Molecular Ecology Notes (2007) 7, 583–587 doi: 10.1111/j.1471-8286.2007.01744.x

PROGRAM NOTE

TRAMPR: an R package for analysis and matching of terminal-restriction fragment length polymorphism (TRFLP) profiles

RICHARD G. FITZJOHN and IAN A. DICKIE
Landcare Research, PO Box 40, Lincoln 7640, New Zealand

Abstract

TRAMPR (TRFLP analysis and matching package for R) is a package for matching multiple terminal restriction fragment length polymorphism (TRFLP) profiles between unknown samples and a database of known TRFLP profiles in order to infer the presence of species in environmental samples. It permits simultaneous analysis of multiple samples and facilitates direct workflow from electrophoresis output through to community analyses. TRAMPR also resolves the issues of multiple TRFLP profiles within a species and (conversely) shared TRFLP profiles across species.

Keywords: community analysis, pattern matching, R, TRFLP

Received 24 November 2006; revision accepted 7 February 2007

Terminal restriction fragment length polymorphism (TRFLP, sometimes T-RFLP) was initially developed as a technique for assessing diversity and shifts in bacterial communities (Liu *et al*. 1997; Marsh 1999), but has since been adapted for the identification of species by using multiple TRFLP profiles to identify species, particularly of fungi (Dickie *et al*. 2002; Zhou & Hogetsu 2002; Avis & Feldheim 2005; Burke *et al*. 2005; Edwards & Turco 2005). This application, which we term 'database TRFLP', involves matching multiple electropherogram profiles (using multiple tagged primers and/or multiple restriction digests) from environmental samples against a database of known profiles to infer the presence of species (Dickie *et al*. 2002). Database TRFLP simultaneously increases the information obtained and avoids many of the potential pitfalls of traditional TRFLP applications (Avis *et al*. 2006).

The number of peak comparisons required when comparing profiles can become very large, requiring computer software to automate the process for all but small collections of sample and known TRFLP profiles. A number of TRFLP analysis programs are available, including programs using virtual digests of sequence databases to determine predicted fragment lengths such as 'TAP-TRFLP' (Marsh *et al*. 2000) for ribosomal DNA and 'TRFCUT' (Ricke *et al*. 2005) for functional marker genes, programs that match virtual digests of sequence data to observed TRFLP profiles from samples ('T-RFLP FRAGSORT'; http://www.oardc.ohio-state.edu/trflpfragsort), programs for calculating profile similarity such as 'T-RFLP PROFILE' (http://rdp8.cme.msu.edu/cgis/trflp.cgi) or for clustering T-RFLP profiles from multiple communities ('T-RFLP STATS'; Abdo *et al*. 2006; 'T-RFLP PROFILE MATRIX'; http://rdp8.cme.msu.edu/cgis/trflp.cgi), and programs to find consensus profiles from multiple T-RFLP profiles ('T-ALIGN'; Smith *et al*. 2005). None of these programs were optimized for matching a database of eukaryotic T-RFLP profiles from multiple restriction digests to environmental samples. To accomplish this, a rudimentary Excel (Microsoft) spreadsheet 'TRAMP' (TRFLP analysis and matching program) was first developed and used in Dickie *et al*. (2002) and subsequently applied by other researchers (Edwards *et al*. 2004; Allmer *et al*. 2006; Genney *et al*. 2006). However, TRAMP only permitted analysis of a single sample at a time, had little flexibility, and required extensive reformatting of data from sequencer output files, limiting its utility. In addition, TRAMP had no way of dealing with known species that had multiple profiles, or where multiple species had indistinguishable profiles. A very recent addition to the field, FRAGMATCH (Saari *et al*. 2007), which was developed for the same purposes as TRAMP, improves on the TRAMP interface. Nonetheless, FRAGMATCH

Correspondence: I. A. Dickie, Fax: +64 3321 9998; E-mail: dickiei@landcareresearch.co.nz

shares many of the limitations of the original TRAMP software.

We have therefore re-implemented and extended the underlying algorithms of TRAMP in R (R Development Core Team 2006), producing the package 'TRAMPR' (TRFLP analysis and matching package for R). While TRAMPR uses the same core logic as the original TRAMP, it represents a complete overhaul. Major features of TRAMPR are:

1 The ability to directly load data from ABI output files
2 The ability to automatically and simultaneously analyse multiple samples
3 Flexibility in the method of determining if a known pattern is present or not
4 Automatic grouping of knowns that share species names or highly similar TRFLP patterns
5 Automated detection of putative 'unknown knowns'
6 Direct output of a species presence/absence matrix that can be used for community analysis.

TRAMPR works by first generating a matrix containing the distances (in base pairs) between peaks in a collection of samples and a database of knowns across several enzyme/primer combinations. Each known TRFLP profile may have only a single peak per enzyme/primer combination, while samples may have many peaks per combination (perhaps representing different species present). For s samples, k knowns and n enzyme/primer combinations, this generates a three-dimensional $s \times k \times n$ matrix, each element of which is the minimum distance between the single known peak and any sample peak for that enzyme/primer combination. This difference matrix is then summarized to determine if a particular known species might be present, by default calculating whether the maximum of $|p_i - q_i|$ (where p_i and q_i are the sizes of peaks for the ith enzyme/primer combination for a sample and known, respectively) is larger than a threshold acceptable matching error (by default 1.5 base pairs). Euclidean and Manhattan distances can also be used by modifying the default settings. With the exception of detection of putative 'unknown knowns', TRAMPR uses only the sizes of peaks in base pairs and does not evaluate the relative heights of peaks. Excluding peak heights avoids overlapping peak sizes masking the possible presence of a species.

The TRFLP profile of a single species can have variation in peak sizes due to DNA sequence variation, with profiles for different individuals slightly or completely different (Avis *et al.* 2006). If not accounted for, multiple TRFLP patterns could inflate the number of species recorded as present within a sample. Alternatively, two or more different known species may share a similar TRFLP profile and therefore be indistinguishable using TRFLP. If these patterns are not grouped, two species will be recorded as present wherever either is present. TRAMPR resolves both types of error by automatically grouping knowns together where knowns (i) share a common species name or (ii) share highly similar TRFLP profiles (based on clustering). Knowns are grouped together iteratively, applying the two rules until no groups are changed; in certain cases this may chain together seemingly unrelated groups. The clustering parameters are configurable, but the defaults are designed to match the behaviour of the matching algorithm. It is useful to plot the knowns data to observe how clustering occurs (Fig. 1). Analysis of community patterns should be done on knowns groups, rather than on individual TRFLP profile matches.

In addition to analysing unknown samples against a database of knowns, TRAMPR permits knowns databases to be developed from samples. This can be done either from known tissue samples (e.g. cultures, sporocarps, spores, sclerotia), or sample data can be mined for 'unknown knowns' using the build.knowns() function. Using build.knowns(), unknown knowns are defined as TRFLP profiles with only one dominant peak in each enzyme/primer pair, with dominance being defined as a ratio of peak height of the highest to the second highest peak greater than a threshold minimum (default 3). TRAMPR also permits interactive addition of knowns and removal of suspect matches using the update() function.

TRFLP output can be directly loaded from GeneMapper (Applied Biosystems) output files, using load.abi(). At present the automated file conversion convenience functions are only implemented for ABI output files, but data from other systems can be loaded manually using TRAMPsamples() and TRAMPknowns() after converting data into the correct format (see documentation). We welcome user contributions of convenience functions to streamline data loading. In addition, TRAMPR includes sample data that can be used to demonstrate functionality; the data may be loaded by using data(demo.knowns) and data(demo.samples).

The output of TRAMPR includes a samples × knowns presence/absence matrix, which can show either the presence of individual knowns or (more usefully) the presence of knowns groups. TRAMPR can also provide graphical output of the samples and of the match between samples and knowns (Fig. 2). Once a presence/absence matrix is obtained from TRAMPR, it can be directly analysed using R, or exported and analysed using other statistical packages. An example of a complete analysis from raw data to community analysis is provided in the package demo: once TRAMPR is installed, type vignette('TRAMPRdemo') to view the demonstration. TRAMPR also includes a comprehensive reference manual.

Both TRAMPR and R are released under a GPL licence, and may be freely used and modified. The demonstration sample data and knowns database are provided under a

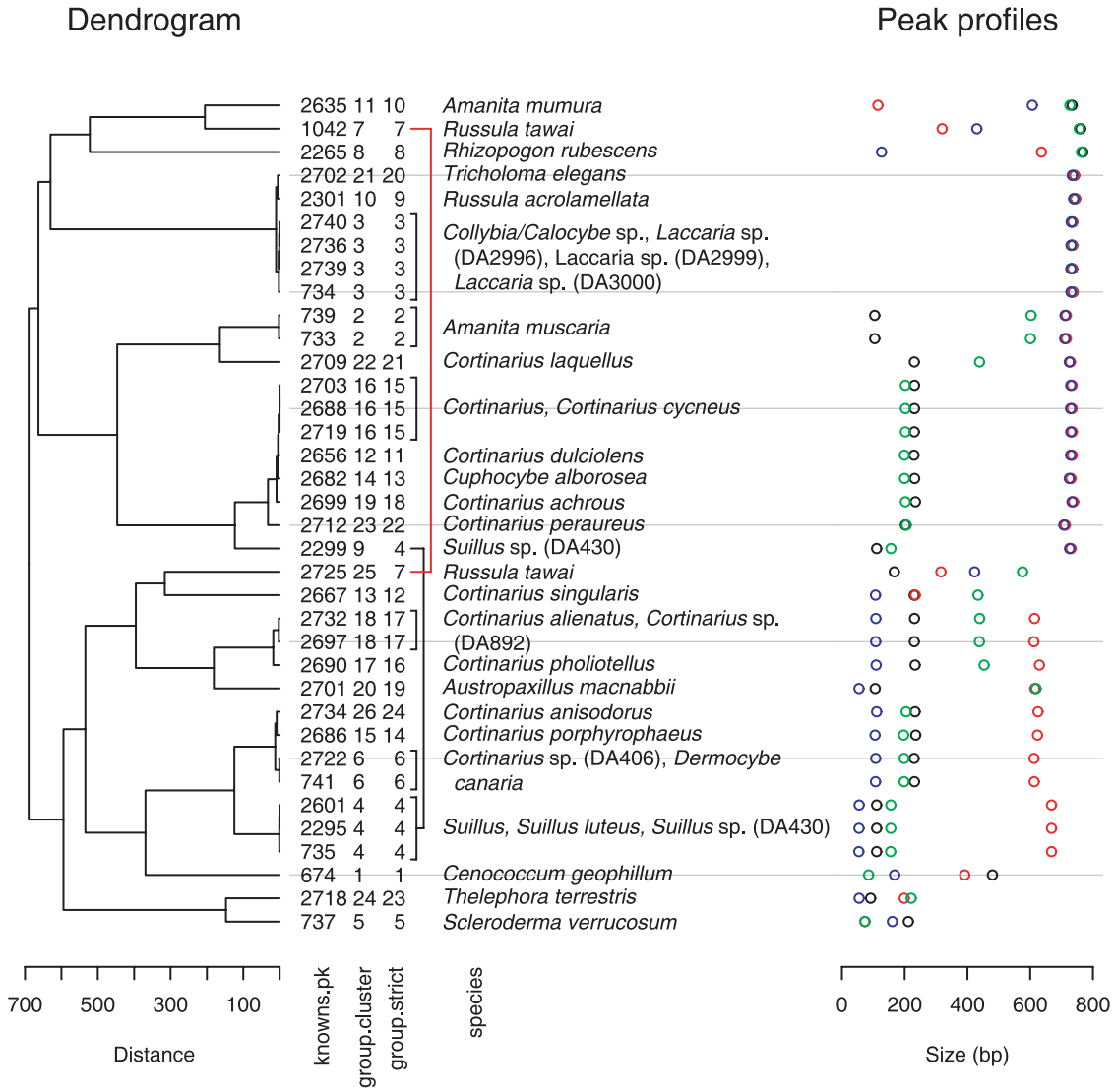

Fig. 1 Labelled knowns plot, showing (from left) the clustering of knowns based on profile similarity, sample identities and group membership, clustering of knowns based on shared names, names of groups, and TRFLP profiles of each known. This plot was generating using plot(demo.knowns).

 Because TRAMPR is open source, the implementation can be easily inspected and modified. Both TRAMPR and R work on all major computer platforms. TRAMPR is available from CRAN (http://cran.r-project.org), and R is available from http://www.r-project.org. TRAMPR requires a recent version of R (2.4.0 or higher).

Acknowledgements

The authors were supported by research funds from the Foundation for Research Science and Technology of New Zealand. P. Avis and S. Branco kindly provided data used in testing the program. A grant from the International Science and Technology (ISAT) Linkages Fund of the Royal Society of New Zealand to I.A.D. was instrumental in developing this program.

References

Abdo Z, Schüette UME, Bent SJ, Williams CJ, Forney LJ, Joyce P (2006) Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environmental Microbiology*, **8**, 929–938.

Allmer J, Vasiliauskas R, Ihrmark K, Stenlid J, Dahlberg A (2006) Wood-inhabiting fungal communities in woody debries of Norway spruce (*Picea abies* (L.) Karst.), as reflected by sporocarps,

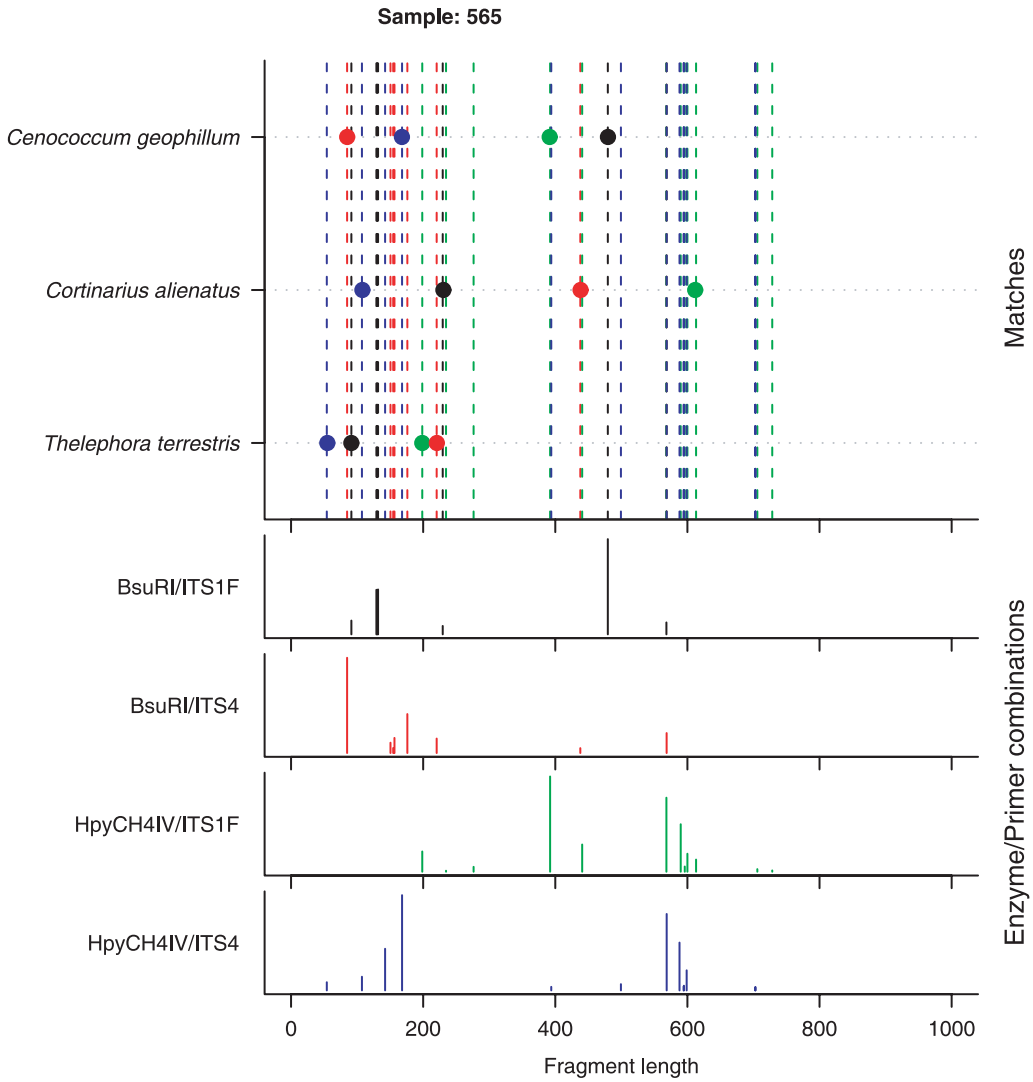

Fig. 2 Example plot showing the TRFLP profile of an unknown species for each enzyme/primer combination (at bottom) and matches to knowns (at top). This plot was generated using fit < – TRAMP(demo.samples, demo.knowns); plot (fit, 565).

mycelial isolations, and T-RFLP analysis. *FEMS Microbiology Ecology*, **55**, 57–67.

Avis PG, Dickie IA, Mueller G (2006) A 'dirty' business: testing the limitations of TRFLP analysis of soil fungi. *Molecular Ecology*, **15**, 873–882.

Avis PG, Feldheim KA (2005) A method to size DNA fragments from 50 to 800 bp on a DNA analyzer. *Molecular Ecology Notes*, **5**, 969–970.

Burke DJ, Martin KJ, Rygiewicz PT, Topa MA (2005) Ectomycorrhizal fungi identification in single and pooled root samples: terminal restriction fragment length polymorphism (TRFLP) and morphotyping compared. *Soil Biology and Biochemistry*, **37**, 1683–1694.

R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (available from: http://www.R-project.org).

Dickie IA, Xu W, Koide RT (2002) Vertical niche differentiation of ectomycorrhizal hyphae in soils as shown by T-RFLP analysis. *New Phytologist*, **156**, 527–535.

Edwards IP, Cripliver JL, Gillespie AR *et al*. (2004) Nitrogen availability alters macrofungal basidiomycete community structure in optimally fertilized loblolly pine forests. *New Phytologist*, **162**, 755–770.

Edwards IP, Turco RF (2005) Inter- and intraspecific resolution of nrDNA TRFLP assessed by computer-simulated restriction analysis of a diverse collection of ectomycorrhizal fungi. *Mycological Research*, **109**, 212–226.

Genney DR, Anderson IC, Alexander IJ (2006) Fine-scale distribution of pine ectomycorrhizas and their extramatrical mycelium. *New Phytologist*, **170**, 381–390.

Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S

rRNA. *Applied and Environmental Microbiology*, **63**, 4516–4522.

Marsh TL (1999) Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Current Opinion in Microbiology*, **2**, 323–327.

Marsh TL, Saxman P, Cole J, Tiedje J (2000) Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. *Applied and Environmental Microbiology*, **66**, 3616–3620.

Ricke P, Kolb S, Braker G (2005) Application of a newly developed ARB software-integrated tool for *in silico* terminal restriction fragment length polymorphism analysis reveals the dominance of a novel pmoA cluster in a forest soil. *Applied and Environmental Microbiology*, **71**, 1671–1673.

Saari TA, Saari SK, Campbell CD, Alexander IJ, Anderson IC (2007) FRAGMATCH — a program for the analysis of DNA fragment data. *Mycorrhiza*, in press.

Smith CJ, Danilowicz BS, Clear AJ, Costello FJ, Wilson B, Meijer WG (2005) T-Align, a web-based tool for comparison of multiple terminal restriction fragment length polymorphism profiles. *FEMS Microbiology Ecology*, **54**, 375–380.

Zhou ZH, Hogetsu T (2002) Subterranean community structure of ectomycorrhizal fungi under *Suillus grevillei* sporocarps in a *Larix kaempferi* forest. *New Phytologist*, **154**, 529–539.