MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2009) **9**, 1359–1363 doi: 10.1111/j.1755-0998.2009.02609.x

COMPUTER PROGRAM NOTE

# StatFingerprints: a friendly graphical interface program for processing and analysis of microbial fingerprint profiles

R. J. MICHELLAND,*†‡ S. DEJEAN,§ S. COMBES,* L. FORTUN-LAMOTHE* and L. CAUQUIL*

**INRA, UMR 1289 TANDEM, Tissus Animaux, Nutrition, Digestion, Ecosystème et Métabolisme, Université de Toulouse, F-31326 Castanet-Tolosan cedex, France, †INPT-ENSAT, UMR 1289 TANDEM, Université de Toulouse, F-31326 Castanet-Tolosan cedex, France, ‡ENVT, UMR 1289 TANDEM, F-31076 Toulouse cedex, France, §Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse cedex 9, France*

## Abstract

**Molecular fingerprint methods are widely used to compare microbial communities in various habitats. The free program StatFingerprints can import, process, and display fingerprint profiles and perform numerous statistical analyses on them, and also estimate diversity indexes. StatFingerprints works with the free program R, providing an environment for statistical computing and graphics. No programming knowledge is required to use StatFingerprints, thanks to its friendly graphical user interface. StatFingerprints is useful for analysing the effect of a controlled factor on the microbial community and for establishing the relationships between the microbial community and the parameters of its environment. Multivariate analyses include ordination, clustering methods and hypothesis-driven tests like 50–50 multivariate analysis of variance, analysis of similarity or similarity percentage procedure and the program offers the possibility of plotting ordinations as a three-dimensional display.**

*Keywords*: diversity, ecology, fingerprint, processing, statistic

*Received 5 December 2008; accepted 21 January 2009*

Molecular fingerprint methods and especially those based on capillary electrophoresis (CE), such as terminal restriction fragment length polymorphism (T–RFLP), single-strand conformation polymorphism (SSCP) or automated ribosomal intergenic spacer analysis (ARISA) offer, at little cost, a rapid high resolution snapshot of the overall community in a sample (Hori *et al*. 2005; King *et al*. 2005; Hong *et al*. 2007; Zinger *et al*. 2007). Consequently, they have been widely used in various environments to compare the communities of numerous samples (Osborn *et al*. 2000; Leclerc *et al*. 2004; King *et al*. 2005; Tatsuoka *et al*. 2006; Babendreier *et al*. 2007; Guo *et al*. 2007; Kent *et al*. 2007). However, the fingerprint profiles generated are difficult to analyse. First, they need to be processed before they can be compared, and second, the large data sets generated require multivariate statistical methods. Dedicated programs like GeneMarker (SoftGenetics Inc.), Safum (Zemb *et al*. 2007), DAx (Van Mierlo program) or GeneScan (Applied Biosystems) provide no or only rudimentary statistical tools, with limited, if any, export possibilities to perform further analyses. They are also expensive. Furthermore, pre-programmed algorithms cannot be parameterized sufficiently to be satisfactorily used for ecological purposes (Zemb *et al*. 2007). For example, it was recently demonstrated that the area under the peaks is informative and must be taken into account for statistical analysis; hence, programs need to be able to retain this background information (Loisel *et al*. 2006).

Therefore, due to the wide use of fingerprint methods and the inadequacy of current programs for processing and analysing fingerprint profiles, we developed a free program called StatFingerprints. No programming knowledge is required to use it, thanks to its friendly graphical user interface (GUI). It can i) import raw files containing fingerprint profiles (FSA, ASCII), ii) process fingerprint profiles (alignments, normalization, etc.), iii) estimate diversity indexes, iv) perform univariate and a

Correspondence: Laurent Cauquil, Fax: 33 (5) 61 28 53 18; E-mail: laurent.cauquil@toulouse.inra.fr

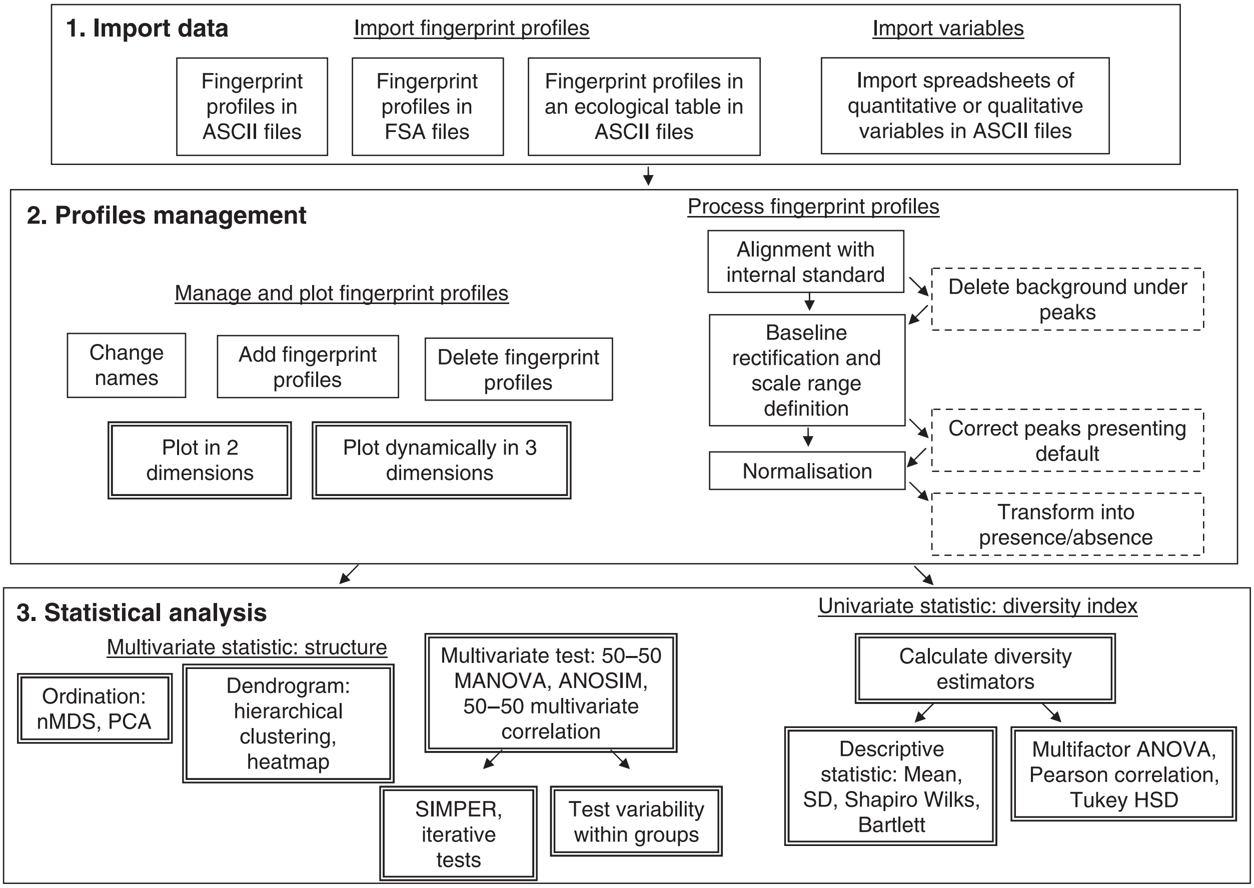
1. Import data
Import fingerprint profiles
Fingerprint profiles in ASCII files
Fingerprint profiles in FSA files
Fingerprint profiles in an ecological table in ASCII files
Import variables
Import spreadsheets of quantitative or qualitative variables in ASCII files
2. Profiles management
Manage and plot fingerprint profiles
Change names
Add fingerprint profiles
Delete fingerprint profiles
Plot in 2 dimensions
Plot dynamically in 3 dimensions
Process fingerprint profiles
Alignment with internal standard
Delete background under peaks
Baseline rectification and scale range definition
Correct peaks presenting default
Normalisation
Transform into presence/absence
3. Statistical analysis
Multivariate statistic: structure
Ordination: nMDS, PCA
Dendrogram: hierarchical clustering, heatmap
Multivariate test: 50–50 MANOVA, ANOSIM, 50–50 multivariate correlation
SIMPER, iterative tests
Test variability within groups
Univariate statistic: diversity index
Calculate diversity estimators
Descriptive statistic: Mean, SD, Shapiro Wilks, Bartlett
Multifactor ANOVA, Pearson correlation, Tukey HSD

**Fig. 1** Structure of the StatFingerprints program. Three main parts guide the user: (1) importation, (2) management, and (3) statistical analysis of the fingerprint profiles. Arrow indicates successive steps. A step with double borders offers the possibility of exporting its numerical or graphical results. Steps with a dashed border are optional.

wide range of multivariate statistical tests, and v) plot in three dimensions with dynamic control. The program StatFingerprints works with R: a free program providing an environment for statistical computing and graphics (R development Core Team 2008). StatFingerprints and its user guide can be downloaded from the comprehensive R archive network at http://cran.r-project.org/web/packages/StatFingerprints/index.html. In this note, we describe some possibilities of the program StatFingerprints.

## Import and export fingerprint profiles

Fingerprint profiles from CE can be imported into StatFingerprints using all ASCII formats (Fig. 1) and also directly from the output of ABI PRISM sequencers (Applied Biosystems) using the FSA files. Data generated by traditional gel-based methods like denaturing gradient gel electrophoresis (DGGE) or temporal temperature gel electrophoresis (TGGE) can also be imported as chromatograms. All results can be exported in all ASCII formats and plots can be exported in Metafile, Postscript, PDF, png, bmp, TIFF or jpeg formats.

## Process fingerprint profiles

CE raw fingerprint profiles can be aligned with a safe manual control of the detection of the internal standard peaks. The baseline of the fingerprint profiles can also be aligned and horizontally re-oriented. The background under peaks of fingerprint profiles, which is often considered as noise, can be deleted. The area under each fingerprint profile can be normalized. Processing the fingerprint profiles in such a way is required to compare them for further statistical analysis. Some authors prefer to transform fingerprint profiles into binary fingerprint profiles due to the semi-quantitative information given by fingerprint methods (Blackwood *et al.* 2003; Green *et al.* 2004; Ramírez-Moreno *et al.* 2005). The program StatFingerprints can detect peaks in profiles and then transform areas with and without peaks to one and zero respectively to obtain such binary profiles.

## Estimation and analysis of diversity indexes

Diversity indexes are frequently estimated from fingerprint profiles. They summarize all the scans of a fingerprint profile

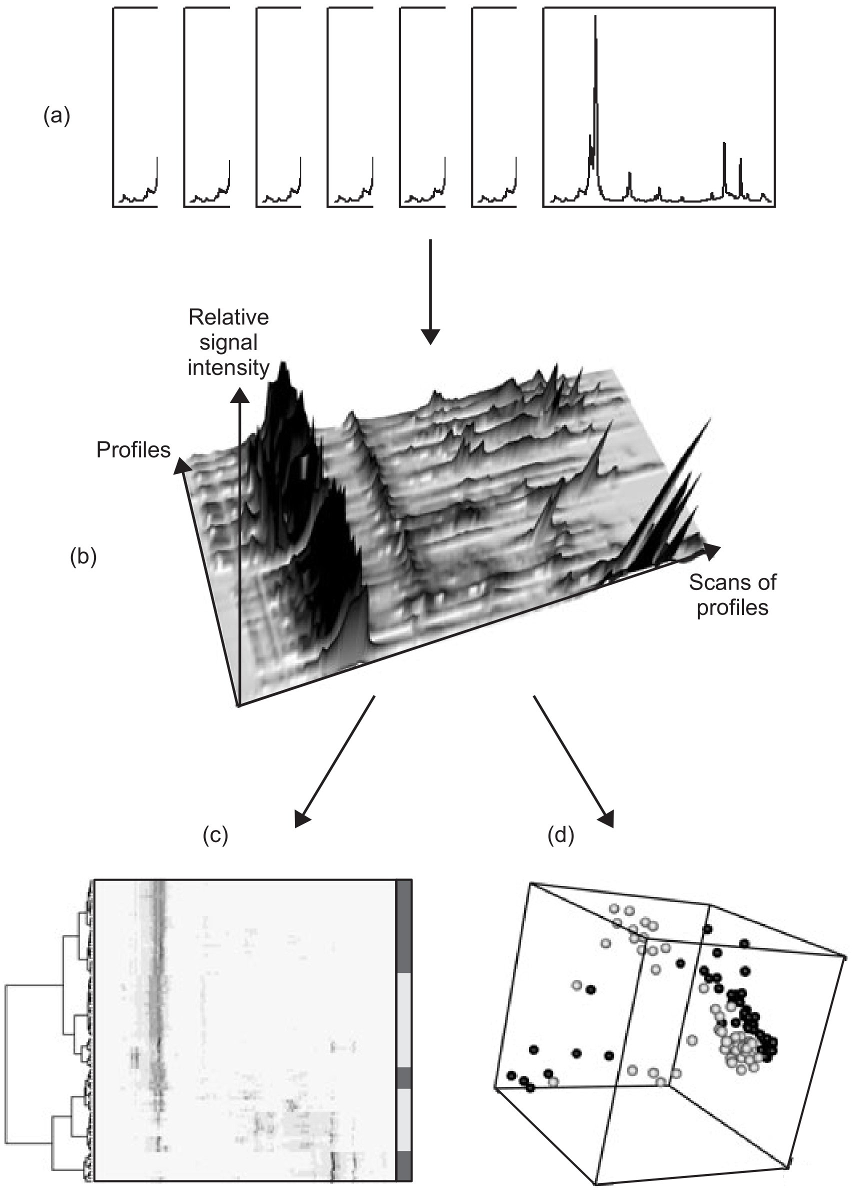

**Fig. 2** Analysis of CE-SSCP profiles of the archaeal community of the digestive tract of the rabbit and the cow using StatFingerprints. Raw CE-SSCP profiles (a) require processing before they can be compared (b). Then they can be compared using heatmap (c) or three-dimensional nMDS (d). The black and grey colours correspond to rabbit and cow, respectively. Heatmap was calculated using Euclidean distances and the Ward algorithm. The nMDS was computed with 10 000 random starts using Euclidean distances. The stress of the nMDS was equal to 0.050.

into a single value by taking into account the number of peaks and their relative abundance. StatFingerprints can estimate several diversity indexes: the richness, Simpson's negative logarithm, one minus Simpson, Shannon, Buzas and Gibson's evenness and equitability (Rosenzweig 1995; Begon *et al*. 1996; Magurran 2004). The algorithm for the detection of the peaks in the fingerprint profiles can be easily and fully parameterized. For example, artefact peaks can be deleted using a threshold and the peaks' abundance can be calculated using their heights or their areas. StatFingerprints can perform some commonly used univariate tests: multifactor ANOVA, Tukey's honest significant difference, Pearson's correlation, Bartlett's test or Shapiro–Wilk's test. These tests can be used to calculate basic statistics either on the diversity index or on other imported variables.

## Analysis of the structure of community with multivariate statistics

The proximities between pairs of fingerprint profiles can be calculated using 13 proximity measures: Euclidean distance, maximum distance, Manhattan distance, Canberra distance, Minkowski distance, Bray–Curtis similarity index, chi-squared correlation index, Ruzicka similarity index, Roberts similarity index, Jaccard binary index, Dice-Sørensen binary index, Ochiai binary index, Stainhaus binary index (Wolda 1981; Legendre & Legendre 1998; Magurran 2004). The proximity measures between fingerprint profiles can be explored using either ordination or dendrogram methods (Fig. 2). Ordination methods available in the program StatFingerprints are principal components analysis

(PCA) and random starts nonmetric multidimensional scaling (nMDS, Cox & Cox 2003; Borg & Groenen 2005). These analyses produce a display in two or three dimensions in which each point represents one fingerprint profile. From the proximity matrix, nMDS plots points so as to respect as much as possible the proximity measures between each pair of profiles. On the other hand, PCA does not use the proximity matrix. PCA calculates the linear combination with the largest variance within scan values of all profiles to produce synthetic variables. These synthetic variables determine the axis of the space where fingerprint profiles are plotted as points (Hotelling 1933). Dendrogram methods available in StatFingerprints are hierarchical clustering and heatmap. Seven dendrogram construction algorithms are available: Ward, nearest neighbour, unweighted pair-group average, average, Mc Quitty, median and centroid (Murtagh 1985; Gordon 1999). Heatmap is a hierarchical clustering coupled with a simplified representation of the profiles using a colour data set (the signal intensity is translated into a gradient of colours), which facilitates the visual interpretation of the plot.

The influence of a factor on the structure of the microbial community has often been studied by a subjective observation of the clusterings on the ordination plot but has rarely been statistically determined using a hypothesis-driven test. StatFingerprints offers two hypothesis-driven tests: 50–50 multivariate ANOVA with rotations (Langsrud 2002; Langsrud 2005) and analysis of similarity (ANOSIM) with Monte Carlo permutations (Clarke 1993; Chapman & Underwood 1999). ANOSIM analysis can be based on any of the 13 proximity measures available but can test the effect of only one factor. In the other hand, the 50–50 multivariate ANOVA is not based on proximity measures but can test the effect of two or more factors and their interactions. The multivariate post-hoc test can be performed using pairwise ANOSIM (Clarke 1993) to understand which levels differ from the others within a factor. To go further into the analysis and to understand which scans along a fingerprint profile explain the difference between two groups of fingerprint profiles (two levels within a factor), StatFingerprints offers either the similarity percentage procedure (SIMPER; Clarke 1993) or iterative tests. The SIMPER calculates the relative contribution of each scan of a fingerprint profile to the dissimilarities between the two groups, while the iterative test performs a univariate test between the two groups for each scan of the profile. The univariate test can be a *T*-test, Mann–Whitney or Fisher's exact tests for normal, non-normal, nominal (binary fingerprint profiles) variables respectively. StatFingerprints can also calculate the variability within each group of fingerprint profiles and test whether this variability differs significantly between groups. Establishing the relationship between the structure of the microbial community and environmental parameters, for example, a gradient pollution, is often required. For this, StatFingerprints provides a 50–50 multivariate correlation (Langsrud 2002; Langsrud 2005).

In conclusion, StatFingerprints provides a useful tool for scientists in the field of microbial ecology who use molecular fingerprint methods and especially those based on CE. It offers, in a single free program, a way to easily manipulate fingerprint profile output data from the sequencer, and perform numerous statistical analyses on them. A wide range of importing and processing methods, multivariate statistical tests, and the ability to estimate diversity indexes is available, using a user-friendly graphical interface. All the results and figures are exportable and can be easily included in a publication. To our knowledge, StatFingerprints is the only free program to include statistical tests like pairwise ANOSIM, SIMPER, iterative tests or heatmap to analyse the fingerprint data.

## Acknowledgements

We thank Olivier Zemb, Jerôme Hamelin and Laurence Lamothe for helpful ideas and discussion. We are grateful for Mélanie Martignon and Moussa Kimsé for their suggestions.

## References

Babendreier D, Joller D, Romeis J, Bigler F, Widmer F (2007) Bacterial community structures in honeybee intestines and their response to two insecticidal proteins. *FEMS Microbiology Ecology*, **59**, 600–610.

Begon M, Harper JL, Townsend CR (1996) *Ecology: Individuals, Populations and Communities*. Blackwell Science, Oxford, UK.

Blackwood CB, Marsh T, Kim S-H, Paul EA (2003) Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Applied and Environmental Microbiology*, **69**, 926–932.

Borg I, Groenen PJF (2005) *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, New York.

Chapman MG, Underwood AJ (1999) Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests. *Marine Ecology Progress Series*, **180**, 257–265.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, **18**, 117–143.

Cox TF, Cox MAA (2003) Multidimensional scaling. *Technometrics*, **45**, 328–328.

R development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Gordon AD (1999) *Classification*. 2nd edn, Chapman and Hall, London.

Green JL, Holmes AJ, Westoby M *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature*, **432**, 747–750.

Guo Y, Zhu N, Zhu S, Deng C (2007) Molecular phylogenetic diversity of bacteria and its spatial distribution in composts. *Journal of Applied Microbiology*, **103**, 1344–1354.

Hong H, Pruden A, Reardon KF (2007) Comparison of CE-SSCP and DGGE for monitoring a complex microbial community remediating mine drainage. *Journal of Microbiological Methods*, **69**, 52–64.

Hori T, Haruta S, Ueno Y, Ishii M, Igarashi Y (2005) Direct comparison of single-strand conformation polymorphism (SSCP) and denaturing gradient gel electrophoresis (DGGE) to characterize a microbial community on the basis of 16S rRNA gene fragments. *Journal of Microbiological Methods*, **66**, 165–169.

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *The British Journal of Educational Psychology*, **24**, 417–441.

Kent AD, Yannarell AC, Rusak JA, Triplett EW, McMahon KD (2007) Synchrony in aquatic microbial community dynamics. *ISME Journal*, **1**, 38–47.

King S, McCord BR, Riefler RG (2005) Capillary electrophoresis single-strand conformation polymorphism analysis for monitoring soil bacteria. *Journal of Microbiological Methods*, **60**, 83–92.

Langsrud Ø (2002) 50–50 multivariate analysis of variance for collinear responses. *The Statistician*, **51**, 305–317.

Langsrud Ø, (2005) Rotation tests. *Statistics and Computing*, **15**, 53–60.

Leclerc M, Delgenes J-P, Godon J-J (2004) Diversity of the archaeal community in 44 anaerobic digesters as determined by single strand conformation polymorphism analysis and 16S rDNA sequencing. *Environmental Microbiology*, **6**, 809–819.

Legendre P, Legendre L (1998) *Numerical Ecology*, 2 edn. Elsevier, Amsterdam, The Netherlands.

Loisel P, Harmand J, Zemb O *et al*. (2006) Denaturing gradient electrophoresis (DGE) and singlestrand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. *Environmental Microbiology*, **8**, 720–731.

Magurran AE (2004) *Measuring Biological Diversity*. Blackwell publishing, Oxford, UK.

Murtagh F (1985) *Multidimensional Clustering Algorithms*. Physica-Verlag, Vienna, Austria.

Osborn AM, Moore ERB, Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environmental Microbiology*, **2**, 39–50.

Ramírez-Moreno S, Martínez-Alonso M, Méndez-Alvarez S, Gaju N (2005) Seasonal microbial ribotype shifts in the sulfurous karstic lakes Cisó and Vilar, in northeastern Spain. *International Microbiology*, **8**, 235–242.

Rosenzweig ML (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Tatsuoka N, Mohammed N, Mitsumori M *et al*. (2006) Analysis of methanogens in the bovine rumen by polymerase chain reaction single-strand conformation polymorphism. *Animal Science Journal*, **78**, 512–518.

Wolda H (1981) Similarity indices, sample size and diversity. *Oecologia*, **50**, 296–302.

Zemb O, Haegeman B, Delgenes JP, Lebaron P, Godon JJ (2007) Safum: statistical analysis of SSCP fingerprints using PCA projections, dendrograms and diversity estimators. *Molecular Ecology Notes*, **7**, 767–770.

Zinger L, Gury J, Giraud F *et al*. (2007) Improvements of polymerase chain reaction and capillary electrophoresis single-strand conformation polymorphism methods in microbial ecology: toward a high-throughput method for microbial diversity studies in soil. *Microbial Ecology*, **54**, 203–216.