REPK: an analytical web server to select restriction endonucleases for terminal restriction fragment length polymorphism analysis

Roy Eric Collins and Gabrielle Rocap*

School of Oceanography, University of Washington, Seattle WA, USA

Received January 31, 2007; Revised April 18, 2007; Accepted April 30, 2007

ABSTRACT
Terminal restriction fragment length polymorphism (T-RFLP) analysis is a widespread technique for rapidly fingerprinting microbial communities. Users of T-RFLP frequently overlook the resolving power of well-chosen restriction endonucleases and often fail to report how they chose their enzymes. REPK (Restriction Endonuclease Picker) assists in the rational choice of restriction endonucleases for T-RFLP by finding sets of four restriction endonucleases that together uniquely differentiate user-designated sequence groups. With REPK, users can provide their own sequences (of any gene, not just 16S rRNA), specify the taxonomic rank of interest and choose from a number of filtering options to further narrow down the enzyme selection. Bug tracking is provided, and the source code is open and accessible under the GNU Public License v.2, at http://code.google.com/p/repk. The web server is available without access restrictions at http://rocaplab.ocean.washington.edu/tools/repk.

INTRODUCTION
Terminal restriction fragment length polymorphism (T-RFLP) analysis is a microbial fingerprinting technique capable of discriminating microbial communities quickly and relatively inexpensively (1–3). T-RFLP is increasingly used in high-throughput studies of microbial communities in combination with or even in lieu of clone library analysis (4,5). Briefly, the method involves PCR amplification of a gene of interest (often 16S rRNA) with fluorescent dye-labeled primers, followed by multiple single restriction digests done in parallel. The resulting fragments are then separated by capillary electrophoresis with an internal size standard to determine the lengths of the terminal (fluorescently labeled) fragments. Each distinct terminal restriction fragment is considered an operational taxonomic unit (OTU), thus the choice of restriction enzymes can impact the number of OTUs observed in each sample and the calculation of diversity statistics.

When analyzing uncharacterized and very diverse bacterial communities, sufficient community discrimination can often be accomplished with multiple randomly-chosen tetrameric restriction enzymes (6). However, a brief review of the literature indicates that there is still no standard in even this simplified case. We examined 26 papers (1–5,7–26) that were published between 1997 and 2007 and used T-RFLP. Of those papers, 38% used universal bacterial primers combined with a single restriction enzyme, but the choice of enzyme was not consistent. MspI was used most frequently (four studies), followed by TaqI (two studies), and one study each used AluI, CfoI, HhaI and HaeIII. Overall, only three of the 26 papers included a rationalization of enzyme selection (1,2,17).

An alternate approach to T-RFLP can be taken if the microbial community has been characterized (by clone library analysis or by prediction from previous studies) or if a particular taxonomic group is being targeted with specific primers. In this case, a more reasoned choice of restriction enzymes can be conducted. In particular, specific species or microbial taxa of interest to the researcher—particularly closely related taxa that may share some restriction sites—can often be differentiated if the proper restriction enzymes are selected.

There are, however, few resources available to narrow down the selection process. Over 600 Type II restriction enzymes are commercially available, accounting for 262 distinct specificities (27). Existing computer programs for assisting in the choice of restriction enzymes include TAP-TRFLP (28), MiCA Enzyme Resolving Power Analysis (http://mica.ibest.uidaho.edu) and TRF-CUT (29). These programs perform in silico restriction digestions of a predefined sequence database or user-provided sequences, but these results must still be manually examined to determine which enzymes are best suited to discriminate that set of sequences. CLEAVER (30), a stand alone program, provides the above features as well as the ability

*To whom correspondence should be addressed. Tel: 206 685 9994; Fax: 206 685 6651; Email: rocap@ocean.washington.edu

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
to assign sequences to taxonomic groups at multiple levels and to search for enzymes that cut one group but not another group. However, it is limited to comparing only two groups at once. Restriction Endonuclease Picker (REPK) addresses this gap by finding enzymes that are able to discriminate an unlimited number of user-designated sequence groups on the basis of their terminal restriction fragment lengths. If no single enzyme can discriminate all groups, REPK reports sets of four restriction enzymes that together are able to differentiate the groups of interest. An important component of REPK is this ability to specify the taxonomic rank of sequences to be differentiated, which is particularly useful in the case where a diverse microbial community has been characterized by clone library analysis or there is an existing database of several subgroups of sequences that amplify with the same specific primers.

SITE USAGE AND EXAMPLES

A complete manual and example input files are provided on the REPK website (http://rocaplab.ocean.washington.edu/tools/repk). The example shown in Figure 1 was prepared using REPK v. 1.0, with the following operating parameters (also the defaults): example sequence file (alignment5.txt), all commercially available Type IIP enzymes (REBASE Version 704), taxonomic rank = 1, cut-off = 5, min. fragment length = 75, max. fragment length = 900, stringency = ‘automatic’, max. missing groups = 0, max. matches returned = 100.

User input

The user must provide a trimmed FASTA-formatted file with nucleotide sequences beginning at the 5'-end of the labeled primer used for PCR amplification and ending at the 5'-end of the unlabeled primer. Sequence groups can be designated in the description line of the FASTA file, by using a delimiter to separate taxonomic rank terms or optionally taxonomic identifications can be prepended to the description line using an output file from RDPClassifier (31). Figure 1A shows a subset of the example sequence file provided on the website, alignment5.txt. Sequence groups are separated by a single underscore, and in this example ‘taxonomic rank 1’ was chosen, corresponding to the genus of these Archaea.

A selectable list of commercially available enzymes from the latest REBASE database (27) is available and is automatically updated on the first day of each month. The enzymes available for selection include primarily Type IIP enzymes, which have symmetric recognition sequences and cleavage sites. Restriction enzymes of Type IIA (having asymmetric recognition sequences) and Type IIB (cleaving both sides of the recognition sequence on both strands) are at the present time not supported by REPK, although some are included in a separate enzyme file for advanced users willing to perform some manual processing. Users should be aware that some enzymes in the REBASE database may not be suitable for T-RFLP due to methylation specificities or requirements for multiple restriction sites to be present for effective digestion.

Finally, users can define their own custom enzymes if they are not included in the standard list. The default (all standard enzymes) was used for the example in Figure 1. For computational efficiency isoschizomers are grouped by cleavage site.

The final output is refined by setting several options. Some of these, the minimum and maximum allowable fragment lengths and the maximum difference in size between two fragments that will still be considered the ‘same’ fragment, will be dependent on the specifications and resolving power of particular capillary electrophoresis systems. Users can also set the minimum threshold for the number of groups each enzyme must be able to discriminate on its own (the enzyme stringency), and the number of groups allowed to remain undifferentiated in the case that no ‘perfect’ enzyme groups are discovered.

Program operations

Sequences are first digested in both orientations by all selected enzymes to find the shortest labeled restriction fragment; these lengths are output as a table (and a downloadable tab-delimited text file, fragfile.csv), a subset of which is shown in Figure 1B. In this example, the sequences were cut by every enzyme except AasI, which resulted in full-length fragments.

Next, all terminal fragment lengths are binned within the chosen cut-off (here 5 bp) and a binary matrix of pairwise group differentiations is created. Bins containing a single sequence group yield a ‘1’, while bins containing more than one sequence group yield a ‘0’, indicating no differentiation between those groups. In the example in Figure 1, BanII failed to distinguish between sequence groups Sulfurisphaera and Thermofilum because the difference between their fragment lengths (1 bp) was less than the chosen cutoff of 5 bp (Figure 1B). However, AspLEI did distinguish between those groups because the difference in fragment lengths was 188 bp. It is not necessary for sequences from the same sequence group to have similar fragment lengths (e.g. Sulfolobus). Fragment lengths outside the boundaries set by the minimum and maximum fragment length options are binned together without regard for their actual lengths, decreasing the number of sequence groups discriminated by those enzymes (e.g. BmiI). The enzyme stringency filter is then applied to this matrix, allowing only enzymes that discriminate at least the specified fraction of sequence groups to proceed. The passing enzymes are output as a table (and a downloadable tab-delimited text file, enzmatrix.csv), a subset of which is shown in Figure 1C.

For computational efficiency, the enzymes are then sorted into ‘enzyme bins’ that produce identical differentiation patterns, although they may not produce the same terminal fragment lengths. In this example, neoschizomers AspLEI and GlaI produce different fragment lengths but the same differentiation pattern so they were grouped together for the final analysis. It is important to note that the enzyme bins are dependent on the particular sequence file and taxonomic rank selected for the analysis. That is, two enzymes may have equal discriminatory
Figure 1. Schematic summarizing the processing steps performed by REPK using program options detailed in the text, as well as subsets of example input and output files.
power for a particular set of sequence groups but for a different set of sequences, one enzyme may be much better and the two enzymes would be placed in the same bin in the first but not the second case.

Finally, groups of four enzymes (a ‘set’) are logically summed (e.g. $101 + 011 = 111$) to determine the coverage of the set, i.e. the number of sequence groups discriminated by the enzymes in the set. If this number is greater than the total number of sequence groups (less than the max. missing groups, here 0) then the set is saved. A score is calculated for each saved set and all saved sets are sorted before the highest-scoring sets are output to a text file, finalout.txt, a subset of which is shown in Figure 1D. If more than 10,000 sets are found and the enzyme stringency is set to ‘automatic’, it is incremented by 10% (decreasing the number of passing enzymes and thus enzyme sets) and the analysis is repeated. The final output reports and summarizes those enzyme sets that best discriminated the sequence groups.

The final output consists of three parts: ‘successful enzyme sets’, ‘enzyme picker key’, and ‘quick overview’. The successful enzyme sets (Figure 1D.1) consist of a list of enzyme groups in each set, and a score indicating the frequency with which each set discriminated the sequence groups. A perfect enzyme (one that discriminates 100% of the sequence groups) contributes a score of 1, so four perfect enzymes would produce the maximum score of 4. The enzyme picker key (Figure 1D.2) lists the members of each enzyme group, with neoechizomers separated by brackets. Each member of an enzyme group produces the same sequence group differentiation pattern but may differ in recognition site, terminal fragment lengths, etc. The quick overview (Figure 1D.3) histogram summarizes the frequency with which each enzyme group appears in the printed results.

After submission the program generally takes less than 1 min to complete, depending most heavily on the number of sequence groups, the number of enzymes selected and the server load, respectively. The final choice of restriction enzymes is left to the researcher, and is likely to be based on practical factors such as cost, availability, reaction conditions, methylation sensitivity or requirements, star activity and other specifics that are detailed at REBASE. An online manual detailing usage and options, bug activity and other specifics that are detailed at REBASE. An online manual detailing usage and options, bug activity and other specifics that are detailed at REBASE. An online manual detailing usage and options, bug activity and other specifics that are detailed at REBASE. An online manual detailing usage and options, bug activity and other specifics that are detailed at REBASE. An online manual detailing usage and options, bug activity and other specifics that are detailed at REBASE.

CONCLUSIONS
We found that researchers often failed to report their rationale in choosing a particular set of restriction enzymes for T-RFLP analysis, yet this choice is crucial for resolving the microbial community and interpreting the results. We provide REPK in the hope that it will allow microbial ecologists to maximize their ability to discriminate terminal restriction fragments obtained during T-RFLP and thereby take greater advantage of this powerful community fingerprinting technique.

ACKNOWLEDGEMENTS
R.E.C. was supported by NSF award OPP-0327244 and Washington Sea Grant award to J.W. Deming. G.R. was supported by NSF awards OCE-0822026 and OCE-0352190. Funding to pay the Open Access publication charges for this article was provided by National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES


