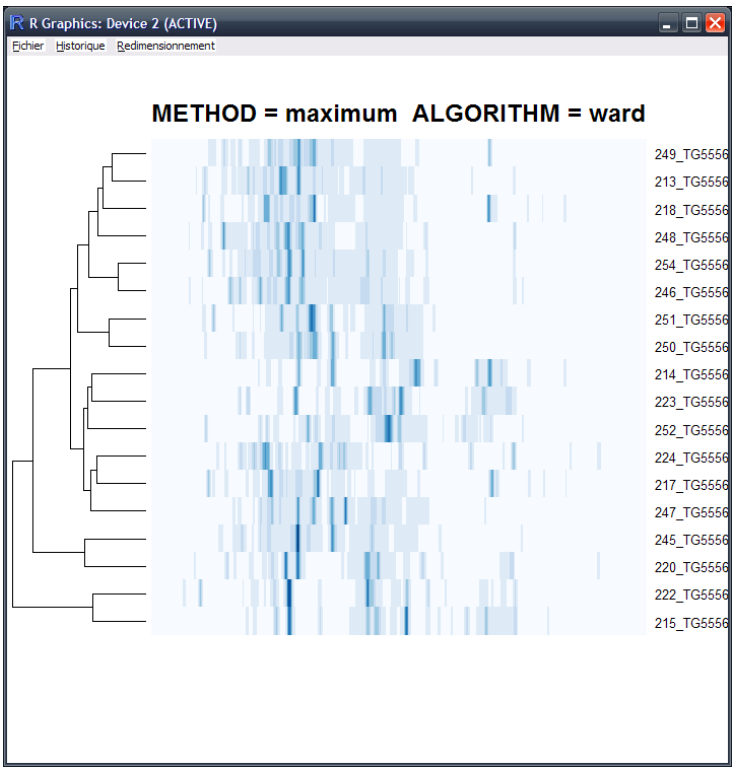


Fig 24: Heatmap

7.7 Global ANOSIM

ANOSIM (ANalysis Of SIMilarities) is a non-parametric statistical test of significant difference between two or more groups, based on any proximity measures available [1]. Fourteen proximity measures can be chosen (see 7.1). In this procedure, groups are designed according to the level of a factor (qualitative variable). In a rough analogy with ANOVA, the tests is based on comparing distances between groups with distances within groups to produce the ANOSIM statistic R:

$$R = \frac{(rB - rW)}{N(N-1)/4}$$

rB and rW are the mean of the ranked similarity BETWEEN groups and WITHIN groups respectively and N is the total number of samples (objects).
This ANOSIM R value ranges from 0 (no difference) to 1 (completely separated groups). The statistical significance of observed ANOSIM R is assessed by Monte Carlo permutations to obtain the empirical distribution of ANOSIM R under the null hypothesis. We advise at least 1000 permutations to produce an accurate p-value. Statistics are displayed on the R console.

7.8 Pairwise ANOSIM

This procedure uses exactly the same algorithm as that described for the global ANOSIM. It indicates which levels of a factor (qualitative variable) differ from the others (when there is a significant differrence according to this qualitative variable using global ANOSIM). First select the factor (qualitative variable), next choose the two levels to compare. When selecting "all pairwise ANOSIM" in one or in the two selection boxes of levels, the returned result is a pairwise ANOSIM (ANOSIM

R and p-values) of each pair of levels within the factor. Statistics are displayed on the R console.

7.9 Within-group variability

This procedure tests whether the within-group variability differs significantly for two or more groups of fingerprint profiles. The groups are defined as levels of a factor (a qualitative variable). The test used by this algorithm consists of a type II ANOVA and a Tuckey HSD test as a retrospective test. Fourteen proximity measures can be chosen (see 7.1). This procedure is normally used when fingerprint profiles differ according to the level of a qualitative variable. It shows whether one or more groups of fingerprint profiles (microbial communities) are more or less homogeneous than the other groups (other microbial communities). Statistics are displayed on the R console and you have to choose between 3 kinds of plots (boxplot, points and SD, line and SD).

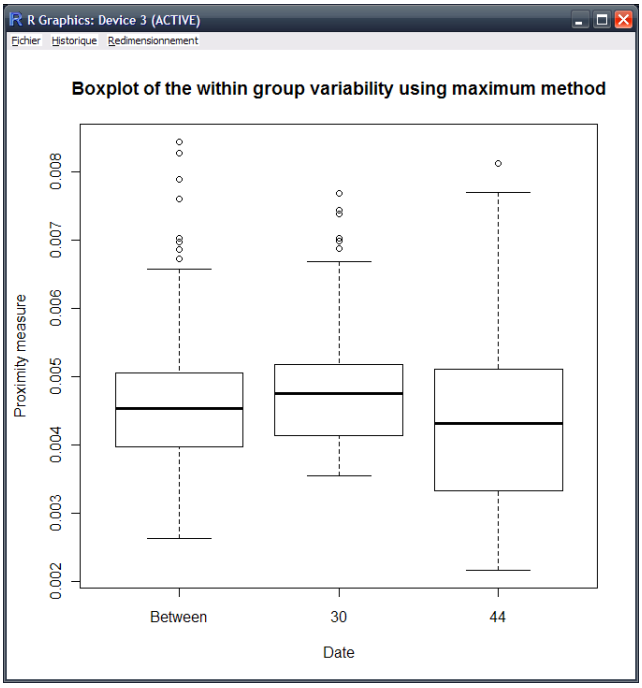


Fig 25: Within-group variability boxplot

7.10 SIMilarity PERcentages procedure

SIMPER (SIMilarity PERcentages procedure) is a simple method for assessing which scans are primarily reponsible for an observed difference between groups of fingerprint profiles. The overall significance of the difference is often assessed by global ANOSIM (see 7.7). In the procedure of this program only the Euclidean distance can be chosen as proximity measure.

First select the qualitative variable; next choose the two levels to compare. The output of SIMPER is a list of the scans sorted in decreasing order of contributing to the overall dissimilarity. Their relative and cumulative contributions as percentages are also provided. All this data can be exported with the EXPORT SIMPER DATA

button. The procedure also displays a graph with a black curve indicating the relative percentages of contribution of scans. The threshold (percentage of contribution in per mill) allows scans above the threshold to be coloured in red in the resulting plot.

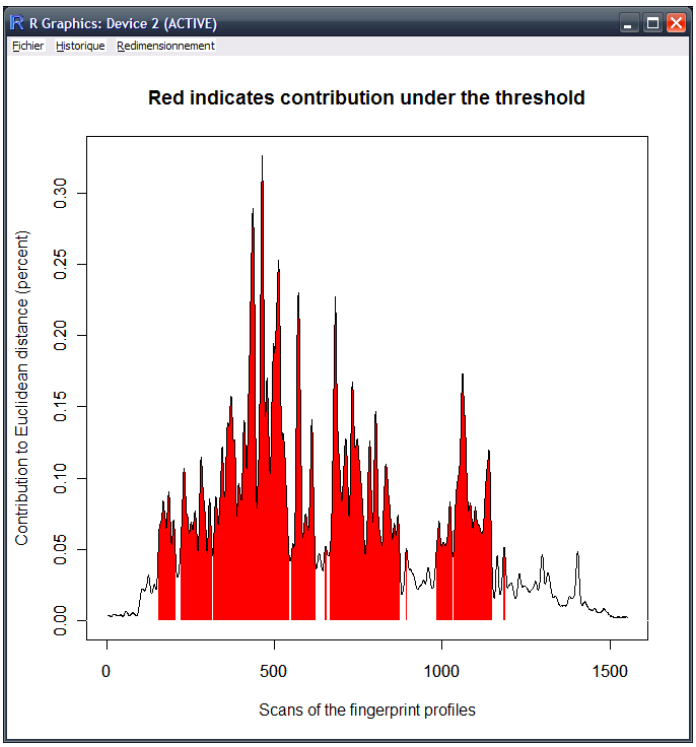
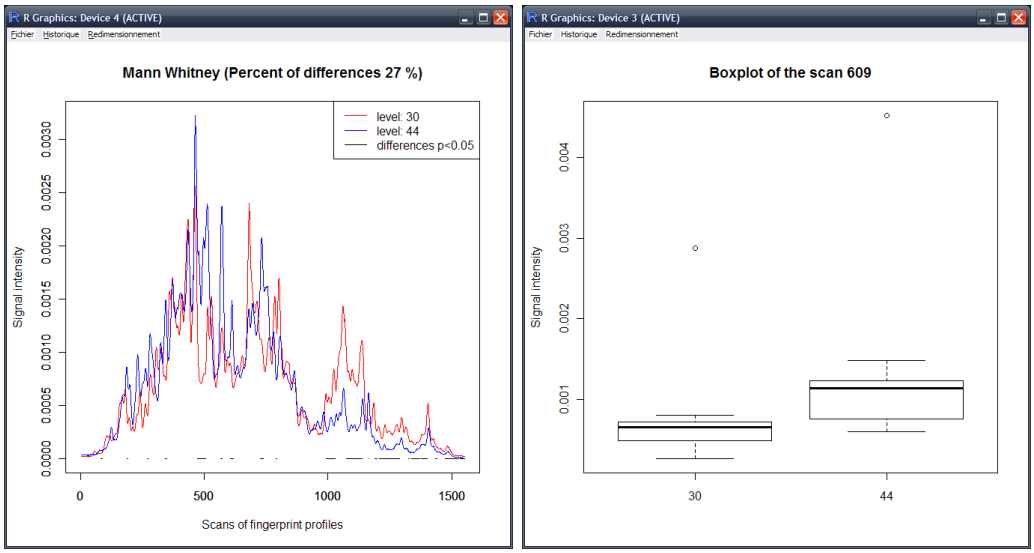


Fig 26: SIMPER function

7.11 Iterative test

An iterative tests shows which scans differ significantly at $p<0.05$ when comparing two groups of fingerprint profiles (two microbial communities). The two groups are defined according to two levels of a qualitative variable. The overall significance of the difference is often assessed by global (7.7) or pairwise ANOSIM (7.8). Fourteen proximity measures can be chosen (see 7.1) and three statistical tests: T-test (parametric), Mann Whitney test (non-parametric) and Fischer's exact test (for binary fingerprint profiles).

First select the qualitative variable; then choose the two levels to compare within this qualitative variable. The result is a plot representing the average fingerprint profiles of each group and the areas of difference as block horizontal lines along the average fingerprint profiles. The percentange as well as the number of significantly diffenrent scans are also provided. The VISUALIZE DISTRIBUTION OF A SCAN button can be useful to check the result of iterative test of a specific scan.



(a) Iterative test plot (b) Boxplot at a specific scan

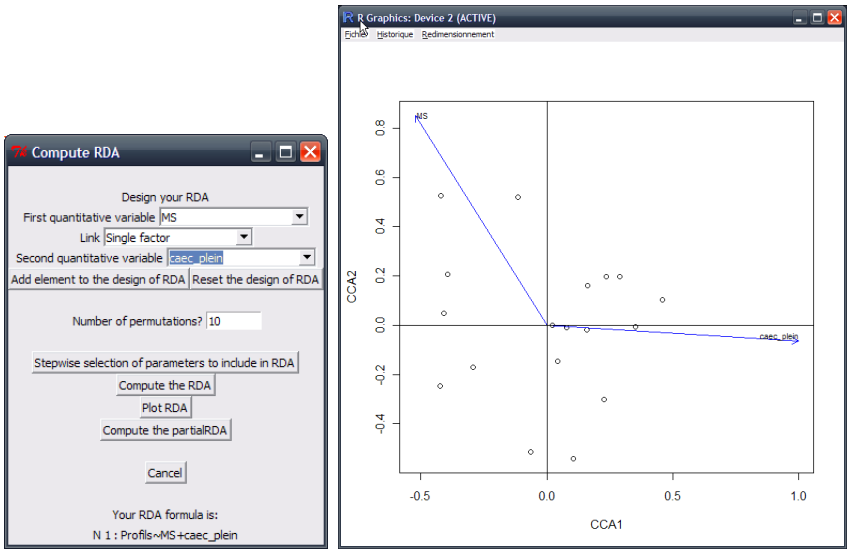
Fig 27: Iterative test function

7.12 Multivariate correlation

This procedure performs a multivariate correlation test using general linear correlation with a single quantitative variable. P-values, ordinary univariate p-values and adjusted p-values using rotation testing are calculated [6]. We advise using at least 1000 rotations to produce accurate p-values. This procedure is typically used to study the effect of the structure of the microbial community according to environmental parameters.

7.13 Redundancy analysis (RDA)

Redundancy analysis is related to principal component analysis. It is based on Euclidean distances and perform linear mapping. Be aware that using the RDA function implies that you do not have any missing values in your parameter file. If there are some, the package offers you the possibility to delete temporarily the samples (deleting the profile and all the parameters for this sample) or the parameters (deleting the corresponding columns) containing missing value.



(a) RDA menu (b) Missing value with RDA

Fig 28: RDA function

7.14 Constrained correspondence analysis (CCA)

Constrained Correspondance Analysis is related to correspondance analysis. It is based on Chi-squared distances and performs weighted linear mapping.

7.15 Constrained analysis of proximities (CAP)

Constrained Analysis of Proximities is related to metric scaling (cmdscale). It can handle any dissimilarity measures and performs a linear mapping.

7.16 Export proximity matrix

This procedure allows to export in csv format any of the proximity matrices available in the package.

8 Advanced modes: internal object and procedure management

8.1 Object invoked

Here is a list of objects that are or can be found in a project saved with StatFingerprints:

- **mat.raw** : this object is a list. It contains two matrices: the imported raw fingerprint profiles in rows of `mat$profile` or `mat[[1]]` and their respective internal standard in rows of `mat$rox` or `mat[[2]]` This object is created when using procedures Convert FSA files and import ,Import fingerprint profiles in ASCII files. When using Import an ecological table

procedure, `mat$rox` is equal to `mat$profile` as the internal standards are not loaded.

- **mat.analyse** : this object is a matrix containing the fingerprint profiles resulting from the last procedure to be executed in the Profile processing menu. All procedures of the menus Plot, Univariate statistics, diversity and Multivariate statistics: structure are executed using this matrix.

- **mat.align** : this object is a matrix containing the aligned fingerprint profiles resulting from the Align procedure one by one procedure.

- **mat.background** : this object is a matrix containing the fingerprint profiles if the Delete background under profiles procedure has been executed; otherwise it is equal to 1.

- **mat.baseline** : this object is a matrix containing the aligned fingerprint profiles resulting from the Define a common baseline for all profiles procedure.

- **mat.range** : this object is a matrix containing the aligned fingerprint profiles resulting from the Define the range of the profiles procedure.

- **mat.normalize** : this object is a matrix containing the aligned fingerprint profiles resulting from the Normalize area under peaks procedure.

- **mat.rebuilt** : this object is a matrix containing the fingerprint profiles resulting from the Rebuild peaks of profiles presenting defaults procedure if used; otherwise it is equal to 1

- **mat.binary** : this object is a matrix containing the fingerprint profiles resulting from the Transform profiles into presence/absence profiles procedure if used; otherwise it is equel to 1

- **rxref** : this object is a vector containing peaks of the reference standard.

- **div** : this object is a matrix containing the diversities values calculated using the Compute diversity index procedure

- **fact** : this object is a matrix containing, in columns, qualitative variables loaded when using the Import variable (quantitative or qualitative) with the option qualitative variable in the file type box.

- **param** : this object is a matrix containing, in columns, quantitative variables loaded when using the Import variable (quantitative or qualitative) with the option quantitative variable in the file type box.

- **alig** : this object is a vector containing the names of the aligned fingerprint profiles which are aligned

8.2 Internal procedures

The function to launch the main GUI interface of the program is `StatFingerprints()` (see 1.1). It displays the general GUI interface. Most procedures of the StatFingerprints program are built with 2 functions. The first is the executive function and the second is the GUI function which displays the GUI window for the executive function. The GUI function is always named with the name of the executive function plus "GUI" added at the end. To list all the function of the package, just type `ls(package:StatFingerprints)` on the R console.

9 Acknowledgement

Authors thank all users for their critical evaluation of the program resulting in numerous improvements. Authors would like to thank Jerôme Hamelin, Kim Milferstedt and Mathieu Bueche for their useful suggestions and work that led to an extended and improved version of the package.

10 References

[1] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[2] Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Austral Ecol 18(1): 117-143

[3] Cox TF, M.A.A. C (2001) Multidimensional Scaling, Chapman and Hall.

[4] Forsythe GE, Malcolm MA, Moler CB (1977) Computer Methods for Mathematical Computations

[5] Gordon AD (1999) Classification, second edn. London.

[6] Langsrud O(2002) 50-50 multivariate analysis of variance for collinear responses. TheStatistician 51: 305-317

[7] Legendre P, Legendre L (1998) Numerical ecology., 2 edn. Amsterdam: Elsevier.

[8] Magurran AE (2004) Measuring biological diversity., Oxford: Blackwell publishing.

[9] Murtagh F (1985) Multidimensional Clustering Algorithms. In Compstat. Wuerzburg

[10] Wolda H (1981) Similarity indices, sample size and diversity. Oecologia 50: 296-302