

Brief report

Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes

Zaid Abdo,^{1,2,4*} Ursel M.E. Schüette,^{3†}
Stephen J. Bent,^{3,4†} Christopher J. Williams,^{2,4‡}
Larry J. Forney^{3,4†} and Paul Joyce^{1,2,4§}

Departments of ¹Mathematics, ²Statistics and ³Biological Sciences, and ⁴Initiative in Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA.

Summary

The analysis of terminal restriction fragment length polymorphisms (T-RFLP) of 16S rRNA genes has proven to be a facile means to compare microbial communities and presumptively identify abundant members. The method provides data that can be used to compare different communities based on similarity or distance measures. Once communities have been clustered into groups, clone libraries can be prepared from sample(s) that are representative of each group in order to determine the phylogeny of the numerically abundant populations in a community. In this paper methods are introduced for the statistical analysis of T-RFLP data that include objective methods for (i) determining a baseline so that 'true' peaks in electropherograms can be identified; (ii) a means to compare electropherograms and bin fragments of similar size; (iii) clustering algorithms that can be used to identify communities that are similar to one another; and (iv) a means to select samples that are representative of a cluster that can be used to construct 16S rRNA gene clone libraries. The methods for data analysis were tested using simulated data with assumptions and parameters that corresponded to actual data. The simulation results demonstrated the usefulness of these methods in their ability to recover the true microbial

community structure generated under the assumptions made. Software for implementing these methods is available at http://www.ibest.uidaho.edu/tools/trflp_stats/index.php.

Introduction

The analysis of terminal restriction fragment length polymorphisms (T-RFLP) of 16S rRNA genes, first introduced by Liu and colleagues (1997), offers a useful means for comparing the composition of microbial communities and is widely used. In most applications of T-RFLP, fluorescently labelled universal primers that anneal to conserved regions of 16S rRNA genes in prokaryotes are used to amplify these regions from genomic DNA isolated from a microbial community. Some laboratories use fluorescently labelled forward and reverse primers, whereas others use one labelled and one unlabelled primer. The resulting mixture of amplicons is digested using restriction enzyme(s), and restriction fragments are electrophoretically resolved. Only the fluorescently labelled terminal restriction fragments are detected, and their sizes are determined by comparison to those of an internal standard consisting of DNA fragments of known length. The data obtained can be used in one of two ways. First, the identity of microbial populations in a community can be tentatively identified by comparing the sizes of terminal restriction fragments in a profile to those that would be predicted from the amplification and digestion of 16S rRNA gene sequences found in databases. Second, the presence or absence of fragments and their abundance in profiles of samples can be scored and these data can be used to compare different communities based on similarity or distance measures.

The use of experimentally determined terminal restriction fragment sizes to assess community composition has several limitations. First, the use of different fluorophors on internal size standards and DNA fragments in the sample introduces errors in accurately determining the sizes of unknown fragments (C. Shyu *et al.*, unpublished).

Received 9 May, 2005; accepted 29 August 2005. *For correspondence. E-mail zabdo@uidaho.edu; Tel. +1 208 885 6742; Fax +1 208 885 5843. Present addresses: [†]413 Brink Hall, Department of Mathematics, University of Idaho, Moscow, ID 83844-1103; [‡]414 Brink Hall, Department of Statistics, University of Idaho, Moscow, ID 83844-1104; [§]282 Life Sciences South, Department of Biological Sciences, University of Idaho, Moscow, ID 83844-3051, USA.

This results from shifts in electrophoretic mobility that are independent of fragment size and caused by chemical differences in the fluorophors used to label the DNA fragments of the internal standard versus those in the sample (C. Shyu *et al.*, unpublished). Second, any given fragment in a T-RFLP profile can arise from more than one phylotype (Marsh *et al.*, 2000). Where a phylotype is defined to be a bacterial 16S rDNA sequence type. This is inherent to the occurrence of conserved DNA sequences in 16S rRNA genes, particularly among phylogenetically related organisms. This lack of sequence variability can be partially overcome by the use of PCR amplicons whose 5' and 3' ends are labelled with different fluorophors. This effectively increases the resolution of the method by testing for the existence of two restriction site polymorphisms in each 16S rRNA gene, which potentially doubles the number of fragments produced from a sample. Third, the number of phylotypes represented in DNA sequence databases is a tiny fraction of the total that exist in nature. Consequently, most organisms found in environmental samples are novel; the sequences of their ribosomal RNA genes are unknown and not represented in databases, and their phylogeny cannot be determined from the size(s) of terminal restriction fragment(s). To classify these organisms based on their phylogeny, one must (at least) determine the nucleotide sequence of their 16S rRNA genes, align the sequence with known sequences and create a genetic distance matrix. Thus, it is often necessary to construct a clone library of 16S rRNA genes from each sample, sequence a large number of randomly chosen clones, and phylogenetically analyse the data. This is costly and time-consuming for studies that involve numerous samples.

T-RFLP, used in conjunction with clone library construction and analysis, is suitable for high-throughput, automated analyses of samples. This offers the possibility of conducting expansive studies to assess spatial and temporal changes in microbial diversity, or to systematically explore the effects of treatments and disturbances on microbial community composition and structure. Several computational tools are needed in order to execute such studies in an objective, statistically extensive and cost-effective manner. These include: (i) an algorithm for determining a baseline so that 'true' peaks in electropherograms can be identified; (ii) a means to compare electropherograms and bin fragments of similar size; (iii) clustering algorithms that can be used to identify communities that are similar to one another and (iv) a means to select samples that are representative of a cluster.

Results and discussion

Identifying 'true' peaks

An important first step in the analysis of T-RFLP data is

to identify 'true' peaks in electropherograms by distinguishing baseline 'noise' from signals resulting from fluorescently labelled DNA fragments. Current peak identification methods largely rely on the researcher's judgement in setting a fluorescence threshold below which all peaks are considered to be noise (Dunbar *et al.*, 2001; Blackwood *et al.*, 2003; Rich *et al.*, 2003). In the method we developed, the variability in the data is used to identify 'true' peaks. To compensate for such factors as variation of the concentration of DNA in various samples, the first step in the analysis is to normalize the data by calculating the relative peak areas. This is done by dividing all peak areas by the total area of all peaks in a sample profile. The data in each profile have two components: background 'noise' and 'signals' resulting from fluorescently labelled DNA fragments. Signals typically have much larger areas (and higher peak heights) than background noise and hence will add more to the variation in the data. The variance is calculated by assuming the true mean of the background fluorescence is 0. Large peaks are considered to be outliers and are progressively eliminated from the dataset. The calculation is done recursively until there are no large peaks to be removed and the remaining variation represents the background 'noise' alone. This filtering algorithm proceeds as follows:

- (1) The standard deviation for the standardized dataset is calculated assuming that the true mean is equal to 0, i.e. use the sum of the squares of the standardized data ($\sum x^2$) to calculate the standard deviation
$$\left(\sqrt{\frac{\sum x^2}{n-1}} \right)$$
. x is the peak areas associated with the noise and the signal.
- (2) Data points that have values larger than three standard deviations are identified.
- (3) These points are removed and the standard deviation of the new dataset is recalculated assuming that the true mean is equal to 0.
- (4) Steps 1–3 are repeated until there are no data points present that exceed the three standard deviation limit.
- (5) The data matrix is reduced to include only fragments that correspond to peaks that were removed as outliers ('true' peaks).

This approach automates the process of identifying the peaks, markedly reduces the time required to process data, and enables investigators to identify 'true' peaks in a more objective manner as compared with other available methods.

Comparing electropherograms

The sizes of the fragments in a sample are determined by comparing their electrophoretic mobilities with that of frag-

ments of known size in an internal standard. Each fragment length is assumed to correspond to a certain bacterial phylotype. To compare communities one must first 'bin' DNA fragments of comparable sizes that are found in different samples to compensate for analytical errors made in estimating the fragment lengths. Fragments within the same bin are considered to be phylotypes that are common to the microbial communities sampled.

Binning is performed by first pooling all the data (fragment lengths) from all the community samples of interest. These fragment lengths are then sorted so that duplicate lengths that appear in multiple samples are identified and eliminated. Hierarchical clustering (Johnson and Wichern, 1992; Johnson, 1998) is performed to identify those fragments with lengths close enough to be grouped in the same length category (or bin). Fragment lengths that are binned together are represented by their average length. Peak areas of binned fragments from the same sample are summed. This approach is similar to that described in Dunbar and colleagues (2001), though we do not enforce a limit on the number of fragments that can belong to any one bin. A matrix is created with the new representative fragment lengths being in the first column, each of the subsequent columns contains the peak areas associated with each of the samples. Every row of this matrix represents a phylotype identified by the representative fragment length in the first cell; the entries in the remaining cells represent the peak areas reflecting the abundance of this phylotype in each sample. Accordingly, this matrix has $(N+1)$ columns, where N represented the number of sampled communities. Further processing of the data involves either standardizing the peak areas again by dividing by the total area within each sample to give the relative abundance of each fragment or reformatting the matrix to a 0–1 format based on the presence or absence of a certain fragment length in a microbial community – depending on the goal of the research.

Clustering of samples

Clustering is one of the most common statistical methods utilized to group samples whose T-RFLP profiles are similar to one another. Typically, the number of different kinds of microbial communities (clusters) is subjectively determined – a process that is fraught with problems. The problem of how to objectively determine a meaningful number of clusters has been studied extensively in the statistics literature (Milligan and Cooper, 1985; Cooper and Milligan, 1988). We identified three algorithms to assess the number of clusters in a dataset (Milligan and Cooper, 1985; Cooper and Milligan, 1988; SAS Institute, 1989): Cubical Clustering Criteria (CCC) introduced by

Sarle (1983), the pseudo F statistic introduced by Calinski and Harabasz (1974), and a statistic that can be transformed to a pseudo T^2 developed by Duda and Hart (1973). Once clusters of similar microbial communities have been identified, clone libraries can be constructed from communities that are representative of all the communities within a cluster. This provides the obvious advantage that it is not necessary to prepare and analyse all the samples within a cluster because fewer samples will suffice to describe the diversity within a cluster, thus conserving time and effort while reducing costs.

The CCC compares the R^2 (the proportion of variance accounted for by the clusters) to its expected value, $E(R^2)$, which is calculated assuming the data are uniformly distributed over a hypercube (Sarle, 1983; Milligan and Cooper, 1985). Equation 1 shows the CCC index as presented in Sarle (1983):

$$CCC = \ln \left[\frac{1 - E(R^2)}{1 - R^2} \right] \frac{\sqrt{np^*}}{(0.001 + E(R^2))} \quad (1)$$

where n is the number of available observations, and p^* is the dimensionality of the variation between clusters. The optimal number of clusters is identified by plotting the CCC index against the number of clusters, and then locating the number of clusters that has the highest positive index value that is greater than 2 because an index value between 0 and 2 does not give sufficient evidence for the existence of significant clusters. Negative values that are decreasing for one or more clusters indicate unimodal or long-tailed distribution of the data. Extreme negative values indicate the presence of outliers (Sarle, 1983).

The pseudo F index (Calinski and Harabasz, 1974) is calculated as follows:

$$F = \frac{\text{trace}[B / (k - 1)]}{\text{trace}[W / (n - k)]} \quad (2)$$

where n is the number of observations in a sample, k is the number of clusters, B is the between cluster sum of squares and cross product matrix, and W is the pooled within cluster sum of squares and cross products matrix. By this method, the optimal number of clusters is determined by plotting the F index against the number of clusters. The highest F index corresponds to the optimal number of clusters.

Duda and Hart (1973) propose a ratio index [$Je(2)/Je(1)$] to identify the optimum number of clusters. $Je(2)$ is the within cluster sum of squares error when the data are divided into two clusters. $Je(1)$ is the sum of square error before division. If the within cluster sum of squares error for the two clusters is less than that for one cluster (within a certain critical value), the one cluster hypothesis is rejected in favour of the two clusters (Milligan and Cooper, 1985). This test can be transformed to a pseudo T^2 test

(SAS Institute, 1989; Johnson, 1998). The optimal number of clusters can be identified by a small value for the pseudo T^2 that is followed by a large value (SAS Institute, 1989).

We propose that the optimal number of clusters be identified based on the pseudo F and the CCC indices, if they are in agreement. If they do not agree, then the outcome of one of them is favoured if it is associated with a higher pseudo T^2 index. The choice of this strategy is due, mainly, to the ease of its automation.

Within cluster sampling

The goal of within cluster sampling is to identify a number of communities (referred to as *communities* hereafter) within a cluster that properly represent the cluster so they can be used to construct clone libraries. Here we propose four methods to achieve this aim. Two of these methods (The Pair-wise Distances method and the Maximum Variation method) utilize the coefficient of variation (CV) as a decision rule to determine the sample size. The coefficient of variation is given in Eq. 3:

$$CV = \frac{\text{Standard Error}}{\text{mean}} = \frac{(\text{Standard Deviation}) / \sqrt{n}}{\text{mean}} \quad (3)$$

where n is the sample size. The other two (the Systematic Cover and the Cover Sampling methods) use the per cent cover as a decision rule to determine the appropriate sample size. The cover is defined as the proportion of phylotypes that might be detected in a sample as compared with the total number of phylotypes detected in the whole cluster. Detailed description of each one of these methods is introduced below.

The decision to use one of these methods depends on how much of the variation in the cluster the researcher wants to explain. The lowest resolution results from using the Systematic Cover method, which focuses on richness alone in choosing a sample. The aim of this method is to identify the species that make up the communities associated with a certain cluster with no regard to how abundant these species might be. The advantage is that smaller sample sizes will be chosen using this method as compared with the other methods. The highest resolution results from choosing a sample using the Maximum Variation method, which aims at explaining as much of the variation in the cluster as possible with a disadvantage of having to deal with large sample sizes compared with the other methods. The Pair-wise Distances method and the Cover Sampling method provide intermediate resolutions. The Maximum Variation, Pair-wise Distances and Cover Sampling methods attempt to provide samples that help the researcher study the abundance of the different species that a cluster is composed of, in addition to identifying which species exist.

Pair-wise distances. The following is an iterative approach to determine the sample size that conforms to a predefined coefficient of variation.

- (1) Sample n communities repeatedly B times at random without replacement from all communities within a cluster. Because at least two communities are needed to calculate pair-wise distances, n should be greater than or equal to 2.
- (2) Calculate the pair-wise distances between the sampled communities and average the results for each of the samples drawn.
- (3) Calculate the CV associated with the sample size n .
- (4) Repeat 1 through 3, adding 1 to the previous sample size, until the coefficient of variation is less than or equal to the predefined lower limit for the coefficient of variation.

The n resulting after the last step is the minimum sample size needed to attain a CV equal or less than a predetermined value. Ideally we would want to construct the sampling distribution for the mean pair-wise distances based on all possible permutations at each sample size of interest and calculate the standard error based on these permutations. This can be computationally prohibitive. The proposed sampling without replacement approximates this sampling distribution and is less computationally intensive. The larger the number of repetitions (B), the more accurate the approximation will be.

Maximum coefficient of variation. In this method we utilize the fragment length with the maximum coefficient of variation to calculate the sample size. Our intention is to find a sample that explains as much of the maximum variability as possible. A sampling without replacement algorithm similar to the one described above is used.

- (1) Calculate the CV corresponding to the peaks associated with each of the fragment lengths in a cluster.
- (2) Identify the fragment length that has the highest CV.
- (3) Repeatedly sample n communities B times, at random without replacement, from all communities within a cluster.
- (4) Calculate the CV, for the fragment length identified in step 2, associated with the sample size n .
- (5) Repeat steps 3 and 4, adding 1 to the previous sample size until the coefficient of variation is less than or equal to the predefined coefficient of variation.

Systematic Cover method. To choose samples using this method we systematically search the communities until we identify samples that provide at least a predefined cover within each cluster. The following steps are included in the algorithm to achieve this goal.

- (1) Find the community that provides the largest cover. This community will serve as a starting point.
- (2) Search for the second community that, when combined with the first, will again provide the largest cover.
- (3) Repeat this last step until the cover of the level of interest is attained.

As indicated above, this method provides the best cover although it might not provide the right representation of the phylotype abundance in a community type. However, the advantage of this method is that it provides the smallest sample size (i.e. number of communities) needed to identify the make-up of a certain community type within a defined cover.

Cover Sampling method. This method aims to adjust the previous method to provide a viable representation of phylotype abundance in the different community types. Two stopping rules are employed for this purpose: The first is the cover (described in the previous method); the second is the frequency of attaining this cover when randomly choosing a certain number of communities (a sample size) that belong to a cluster. The aim is to choose a random sample size that covers a proportion, β , of the phylotypes in cluster with frequency π (95% for example). Similar to the methods presented under the coefficient of variation approach, all permutations of sampled communities need to be found based on a certain sample size to create the distribution of the covers associated with that sample size. A sampling without replacement strategy is also used to approximate such a distribution in this case. The sample size that meets the two rules is chosen as the optimum sample size. The algorithm is as follows:

- (1) Repeatedly sample n communities B times, at random without replacement, from all communities within a cluster.
- (2) The algorithm presented previously (see systematic method) is used to calculate the cover for each of the B samples.
- (3) The calculations are stopped if the proportion of times that the sample cover exceeds the predefined cover, β , is more than or equal π . Otherwise, steps 1 and 2 are repeated after increasing n by 1.

Sampling intensity

A method was developed to identify a lower bound on the number of microbial communities sampled so that all common microbial community types are represented with high probability. For illustrative purposes we denote a community type as *common* if its frequency among the microbial community types is greater than or equal to $p_0 = 10\%$.

The largest sample size will be required when all of the common microbial community types are barely above the

threshold required to be deemed common. In our case, this occurs when there are 10 common community types each with frequency of 10%. Any other configuration would have fewer common community types and the common types would appear in higher frequency. So, consider a sample of size n from a group of community types with 10 types of equal frequency. The probability that a particular community type is not sampled in the first draw is 0.90, and the probability that this type is missed in all n draws from the population is $(0.90)^n$, and therefore the chance that at least one of the 10 community types is missed in each of the n draws is bounded above by $(10)(0.90)^n$, making the chance that all the common types are sampled to be at least $1 - (10)(0.90)^n$. Setting this probability equal to 0.99, virtually assures that all the common types will be sampled. This is done by solving for n in the following equation $(10)(0.9)^n = 0.01$, which implies

that $n = \frac{\ln(0.01/10)}{\ln(0.90)} \approx 65$. In general, if p_0 is the minimum

frequency of a common community type and $1 - \alpha$ is the probability that all the common types are sampled, then the general formula for the sample size n is given by

$n = \frac{\ln(\alpha/p_0)}{\ln(1-p_0)}$. Yu and Williams (1991) present a special

case of this sampling formula that is concerned with one community type only. While the above formula virtually guarantees that all common types will be represented in the sample, it does not make any predictions about the sample frequency of each of these types. It is possible that a common type will appear in the sample with low frequency, simply due to sampling error. It is relatively straightforward to demonstrate that any type whose frequency in the population is 10% or higher will appear in the sample (with 0.95 probability) at frequency of 4% or higher. The general inclusion rule would be to include all types with sample frequency \hat{p} or higher where

$\hat{p} = p_0 - 1.645\sqrt{\frac{p_0(1-p_0)}{n}}$. Based on this logic we

consider a sample of size 65 to adequately sample all common community types, where a common type has frequency of at least 10%. To guard against sampling error we will include any type with sample frequency of 4% or greater as common.

Analysis of simulated data

Using the Jaccard measure of similarity (Milligan, 1981), we compared the clusters that arose from analysis of the simulated data that was generated, as described in the *Experimental procedure* section below, to the true clusters of the data (Table 1). The data showed a clear improvement in identifying the correct number of clusters as the range of similarity used for binning the fragments was increased from 0.1 bp to 1 bp regardless of the scenario

Table 1. Results of the Jaccard measure of similarity based on comparing the outcome of the simulated data to the truth.

Per cent difference	Radius	200 fragments			100 fragments			50 fragments		
		25 SD	50 SD	75 SD	25 SD	50 SD	75 SD	25 SD	50 SD	75 SD
10%	0.1 bp	Scenario 1 0% (0.000)	Scenario 2 0% (0.000)	Scenario 3 0% (0.000)	Scenario 4 0% (0.000)	Scenario 5 0% (0.000)	Scenario 6 0% (0.000)	Scenario 7 0% (0.000)	Scenario 8 0% (0.000)	Scenario 9 0% (0.000)
	0.5 bp	32.7% (0.015)	19.3% (0.013)	13.6% (0.011)	32.2% (0.015)	19.1% (0.012)	14.4% (0.011)	29.9% (0.015)	18.1% (0.012)	14.8% (0.011)
	1 bp	61.2% (0.015)	49.1% (0.016)	41.7% (0.016)	61.1% (0.015)	52.0% (0.016)	43.1% (0.016)	57.8% (0.016)	46.7% (0.016)	36.2% (0.015)
20%	0.1 bp	Scenario 10 0% (0.000)	Scenario 11 0% (0.000)	Scenario 12 0% (0.000)	Scenario 13 0% (0.000)	Scenario 14 0% (0.000)	Scenario 15 0% (0.000)	Scenario 16 0% (0.000)	Scenario 17 0% (0.000)	Scenario 18 0% (0.000)
	0.5 bp	32.6% (0.015)	19.3% (0.013)	12.4% (0.013)	37.2% (0.015)	20.0% (0.013)	13.4% (0.011)	34.2% (0.015)	18.7% (0.012)	0.6% (0.002)
	1 bp	58.9% (0.016)	48.7% (0.016)	41.4% (0.016)	63.0% (0.015)	50.8% (0.016)	45.6% (0.016)	60.1% (0.016)	47.1% (0.016)	7.6% (0.008)

Percentages represent the proportions of exact matches between clusters resulting after processing and clustering and those of the truth. These percentages are associated with 10% and 20% community difference. Numbers between parentheses represent a conservative estimate of the standard deviation of the simulation for these proportions. SD, standard deviation.

under consideration. The analysis also indicated that as the level of background noise increased, the ability to recover the true clusters decreased. The performance of the analytical procedures was about the same for communities that contained 100 or 200 true phylotypes, and somewhat worse for smaller communities with 50 true phylotypes. The proportions of perfect match for scenarios conforming to the same variation and cut-off limits were within three standard deviations from one another. This lack of significant distinction between true and calculated community types held when there was 10% or 20% difference between community types.

Examples of the distributions of the numbers of recovered clusters for scenarios 4, 5, 6 and 13, 14, 15 are presented Figs 1 and 2 respectively. These scenarios differ only in the per cent difference between the community types (10% in Fig. 1 versus 20% in Fig. 2). The data highlight the importance of choosing a correct range for peak alignment when comparing the different communities; choosing a range of 0.1 bp usually did not result in the correct number of clusters, while increasing the range to 0.5 or 1 bp improved the ability to recover the true number of clusters (Figs 1 and 2). Increasing the fragment size range from 0.5 bp to 1 bp also resulted in tighter distributions that centred on the true number of clusters (7). The ability to estimate the true number of clusters diminished as the background noise was increased from 25 to 75 (area units), and there was a tendency to overestimate the number clusters as the noise was increased. This conforms to the results deduced from Table 1. The results of analyses done on other scenarios were similar and reinforce these conclusions.

Table 2. Simulation scenarios used to validate the proposed data processing and analysis methods.

Scenario	Percent community difference (%)	Fragment sample size	Noise standard deviation
1	10	200	25
2	10	200	50
3	10	200	75
4	10	100	25
5	10	100	50
6	10	100	75
7	10	50	25
8	10	50	50
9	10	50	75
10	20	200	25
11	20	200	50
12	20	200	75
13	20	100	25
14	20	100	50
15	20	100	75
16	20	50	25
17	20	50	50
18	20	50	75

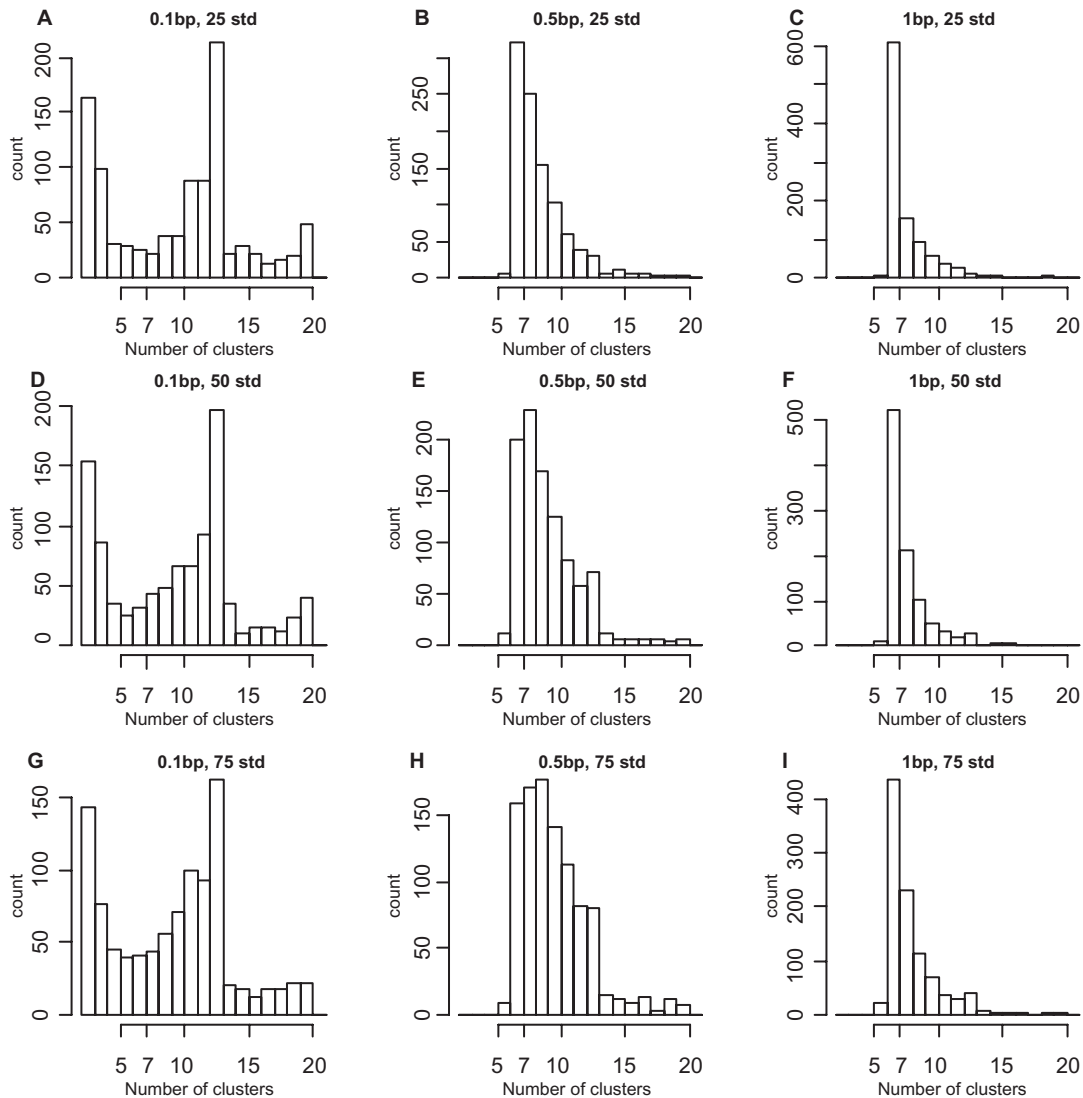


Fig. 1. Distribution of the number of clusters recovered from simulated data based on 10% difference between community types and 100 true species (scenarios 4, 5 and 6). A–C are associated with background variation based on a standard deviation of 25 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively. D–F are associated with background variation with standard deviation of 50 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively. G–I are associated with background variation based on a standard deviation of 75 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively.

Experimental procedure

Data simulation

Data with known history and assumptions must be used to evaluate methods for processing and clustering of T-RFLP data (summarized in Table 2). Such data were generated by simulation and were then analysed using the methods described above. The performance of these methods was measured by their ability to reconstruct the truth.

Our simulated microbial communities were designed so that they would belong to one of seven distinct microbial community types and form seven clusters. Each community

type (hence each cluster) was identified by a set of fragment lengths that are considered to be the phylotypes that constitute a particular community and an associated set of peak areas representing the phylotypes in the community. The sets of fragment lengths corresponding to each of the community types overlapped to accommodate the fact that different microbial community types may share similar microbes in their composition. We used two overlap settings: at least 80% (corresponding to 20% difference) and at least 90% (corresponding to 10% difference), to assess the effect of the differences among community types on recovering the true clusters. The sets of peak areas were generated using a lognormal distribution (May, 1975; Pielou, 1975) with

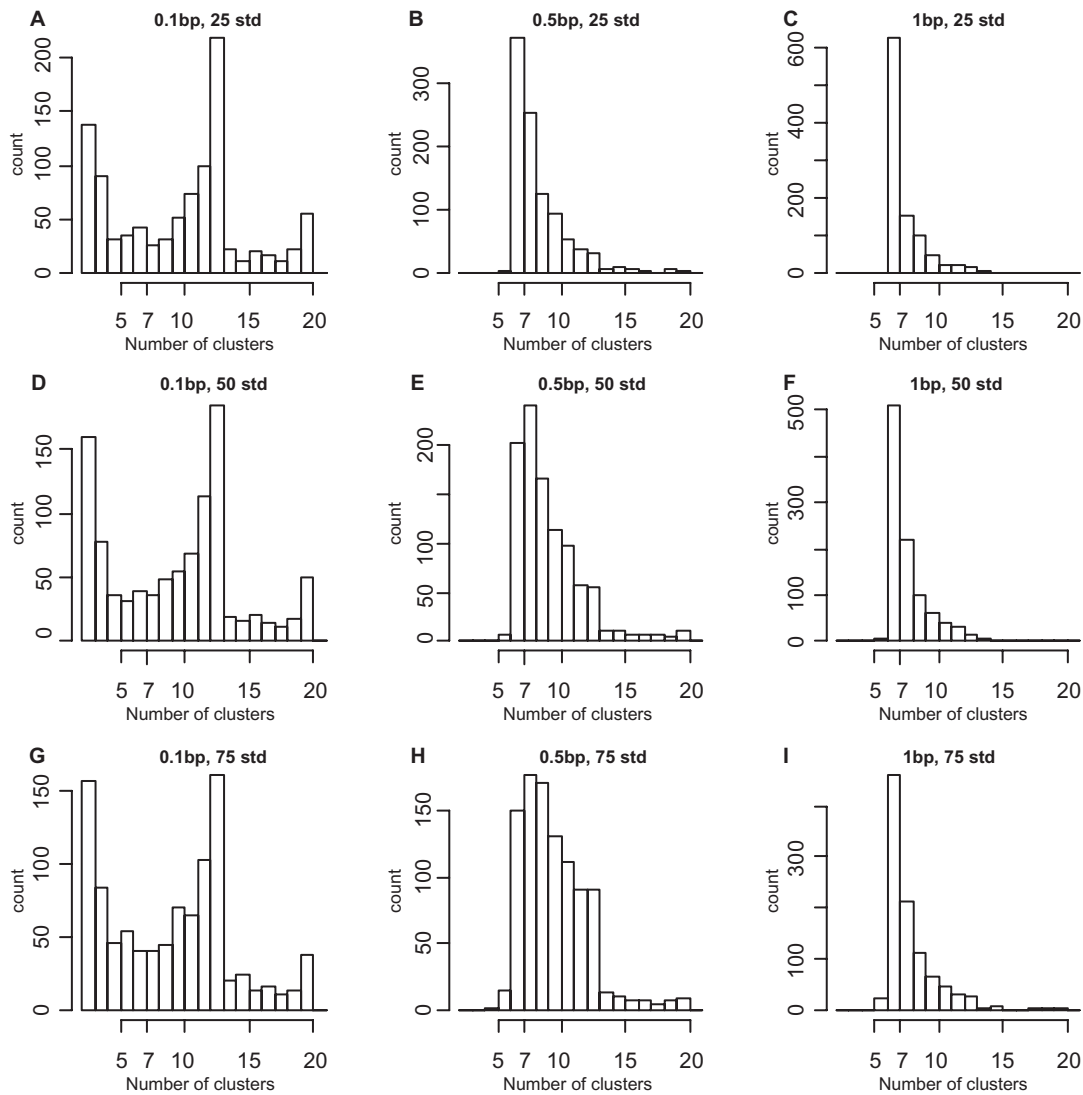


Fig. 2. Distribution of the number of clusters recovered from simulated data based on 20% difference between community types and 100 true species (scenarios 13, 14 and 15). A–C are associated with background variation based on a standard deviation of 25 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively. D–F are associated with background variation with standard deviation of 50 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively. G–I are associated with background variation based on a standard deviation of 75 from a mean equal to 0 and corresponds to clustering using 0.1 bp, 0.5 bp and 1 bp binning ranges respectively.

parameters mean-log = 4.5, and variance-log = 4 that allowed the resulting data to be within the range observed for vaginal microbial communities (X. Zhou *et al.*, unpubl. data). The number of peak areas in each of these sets matched the number of fragments in one of the fragment lengths sets, and each peak area was randomly assigned to one fragment length in the corresponding fragment lengths set. Accordingly, community types differed from one another on the basis of the phylotypes present (fragment lengths) or in terms of the abundances (peak areas) of a given phylotype found in more than one community. We tested communities with 50, 100 and 200 true phylotypes to study the effect of increasing the number of true peaks in a profile on the recovery of the true clusters.

Samples from each community type were generated by using its identifying fragment lengths sets as a mean of a multivariate normal distribution, with diagonal variance-covariance matrices, to generate fragment lengths. The diagonal elements of these matrices were equal to 0.01 (equivalent to a standard deviation of 0.1 bp), which was determined based on data from an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) provided by S. Bent (unpubl. data). Then, using the peak areas as a mean to a multivariate normal distribution, we generated the peak areas corresponding to the aforementioned fragment lengths samples. The variance-covariance matrices for these normal distributions was also diagonal though the variances on the diagonal were determined using a coefficient of variation of 0.1.

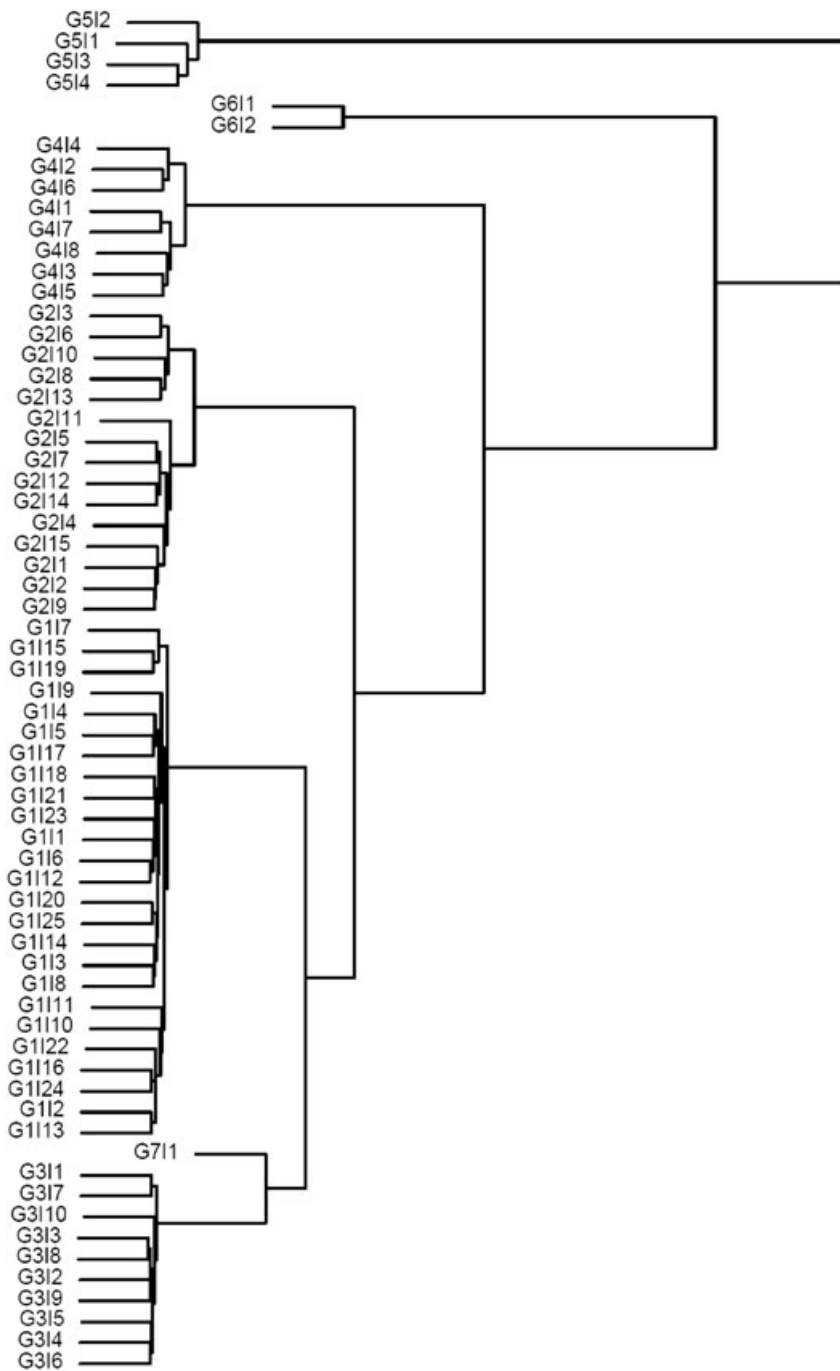


Fig. 3. Dendrogram based on raw peak data generated under the assumption that 200 true species existed in a community type and that these community types had 20% difference between them. G111, G112, ..., G711 represent the names of the clusters and the names of individuals belonging to those clusters. For example, G112 represents the second individual of the first cluster.

Accordingly, the variation in the peak areas depended on the areas of these peaks (the larger these areas, the larger the associated variance was); this is in accordance with our observations based on the dataset we studied. Figure 3 presents a dendrogram based on peak data generated under the assumption that 200 true phylotypes constitute a community type, and that the community types differ by 20%. G111, G112, ..., G711 denote individuals in specific clusters. For example, G112 represents the second individual of the first cluster.

We added noise to each of the samples by simulating 750 data points under a normal distribution with a mean equal to 0. Given that background noise is pooled and subtracted from the dataset, we contend that 750 data points representing the noise in each sample were sufficient for the purposes of this simulation. Three different background variation scenarios [standard deviations of 25, 50 and 75 (area units)] were used to study the effect of the relative noisiness of the data (as compared with the peak sizes). To simulate the discrepancy due to the inconsistencies of varying the amount of DNA

per sample or under sampling of DNA fragments in a sample we multiplied the different sample peak sizes by a factor generated using a uniform distribution with values between 0.5 and 1.5 and then dropping peaks with area smaller than 5. This reduces the number of peaks in the dataset, by eliminating those which are too small to distinguish from the background noise. We sampled 65 communities that were distributed as follows: 25 belonging to the first community type; 15 belonging to the second; 10 belonging to the third, eight belonging to the fourth; and four, two and one belonging to the fifth, sixth and seventh community types respectively. In total, 18 scenarios (two different per cent overlap between community types; three different true numbers of phylotypes associated with each community type; and three different background noise levels resulting in different relative peak-noise sizes) were tested. For each scenario 1000 datasets were simulated, for a total of 18 000 datasets.

For each dataset, the peak areas were normalized and true peaks were identified using the algorithms described above. Peaks were aligned using three size ranges, 0.1 bp, 0.5 bp and 1 bp. The outcome was three separate data matrices per original data set to yield 54 000 data matrices in total that were processed through SAS® to cluster and identify the optimum number of clusters using the above described methods.

Software and algorithms

R (R Development Core Team, 2003) functions that implement the described procedure for identifying the 'true' peaks, binning the different fragment lengths, and for within cluster sampling are available at http://www.ibest.uidah.edu/tools/trflp_stats/index.php along with a SAS® macro that automates the decision procedure used to identify the optimal number of clusters in cluster analysis.

Acknowledgements

We would thank Dr David A. Stahl for handling this manuscript and two anonymous reviewers for their thoughtful comments. Funding for this research was provided by NSF EPSCoR Grants EPS-0080935 and EPS-0132626, as well as NSF Grants DMS-0072198 and DEB-0089756 (to PJ), and NIH Grant 1P20PR016448-01 (to LF). Special thanks for X. Zhou for advice and the use data on vaginal microbial communities, and to K. Blair and his team for maintaining the University of Idaho Beowulf cluster computers.

References

Blackwood, C.B., Marsh, T., Kim, S., and Paul, E.A. (2003) Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Appl Environ Microbiol* **69**: 926–932.

- Calinski, R.B., and Harabasz, J. (1974) A dendrite method for cluster analysis. *Comm Stat* **3**: 1–27.
- Cooper, M.C., and Milligan, G.W. (1988) The effect of error on determining the number of clusters. In *Proceeding of the International Workshop on Data Analysis, Decision Support and Expert Knowledge Representation in Marketing and Related Areas of Research*, pp. 319–328.
- Duda, R.O., and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*. New York, NY, USA: John Wiley and Sons.
- Dunbar, J., Ticknor, L.O., and Kuske, C.R. (2001) Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment length profiles of 16S rRNA genes from bacterial communities. *Appl Environ Microbiol* **67**: 190–197.
- Johnson, D.E. (1998) *Applied Multivariate Methods for Data Analysts*, 1st edn. Pacific Grove, CA, USA: Duxbury Press.
- Johnson, R.A., and Wichern, D.W. (1992) *Applied Multivariate Statistical Analysis*, 3rd edn. New Jersey, USA: Prentice Hall.
- Liu, W., Marsh, T.L., Cheng, H., and Forney, L.J. (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S RNA. *Appl Environ Microbiol* **63**: 4516–4522.
- Marsh, T.L., Saxman, P., Cole, J., and Tiedje, J. (2000) Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. *Appl Environ Microbiol* **66**: 3616–3620.
- May, R.M. (1975) Patterns of species abundance and diversity. In *Ecology and Evolution of Communities*. Cody, M.L., and Diamond, J.M. (eds). Cambridge, MA, USA: The Belknap Press of Harvard University Press, pp. 81–120.
- Milligan, G.W. (1981) A Monte Carlo study of thirty criterion measures for cluster analysis. *Psychometrika* **46**: 187–199.
- Milligan, G.W., and Cooper, M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**: 159–179.
- Pielou, E.C. (1975) *Ecological Diversity*. New York, NY, USA: John Wiley and Sons.
- R Development Core Team (2003) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <http://www.R-project.org>
- Rich, J.J., Heichen, R.S., Bottomley, P.J., Cromak, K., Jr and Myrold, D.D. (2003) Community composition and functioning of denitrifying bacteria from adjacent and forest soils. *Appl Environ Microbiol* **69**: 5974–5982.
- Sarle, W.S. (1983) *The Cubic Clustering Criterion*. SAS Technical Report A-108. Cary, NC, USA: SAS Institute.
- SAS Institute Inc. (1989) *SAS/STAT® User's Guide*, Version 6, 4th edn, Vol. 1. Cary, NC, USA: SAS Institute.
- Yu, L., and Williams, C.J. (1991) Estimation of sample size and assurance probability in library screening. *Biotech* **10**: 776–777.