

## Assessment of statistical methods used in library-based approaches to microbial source tracking

Kerry J. Ritter, Ethan Carruthers, C. Andrew Carson, R. D. Ellender, Valerie J. Harwood, Kyle Kingsley, Cindy Nakatsu, Michael Sadowsky, Brian Shear, Brian West, John E. Whitlock, Bruce A. Wiggins and Jayson D. Wilbur

### ABSTRACT

Several commonly used statistical methods for fingerprint identification in microbial source tracking (MST) were examined to assess the effectiveness of pattern-matching algorithms to correctly identify sources. Although numerous statistical methods have been employed for source identification, no widespread consensus exists as to which is most appropriate. A large-scale comparison of several MST methods, using identical fecal sources, presented a unique opportunity to assess the utility of several popular statistical methods. These included discriminant analysis, nearest neighbour analysis, maximum similarity and average similarity, along with several measures of distance or similarity. Threshold criteria for excluding uncertain or poorly matched isolates from final analysis were also examined for their ability to reduce false positives and increase prediction success. Six independent libraries used in the study were constructed from indicator bacteria isolated from fecal materials of humans, seagulls, cows and dogs. Three of these libraries were constructed using the rep-PCR technique and three relied on antibiotic resistance analysis (ARA). Five of the libraries were constructed using *Escherichia coli* and one using *Enterococcus* spp. (ARA). Overall, the outcome of this study suggests a high degree of variability across statistical methods. Despite large differences in correct classification rates among the statistical methods, no single statistical approach emerged as superior. Thresholds failed to consistently increase rates of correct classification and improvement was often associated with substantial effective sample size reduction. Recommendations are provided to aid in selecting appropriate analyses for these types of data.

**Key words** | discriminant analysis, fecal coliform, microbial source tracking, similarity, statistical analysis, water quality

**Kerry J. Ritter** (corresponding author)  
Southern California Coastal Water Research Project,  
Westminster, CA 92863, USA  
Tel: +1/7143729210  
Fax: +1/8949699  
E-mail: [kerryr@sccwrp.org](mailto:kerryr@sccwrp.org)

**Ethan Carruthers**  
**Michael Sadowsky**  
Department of Soil, Water & Climate,  
University of Minnesota,  
St Paul, MN 55108, USA

**C. Andrew Carson**  
**Brian Shear**  
Department of Veterinary Pathobiology,  
University of Missouri,  
Columbia, MO 65211, USA

**R. D. Ellender**  
University of Southern Mississippi,  
Hattiesburg, MS 39406, USA

**Valerie J. Harwood**  
Department of Biology,  
University of South Florida,  
Tampa, FL 33620, USA

**Kyle Kingsley**  
**Brian West**  
Applied-Maths, Inc., 512 East 11th Street,  
Suite 207, Austin, Texas 78701, USA

**Cindy Nakatsu**  
Department of Agronomy, Purdue University,  
West Lafayette, IN 47907 USA

**John E. Whitlock**  
Division of Math and Science,  
Hillsborough Community College,  
Tampa, FL 33614, USA

**Bruce A. Wiggins**  
Department of Biology,  
James Madison University,  
Harrisonburg, VA 22807, USA

**Jayson D. Wilbur**  
Department of Mathematical Sciences,  
Worcester Polytechnic Institute,  
Worcester, MA 01609, USA

### INTRODUCTION

Many microbial source tracking (MST) methods rely on libraries of indicator organisms cultivated from known sources of fecal contamination to identify unknown

sources (see Simpson *et al.* 2002 and Scott *et al.* 2002 for a recent review of these methods). These library-based methods involve the assembly of a variety of 'fingerprints'

from indicator organisms for several known animal sources (e.g. cow, human and seagull). These fingerprints are stored as libraries that are used to compare with fingerprints from these same indicators isolated from water presumed contaminated with fecal material. In this way, the source of the unknown indicator bacterium can be identified, or at least predicted, based on similarity to members of the known-source libraries.

Library-based MST methods may be based on either genotypic or phenotypic 'fingerprints' of fecal indicator organisms, frequently *E. coli* or *Enterococcus* spp. Antibiotic resistance analysis (ARA) is a phenotypic MST method that uses profiles of resistance to various antibiotics at different concentrations (Wiggins 1996; Hagedorn *et al.* 1999; Harwood *et al.* 2000; Whitlock *et al.* 2002; Wiggins *et al.* 2003). The underlying assumption of ARA is that differential exposure of humans and animals to a variety of antibiotics will elicit specific resistance patterns for associated flora of host intestines. Subsequently, the antibiotic resistance patterns of indicators from unknown sources can be compared with a library of ARA profiles of indicators from known sources.

Rep-PCR is a genotypic method that uses the polymerase chain reaction and primers based on conserved extragenic repetitive sequences to amplify specific portions of the microbial genome (Versalovic *et al.* 1991, 1994). Following electrophoresis and staining, a banding pattern or fingerprint is revealed that can be used for strain identification. The underlying assumption of this technology is that organisms having indistinguishable banding patterns can be regarded as being identical or nearly identical, and those having similar banding patterns are regarded as genetically related. As a result, hosts for the bacteria may be identified by comparing presence/absence of bands with those from known source fingerprints (i.e. band-matching). Rep-PCR DNA fingerprinting of *E. coli* has been previously used for MST (Dombek *et al.* 2000; Carson *et al.* 2003).

Several statistical methods, including discriminant analysis (DA) (Wiggins 1996; Harwood *et al.* 2000; Carson *et al.* 2001; Whitlock *et al.* 2002), and maximum or average similarity (MS and AS) (Dombek *et al.* 2000; Carson *et al.* 2003) have been used to classify sources. These statistical approaches differ with respect to distributional assump-

tions, measures of distance or similarity, and strategies for prediction. Many approaches, such as DA and AS, take into account the central tendency and variability of each source group as a whole. Other methods, such as nearest neighbour (NN) and MS predict source membership based on similarity to an individual isolate within each source. Consequently, rates of correct classification may differ depending on the method used.

To reduce error inherent in these statistical methods, some have suggested that uncertain or poorly matched isolates should be removed from classification. This is especially relevant when false positives are a concern, such as when response to the presence of specific sources of fecal contamination results in costly management action. Some have argued that imposed thresholds decrease noise and may eliminate false positives resulting from statistical or measurement error (Wiggins, personal communication, 2003). Others have suggested that threshold values should be based on an average rate of misclassification of sources estimated from known library isolates (Whitlock *et al.* 2002) or effects of inter-gel variability on similarity between identical control isolates (Wheeler *et al.* 2002; Sadowsky, personal communication, 2003). Others advocate exact matching, removing any isolates whose fingerprints were not represented in the library (Samadpour, personal communication, 2003).

While statistical methods for classification and threshold criteria to reduce error have been used extensively with these library-based methods, there is no widespread consensus on when each of these methods or criteria is appropriate. Further, little attention has been given to the consequences associated with applying these various statistical approaches to microbial data. In this study we consider several statistical approaches for identifying source membership using these (binary) data resulting from rep-PCR and ARA fingerprints. Statistical methods were selected based on their popularity and availability in standard software packages. Following a brief review of each method, we assess the ability of each statistical technique to successfully identify sources of fecal contamination from blind test samples. Further, we investigate the use of threshold criteria to reduce rates of false positive classifications within several of the statistical methods.

## METHODS

Six libraries of indicator bacteria were used to assess the utility of various statistical approaches. Three of the libraries (A1, A2 and A3) were constructed using antibiotic resistance analysis and three were constructed using rep-PCR (R1, R2 and R3). In this study, only BOXA1R primer was used to generate fingerprints from *E. coli* isolates. Neither the three ARA libraries nor the three rep-PCR libraries were congruent in any way other than having a roughly similar library size, aliquots of the same starting fecal materials and the same sources of blind samples. Library samples consisted of swab fecal samples of human, dog, cow and seagull. Blind test samples consisted of pure (100%) sources of human, cow and seagull, composited from the same fecal samples used to create the library samples. None of the test samples contained pure dog sources. Around 60 isolates per known source sample were used to create libraries and approximately 50 isolates per unknown source sample were used for testing the rates of correct classification by the method. All libraries were constructed using *E. coli* except one (AR2) that used *Enterococcus* spp. For a more detailed description of the study design see Griffith *et al.* (2003).

Seven statistical analyses were performed on each of the six libraries. These included DA (pooled and non-pooled estimates of covariance), NN (Mahalanobis and Euclidean distances), MS (Jaccard), AS (Jaccard) and ID Bootstrap maximum similarity (Jaccard). For three of the statistical methods (DA, MS and ID Bootstrap), threshold criteria for excluding isolates from source identification were applied and resulting changes in percentage correct classification (%CC) were determined for each blind test sample. Threshold criteria for DA and ID Bootstrap were based on estimates of posterior probabilities, or the probability of correct classification. Threshold criteria for ID Bootstrap were based on likelihood of correct classification relative to chance. Threshold criteria for MS were based on a quality quotient. These criteria, along with the corresponding statistical methods, are described in more detail in the following section.

Both DA and NN analyses were performed using PROC DISCRIM in SAS (Cary, North Carolina) while MS, AS and ID Bootstrap were performed using the identification

function in BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium). These software packages were chosen for their versatility, usefulness and popularity among the microbial source tracking community. BioNumerics is a software package predominantly chosen by those using rep-PCR, while the choice of SAS is more common among those using ARA. All rep-PCR patterns were processed in BioNumerics prior to statistical analyses.

The predictive abilities of the various statistical methods for each of the library methods were assessed by calculating the percentage of correctly classified isolates (%CC) from each of the three blind test samples (human, cow and seagull). Trade-offs for applying the different threshold criteria for removing uncertain isolates were examined by calculating both the change in %CC and the proportion of the sample (i.e. percentage of isolates) eliminated from final analyses. Ties, or those isolates that were predicted by the statistical algorithm to belong to more than one source, were excluded from final analyses.

## STATISTICAL ANALYSES

### Discriminant analysis

Discriminant analysis (DA) is a commonly used technique for classifying unknown samples into predefined groups and is especially popular among microbiologists using ARA. In DA, a classification rule is developed from a calibration dataset where group membership among the samples is known (the library). In SAS's PROC DISCRIM the rule is based on estimated posterior probabilities, or the probability that an isolate belongs to a specific group. The isolate is classified into the source group yielding the highest estimated posterior probability among all source categories. Given a normal distribution and assuming equal covariances among groups, classification using posterior probabilities is equivalent to placing the observation into its closest group. By default, distances to groups are defined by Mahalanobis distance, which takes into account both distance of each observation to the mean and the variability within the group. The 'pool=yes' option in PROC DISCRIM estimates a single covariance structure for all groups while the 'pool=no' option allows

for estimation of covariances separately for each group (for more detailed description see SAS/STAT User's Guide, Vol. 1. Chapter 20).

In SAS's PROC DISCRIM a 'threshold' option is available for excluding observations from classification in discriminant analyses based on a minimum threshold for posterior probabilities. If an isolate's posterior probability for its predicted group falls below the threshold value, the isolate is classified into an 'other' category and excluded from classification into one of the known groups.

### Maximum similarity

Maximum similarity (MS) is another commonly used statistical algorithm that is particularly popular among those using rep-PCR. In MS, observations are classified into the group to which its closest or most similar known member belongs. BioNumerics offers several alternative measures of similarity, including Jaccard, Dice and simple matching. We chose to use the Jaccard similarity coefficient for our analyses because Jaccard targets only those bands that are present in at least one of the pairs being compared, ignoring potentially large numbers of missing bands that may dilute or mask differences. The Jaccard similarity coefficient is given by:

$$\frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (1)$$

where  $N_{AB}$  is the number of shared bands,  $N_A$  is the total number of bands in pattern A, and  $N_B$  is the total number of bands in pattern B.

BioNumerics also offers a quality quotient (QQ) option that qualifies the relative uncertainty of correct classification for each isolate and may be used for eliminating potential false positives. The QQ is the ratio of the average distance between the unknown and the library source members and the average internal distance where:

$$QQ = D_{UN}/D_{LU} \quad (2)$$

$$D_{UN} = \frac{\sum_{i=1}^n (1-s_i)}{n} \quad (3)$$

$$\text{and } D_{LU} = \frac{2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n (1-s_{i,j})}{n(n-1)}, \quad (4)$$

$n$  is the number of entries in the library unit and  $s$  is the similarity. If  $QQ \leq 1$ , it can be inferred that the unknown fits into the library source as well as or better than other members of that source group; whereas if  $QQ \gg 1$ , a poor fit is indicated. Qualitative scores or 'grades', each representing a numerical range of the QQ for each isolate, may be exported to a data file along with similarity and scores and predicted source classifications. The numerical range of these grades is as follows: 0–0.5 = A, 0.51–1.0 = B, 1.01–1.5 = C, 1.51–2 = D and  $>2$  = E. Recently a script file has been made available to BioNumerics users that allows numerical values to be exported as well; however, the program was not available at the time of this study.

### Average similarity

Average similarity (AS), a common alternative to MS, is available in BioNumerics, and is also very popular among microbiologists using rep-PCR. Rather than classifying unknown isolates into source groups based on proximity to a single known isolate, AS classifies an unknown isolate into the source category yielding the largest average similarity to all library isolates within that source.

### Nearest neighbour

Nearest neighbour (NN) is a nonparametric alternative to DA and is available in SAS using PROC DISCRIM. Source assignment is based on nonparametric estimates of posterior probabilities based on  $k$ -nearest neighbours. With the  $k = 1$  option, as was specified in this study, NN assigns source membership based on proximity to closest known individual. Therefore, NN is analogous to MS, where NN defines proximity using Mahalanobis or Euclidean distances, rather than Jaccard, Dice or simple matching similarity measures.

### ID Bootstrap

ID Bootstrap is a script file designed to improve correct classification rates and has only recently been made

available to BioNumerics users (see [www.applied-maths.com/bn/bn.htm](http://www.applied-maths.com/bn/bn.htm)). The script applies a bootstrap algorithm to a matrix of similarity values in order to estimate the likelihood of obtaining an observed similarity score relative to chance. A similarity matrix is first calculated for all known samples, each of which is assigned to a particular group, and then each unknown is compared with each group of known samples, providing an average (or maximum) similarity for each unknown to each group. Each unknown is tentatively identified as belonging to the group with which it has the highest average (or maximum) similarity. The distribution of similarities between group non-members and group members is then approximated by resampling, with replacement, group members and non-members. Each bootstrap iteration involves the random selection of 30 or more group members and a single non-member. The proportion of bootstrap iterations in which the unknown is more similar to the re-sampled group than the known non-member approximates the likelihood that the unknown belongs to the group relative to chance. In this study, the script was applied to MS and 1,000 iterations were specified. Thresholds for ID bootstrap were based on estimated relative likelihoods of correct identification.

## RESULTS

The ability of each MST method to correctly predict source membership relied heavily on which statistical analysis was used (Table 1). Depending on the choice of statistical method, %CC could be quite high; at other times, %CC fell well below chance. In the half of the cases where blind test samples were classified, %CC changed by more than 40 percentage points as a result of varying the statistical approach. In one case %CC increased from 0% to 90% by switching the statistical analysis from average similarity to maximum similarity. Even simply changing the measure of distance often resulted in substantial changes in the proportion of correctly identified isolates. For example, in NN, changing from Euclidean to Mahalanobis distance increased the percentage of correctly identified human isolates for researcher A1 from 29% to 75%.

The consistently high %CC for human and seagull across the various statistical methods for researcher A2 was due primarily to the fact that variability of fingerprints was low within each of these unknown samples. In fact, all isolates analysed in the human blind sample for this researcher had exactly the same antibiotic fingerprint (so that only 100% or 0% were possible for %CC) and only a handful of fingerprints were found among the blind seagull isolates.

Despite the large differences in correct classification rates among the statistical methods, no single statistical approach emerged as superior across any of the MST methods or sources. Improvements in correct classification rates for one source as a result of applying an alternative statistical analysis were often followed by a decrease in correct classification rates for another source. For example, for researcher R2, switching statistical analysis from DA (non-pooled) to NN (Mahalanobis) resulted in a 60 percentage point increase in %CC for human and a decreased %CC for cow and seagull by approximately 40 percentage points.

In the majority of the cases, the 'right' choice of statistical method improved substantially the researcher's ability to identify unknown sources of fecal contamination. However, in some cases, none of the statistical methods was particularly satisfying. In many cases, %CC fell below 60%, regardless of which statistical method was applied. In one case (A1), the maximum %CC for human was less than 4% regardless of the statistical approach (less than expected by chance alone).

Attempting to reduce false positives by applying various threshold criteria to exclude uncertain isolates from classifications did not always result in improved %CC (see Tables 2–4). Using thresholds based on estimated DA posterior probabilities between 80% and 95% resulted in changes in %CC for individual samples anywhere from –16 to +33 percentage points depending on the library and source (Table 2). Thresholds based on QQ scores between B and D (there were no As) resulted in changes in %CC for individual samples between –49 to +44 percentage points (Table 3). For ID Bootstrap, relative likelihood thresholds between 80% and 95% resulted in changes in %CC of between –35 and +35 percentage points (Table 4). For one researcher (A3), estimated

**Table 1** | Percentage correct classification for seven statistical methods

Source	StatMethod	RepPCR			ARA		
		R1	R2	R3	A1	A2	A3
Human	DA <sup>P</sup>	50	66	27	2	100	9
	DA <sup>NP</sup>	35	20	12	0	100	25
	NN <sup>M</sup>	43	<b>80</b>	33	2	100	<b>33</b>
	NN <sup>E</sup>	57	74	<b>51</b>	0	100	13
	Max. sim	65	74	22	<b>3</b>	100	13
	Ave. sim	<b>74</b>	66	12	0	100	17
	ID Boot max	65	74	22	<b>3</b>	100	13
Cow	DA <sup>P</sup>	30	16	<b>71</b>	25	51	51
	DA <sup>NP</sup>	39	<b>58</b>	39	70	59	33
	NN <sup>M</sup>	17	20	27	<b>75</b>	27	38
	NN <sup>E</sup>	31	16	24	29	41	35
	Max. sim	53	18	47	29	<b>90</b>	37
	Ave. sim.	<b>59</b>	0	53	4	0	<b>60</b>
	ID Boot max	53	18	47	29	<b>90</b>	39
Gull	DA <sup>P</sup>	53	8	<b>34</b>	41	93	57
	DA <sup>NP</sup>	21	<b>42</b>	23	44	93	25
	NN <sup>M</sup>	47	0	28	33	93	47
	NN <sup>E</sup>	62	8	38	33	98	52
	Max. sim	<b>67</b>	5	29	36	98	<b>65</b>
	Ave. sim.	33	0	0	<b>67</b>	<b>100</b>	57
	ID Boot max	<b>67</b>	5	29	35	98	<b>65</b>

Bold=method with highest percentage correct classification.

DA<sup>P</sup>=Discriminant analyses, pooled covariance.

DA<sup>NP</sup>=Discriminant analyses, non-pooled covariance.

NN<sup>M</sup>=Nearest neighbour, Mahalanobis distance.

NN<sup>E</sup>=Nearest neighbour, Euclidean distance.

R1, R2 and R3=rep-PCR.

A1, A2 and A3=ARA.

**Table 2** | Effect of imposing threshold value for posterior probabilities (discriminant analyses, pooled covariance)

Researcher	%Thresh	Human		Cow		Gull	
		%CC	%Remain	%CC	%Remain	%CC	%Remain
R1	None	<b>50</b>	100	<b>30</b>	100	<b>53</b>	100
	80	<b>48</b>	78	<b>42</b>	70	<b>53</b>	68
	90	<b>48</b>	78	<b>45</b>	57	<b>53</b>	57
	95	<b>69</b>	54	<b>46</b>	52	<b>29</b>	32
R2	None	<b>66</b>	100	<b>16</b>	100	<b>8</b>	100
	80	<b>73</b>	88	<b>0</b>	80	<b>6</b>	68
	90	<b>81</b>	62	<b>0</b>	74	<b>7</b>	56
	95	<b>81</b>	62	<b>0</b>	74	<b>10</b>	42
R3	None	<b>27</b>	100	<b>71</b>	100	<b>34</b>	100
	80	<b>26</b>	78	<b>78</b>	73	<b>36</b>	77
	90	<b>28</b>	59	<b>84</b>	61	<b>31</b>	68
	95	<b>26</b>	55	<b>85</b>	51	<b>24</b>	62
A1	None	<b>2</b>	100	<b>25</b>	100	<b>41</b>	100
	80	<b>0</b>	98	<b>25</b>	98	<b>49</b>	80
	90	<b>0</b>	98	<b>26</b>	95	<b>50</b>	78
	95	<b>0</b>	85	<b>32</b>	73	<b>51</b>	76
A2	None	<b>100</b>	100	<b>51</b>	100	<b>93</b>	100
	80	<b>100</b>	100	<b>67</b>	29	<b>93</b>	100
	90	—	0	<b>60</b>	24	<b>93</b>	61
	95	—	0	<b>56</b>	22	<b>93</b>	61
A3	None	<b>9</b>	100	<b>51</b>	100	<b>57</b>	100
	80	<b>0</b>	49	<b>54</b>	44	<b>85</b>	49
	90	<b>0</b>	36	<b>62</b>	38	<b>88</b>	32
	95	<b>0</b>	26	<b>61</b>	33	<b>90</b>	19

Bold=same or increased as a result of threshold.

Bold italic=decreased as a result of threshold.

**Table 3** | Effect of imposing cut-off based on quality quotient (maximum similarity $\geq$ QQ)

Researcher	QQ	Human		Cow		Gull	
		%CC	%Remain	%CC	%Remain	%CC	%Remain
R1	None	<b>65</b>	100	<b>53</b>	100	<b>67</b>	100
	B	<b>16</b>	15	<b>46</b>	55	<b>94</b>	37
R2	None	<b>74</b>	100	<b>18</b>	100	<b>5</b>	100
	C	<b>74</b>	100	<b>18</b>	100	<b>5</b>	97
	B	<b>38</b>	16	<b>0</b>	52	<b>4</b>	59
R3	None	<b>22</b>	100	<b>47</b>	100	<b>29</b>	100
	B	<b>40</b>	31	<b>56</b>	34	<b>27</b>	74
A1	None	<b>3</b>	100	<b>29</b>	100	<b>36</b>	100
	D	<b>0</b>	98	<b>29</b>	100	<b>36</b>	100
	C	<b>0</b>	98	<b>29</b>	100	<b>34</b>	98
	B	<b>0</b>	86	<b>73</b>	27	<b>77</b>	29
A2	None	<b>100</b>	100	<b>90</b>	100	<b>98</b>	100
	C	<b>100</b>	100	<b>88</b>	85	<b>98</b>	100
	B	<b>100</b>	100	<b>83</b>	30	<b>98</b>	100
A3	None	<b>13</b>	100	<b>37</b>	100	<b>65</b>	100
	D	<b>16</b>	81	<b>38</b>	98	<b>65</b>	100
	C	<b>21</b>	45	<b>29</b>	76	<b>67</b>	69
	B	17	19	0	7	<b>100</b>	38

Bold=same or increased as a result of threshold.

Bold italic=decreased as a result of threshold.

relative likelihood from ID Bootstrap were below 80% for all isolates.

Where increases in %CC occurred as a result of applying a threshold for classification, large proportions of isolates were often eliminated from final analyses. Increases in %CC above 10% as a consequence of applying a posterior probability threshold in DA

often resulted in the elimination of more than half of the sample. Similarly, for exclusion of isolates based on QQs, increases in %CC above 10% were typically accompanied by removal of 60–75% of the sample isolates. For ID Bootstrap thresholds, %CC above 10% resulted in the removal of 60–95% of the sample.

**Table 4** | Effect of imposing threshold value for estimated relative likelihood (ID Bootstrap, maximum similarity)

Researcher	%Thresh	Human		Cow		Gull	
		%CC	%Remain	%CC	%Remain	%CC	%Remain
R1	None	<b>65</b>	100	<b>53</b>	100	<b>67</b>	100
	80	<b>86</b>	55	<b>61</b>	61	<b>81</b>	37
	90	<b>91</b>	28	<b>30</b>	20	<b>90</b>	23
	95	<b>100</b>	20	<b>25</b>	8	<b>100</b>	12
R2	None	<b>74</b>	100	<b>18</b>	100	<b>5</b>	100
	80	<b>40</b>	30	<b>17</b>	96	<b>7</b>	61
	90	<b>56</b>	18	<b>17</b>	94	<b>5</b>	43
	95	<b>100</b>	4	18	88	<b>8</b>	27
R3	None	<b>22</b>	100	<b>47</b>	100	<b>29</b>	100
	80	<b>43</b>	29	<b>46</b>	60	<b>25</b>	69
	90	<b>25</b>	16	<b>53</b>	32	<b>25</b>	46
	95	<b>20</b>	10	<b>43</b>	15	<b>20</b>	14
A1	None	<b>3</b>	100	<b>29</b>	100	<b>35</b>	100
	80	<b>0</b>	81	<b>0</b>	39	<b>22</b>	20
	90	<b>0</b>	81	<b>0</b>	9	<b>22</b>	20
	95	<b>0</b>	59	—	0	<b>0</b>	9
A2	None	<b>100</b>	100	<b>90</b>	100	<b>98</b>	100
	80	100	100	<b>92</b>	60	<b>98</b>	100
	90	100	100	<b>100</b>	20	<b>98</b>	100
	95	100	53	<b>100</b>	5	<b>98</b>	98
A3	None	<b>13</b>	100	<b>39</b>	100	<b>65</b>	100
	80	—	0	—	0	—	0
	90	—	0	—	0	—	0
	95	—	0	—	0	—	0

Bold=same or increased as a result of threshold.

Bold italic=decreased as a result of threshold.

—=relative likelihood for classification of all isolates fell below threshold.

Consequences for both %CC and percentage exclusion of sample isolates resulting from applying threshold criteria were not constant across sources. In fact, increases in %CC for one source were often accompanied by decreases in %CC for another source. For researcher R2 using rep-PCR, applying a threshold of 95% for estimated posterior probability for DA increased %CC for human from 66% to 81% and for gull from 8% to 10%, yet decreased %CC for cow from 16% to 0%. For this same researcher, applying a 95% threshold eliminated 38% of the human isolates, 58% of the seagull isolates and 36% of the cow isolates. Similar trade-offs occurred when applying other threshold criteria.

## DISCUSSION

The results of this study demonstrate that the statistical method used to predict host source membership can significantly affect the ability of library-based methods to correctly identify sources of fecal contamination. While no one statistical method consistently performed better than another across all MST methods and sources, some statistical approaches were better suited than others for identifying certain patterns in the data and for assigning source membership based on those patterns. Choice of similarity or distance measure defined fingerprint distribution within each library and these distributions, in turn, affected the ability of the statistical algorithm to differentiate between sources. Clustering of sources, multimodality, overlap and variability of fingerprints within sources had substantial effects on which statistical tool performed best.

The major challenges for each of the statistical methods were lack of library representativeness (no apparent match of unknown) and lack of significant host specificity (overlap between groups and lack of discriminatory characters). Although not addressed in this study, the issue of representativeness is one of the most important issues for library-based MST methods and has recently been addressed with respect to ARA of *Enterococcus* spp. (Wiggins *et al.* 2003). Representativeness is the link between sampling and the successful use of statistical

methods. Even in this particular study where fingerprints from the unknown isolates should be more similar to those of the library isolates because the same sources of fecal material were used in both the test and library samples, factors such as differential survival (Gordon 2002) or cultivability of indicators (Bissonnette *et al.* 1975; Lleo *et al.* 2001; Boualam *et al.* 2002) may have affected the representativeness of the various libraries. While libraries were constructed directly from fecal material (primary habitat for indicators), indicators from the blind samples were isolated from water (secondary habitat). Gordon *et al.* (2002) suggested that a significant change in population structure can take place between these two habitats for *E. coli*. In this study, many of the patterns observed in the blind test samples were distinct from those represented in the library.

Because most of the statistical strategies investigated in this study relied upon either central tendency or similar patterns within the same source, host specificity and similarity of patterns within a source are key to the success of these statistical methods. For the datasets observed in this study there was quite a bit of overlap between various sources (lack of host specificity). Such overlap has been documented by others. For example, McLellan *et al.* (2003) noted that rep-PCR banding patterns for seagulls overlapped with those obtained from sewage samples. Gordon (2001) has suggested that *E. coli* lack sufficient host specificity to be useful in MST methods.

While various threshold criteria have been proposed for decreasing false positive errors and thus improving %CC, our investigation showed that such actions do not always produce favourable results. One explanation for the decrease in %CC as a result of applying a particular threshold criteria is the presence of 'subtypes' of bacteria within a given host source group. In many of the libraries in this study, we saw clustering of fingerprints into multiple subgroups within the same source and these clusters were often interspersed among multiple clusters within other sources. These subgroups often belonged to the same individual within a given source. As a result test samples often contained isolates that were more similar to another source group or subgroup than the one to which they truly belonged. Applying a threshold value then discarded those isolates that were more poorly matched

yet correctly classified while keeping those more similar to the true source yet incorrectly classified.

## RECOMMENDATIONS

Given that choice of statistical method is important and that no statistical method emerged as superior to all others, how do we choose among the various statistical approaches available? In order to choose the approach likely to yield the highest %CC, we recommend a careful examination of the library, including visualizing the data, calculating estimates of predictive success and assessing library representativeness. Repeating this process across the various statistical methods provides the researcher with a measure of potential success and a comparative basis from which to select the most appropriate statistical method. Finally, we recommend options for managing classification ties and fingerprint patterns resulting in all-zeros, which may distort or bias %CC.

### Visualizing the data

Visualizing fingerprint patterns in the library prior to classifying unknown samples identifies the degree of overlap among fingerprints from different sources, which in turn affects the potential of the statistical method. Many software packages (including SAS and BioNumerics) offer several alternatives for graphical representations of the data. Some of the most popular include canonical discriminant analysis and cluster analyses. When using these visual tools, it is important to note that both are dependent on the choice of distance or similarity measure. Canonical discriminant analyses use Mahalanobis distances while cluster analyses may use alternative measures such as Jaccard, Dice and simple matching. Visual display should reflect the distances employed by the particular statistical algorithm under investigation.

In addition to clustering trees and canonical plots where appropriate, we recommend multidimensional scaling (MDS) for visualizing distances between isolates for various measures of distance. MDS produces a 'map' in two (or more) dimensions such that the distance between isolates best approximates the relative distances between

isolates, allowing one to visualize proximity of isolates both within and among source groups. Therefore, MDS plots often allow one to see which measure of distance or similarity is likely to produce the most separation (or clustering).

MDS plots can also point to whether a statistical tool based on averaging or one based on proximity to singletons is most appropriate. In general, where source patterns are clustered about a central location, with limited overlap, the statistical methods based on averages, such as DA or AS tend to perform well. For sources where fingerprints are patchy and clustered around several locations, NN or MS techniques are often a better alternative. For example, the MDS plot for A2 using Jaccard distance shows that, for cow, there are multiple subgroups within the known cow data and these overlap with other known sources (Figure 1). Therefore, maximum similarity was more appropriate than average similarity (%CC for MS = 90% and %CC for AS = 0%). Of course the usefulness of MDS plots depends on how representative and complete the library is.

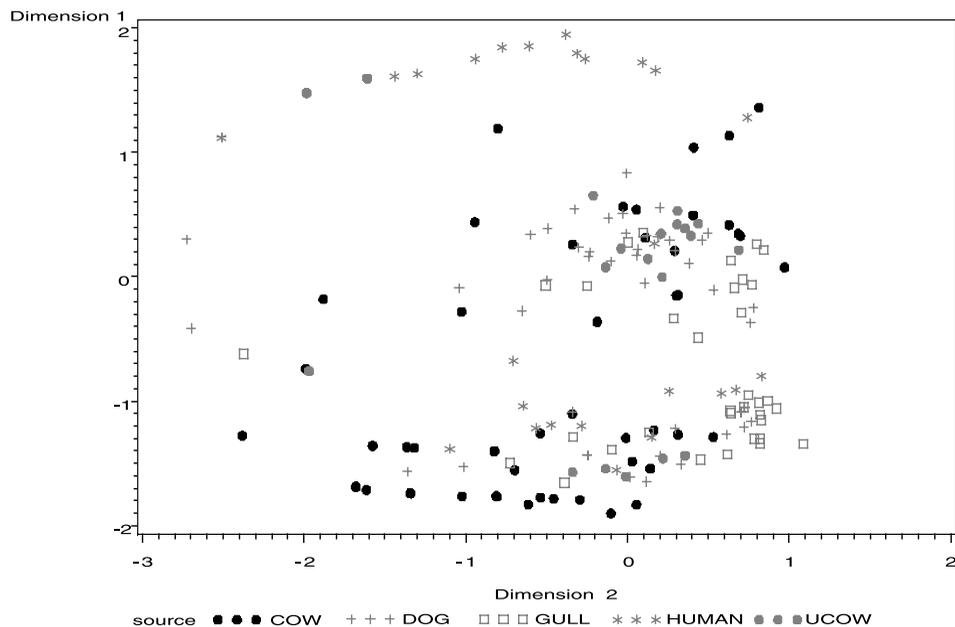
### Estimating predictive success

Another tool to aid the choice of statistical analyses is to compare estimates of predicted success across each of the statistical methods. Among the readily available methods for estimating %CC, we recommend the jackknife. Assuming simple random sampling of the population, jackknifing provides unbiased estimates of correct classification and is readily available in both BioNumerics and SAS. In SAS, jackknife estimates are obtained using the cross validate option in PROC DISCRIM and is preferred over the resubstitution estimates (the default in PROC DISCRIM). In PROC DISCRIM one can also obtain jackknife estimates for percentage correct classifications based on thresholds.

Although not reported in this study, jackknife estimates were calculated using the standard software across each of the statistical methods for each library. For nearly all libraries, jackknife estimates of %CC were higher than those observed. Inflation may be due, in part, to library construction. In this study libraries depended on subsampling of isolates from individuals or groups of

## Jaccard MDS plot (A2)

KNOWN AND UNKNOWN COW



- **Known sources: cow, dog, gull, human**
- **Unknown source: Ucow**

**Figure 1** | Example MDS plot.

individuals within sources rather than on simple random sampling usually required for these standard jackknife procedures. Wiggins *et al.* (2003) suggest an alternative jackknife analysis to hold out entire feces instead of individual isolates that may be more appropriate for sub-sampling designs. Where estimation of predicted success differed substantially from that observed, test fingerprints were underrepresented in the library.

### Assessing library representativeness

Because statistical algorithms used to identify sources rely on similarity to known patterns, it is crucial that libraries sufficiently represent the population of fingerprints within each source. We recommend a thorough investigation of library representativeness prior to classifying unknown sources of fecal contamination. Wiggins *et al.* (2003) offer

several methods for assessing library representativeness. One method compares jackknife estimates of average rates of correct classification within each source to those estimated by the resubstitution method. Recall that the resubstitution method uses all isolates both to build the library and to predict source membership, thereby estimating how well the library can predict itself. A representative library, then, would provide jackknife estimates of average rates of correct classification comparable to those obtained by resubstitution. A second method for assessing library representativeness, also suggested by Wiggins *et al.*, is the 'hold out' method, where a portion of the known samples is used to make up the library (i.e. a calibration dataset) and the remaining portion is used to estimate average rates of correct classifications (i.e. a validation dataset). Again average rates of correct classification are compared with those estimated via resubstitution.

Another potential statistic that may help to test for representativeness is the bootstrap analyses applied to the hold out method. By resampling with replacement the distribution against which the test is performed, we can get a better idea of how variable the rates of correct classification are within each source category. This is especially important when one is using maximum similarity to identify unknowns, since a single outlier can drastically affect the result. The magnitude of variability found in repeated simulations would inversely relate to the representativeness of the library. Further, large variabilities in %CC may indicate multiple clustering within each source population or provide evidence of sampling bias.

### Managing classification ties

Ties occur when the classification rule inherent in the statistical method assigns a given fingerprint to more than one source group. Different software packages have different ways of handling ties and these mechanisms are often unknown or overlooked. In SAS's PROC DISCRIM, for example, the default places ties into an 'other' category, allowing the researcher to decide on the appropriate action. Because SAS calculates %CC based on the entire sample (including ties), excluding these data from the final analyses requires the researcher to adjust reported rates of classification (i.e. the denominator). In contrast, BioNumerics' default systematically assigns ties according to the order in which the source groups are listed in the library. Identifications involving numerous ties are likely to result in severe bias. In this study we found a 30-percentage point increase in %CC for one researcher simply by removing the ties prior to applying MS in BioNumerics. Finally, BioNumerics' jackknife procedure for predicting %CC by default assigns ties to their own known source group, resulting in inflated estimates of predicted success. However, in the more recent version of BioNumerics (version 3.0) there is an option designed to reduce bias towards a single source by spreading ties equally among the source groups through random assignment.

Several options exist for handling ties. Selecting which option to use will depend on available auxiliary

information and penalties associated with false positive or false negative errors. One option is to exclude ties from final analyses. This option should be selected when positive identification of known sources may result in costly penalties relative to consequences of false negatives. For example, beaches can be incorrectly listed as unsafe for swimmers, resulting in severe penalties for sewage dischargers and lost revenue to surrounding business communities, though risk of illness may be low.

A second option for handling ties is to systematically assign ties to a likely source group based on auxiliary information, prior belief or consequences of false identification. PROC DISCRIM has an option for specifying prior probabilities that will bias assignment of ties towards a particular source. For example, previous research may support a distribution of sources within the target area that favours one source over another. By specifying prior probabilities for source classification, one can weigh source assignment towards one particular source more than another. When reliable and current auxiliary data exist, we recommend using this option regardless. In addition to using prior information, systematic assignment of ties could be achieved visually through dendrograms. This procedure is particularly useful for those sources containing multiple clusters or subtypes of fingerprint. Finally, ties can be assigned to a particular source group based on expert knowledge and/or familiarity with the study area. Samples obtained downstream from a dairy farm, for instance, are more likely to belong to cow source group than seagull.

### Managing all-zero patterns

In many cases, fingerprint profiles may yield all-zero patterns. In rep-PCR, all-zero patterns are usually attributed to measurement or laboratory error and are eliminated from further analyses. In ARA, however, all-zero patterns may result from bacteria that are not resistant to any of the antibiotics at the concentrations tested, and thus may contain useful discriminating characteristics. The all-zero patterns present a problem when using similarity measures which target only the number of 'ones' that match in the binary datasets, ignoring all 'zeros' (e.g. Jaccard and Dice).

For example, the Jaccard similarity applied to two isolates whose fingerprints contain all zeros will have a similarity score of '0', yet they are identical. In fact, an all-zero isolate will have zero similarity with all isolates regardless of their pattern. We recommend several alternatives for matching all-zero patterns in ARA data. One alternative is to use exact matching if there is at least one other all-zero value in the library. If there is more than one source in the library that contains an all-zero pattern then one could apply those strategies previously mentioned for managing ties, including using auxiliary information and visualization. Alternatively, one could explore other similarity measures that take into account zeros (e.g. Euclidean distance or simple matching). For the ARA datasets here we chose to report %CC after removing all-zero fingerprints prior to applying MS and AS to Jaccard similarity scores as there was little difference in %CC with all-zeros included.

## CONCLUSION

This study demonstrated that choice of statistical method dramatically influenced the ability of the MST method to correctly predict source membership. Successful predictions relied heavily on which statistical analysis was used. Because no single statistical method emerged as superior across all libraries, we recommend several tools to assist in choosing between available methods, including visual aids, jackknife estimates and library assessments. Recognizing that the default procedures for ties of various software may bias group assignment and estimates of predicted success, we offered several alternatives for dealing with ties. In addition, we suggested techniques for managing fingerprints containing all-zero patterns. Failure of some statistics was attributed to lack of host specificity and library representativeness.

While eliminating isolates based on thresholds can reduce false positives and therefore increase rates of correct classification, results of this study indicate that caution should be exercised in their application. First, correct classification rates did not always improve with threshold. Second, improvement in %CC for one source

was often accompanied by a decline in %CC for another. Finally, since thresholds do not eliminate all sources evenly, thresholds may bias estimates of relative proportions of sources of fecal contamination. Therefore, the choice of whether or not to use a threshold should depend on penalties resulting from false positives and the goals of the study.

Future research is needed to improve statistical techniques for discriminating between sources based on antibiotic resistance or banding patterns. In particular, the measure of similarity is crucial to effectively discriminate between source groups. Additional similarity measures should be evaluated for appropriateness along with the exploration of novel measures of genetic similarity. Finally, statistical procedures for classification may be improved with the removal of nuisance variables (e.g. bands or antibiotics) that do not contain information useful for discriminating between sources. Regardless of which statistical methods are used, testing the correct classification of isolates that are not in the library (proficiencies or blind samples) ultimately determines the utility of any microbial source tracking method. This process should be incorporated into the development of each new library, and should take into consideration factors such as proficiency sample size and pseudoreplication.

## REFERENCES

- Bissonnette, G. K., Jezeski, J. J., McFeters, G. A. & Stuart, D. G. 1975 Influence of environmental stress on enumeration of indicator bacteria from natural waters. *Appl. Microbiol.* **29**, 186–194.
- Boualam, M., Mathieu, L., Fass, S., Cavard, J. & Gatel, D. 2002 Relationship between coliform culturability and organic matter in low nutritive waters. *Wat. Res.* **36**, 2618–2626.
- Carson, C. A., Shear, B. L., Ellersieck, M. R. & Asfaw, A. 2001 Identification of fecal *Escherichia coli* from humans and animals by ribotyping. *Appl. Environ. Microbiol.* **67**, 1503–1507.
- Carson, C. A., Shear, B. L., Ellersieck, M. R. & Schnell, J. D. 2003 Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. *Appl. Environ. Microbiol.* **69**, 1836–1839.

- Dombek, P. E., Johnson, L. K., Zimmerley, S. T. & Sadowsky, M. J. 2000 Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Appl. Environ. Microbiol.* **66**, 2572–2577.
- Gordon, D. M. 2001 Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiol.* **147**, 1079–1085.
- Gordon, D. M., Bauer, S. and Johnson, J. R. 2002 The genetic structure of *Escherichia coli* populations in primary and secondary habitats. *Microbiol.* **148**, 1513–1522.
- Griffith, J. F., Weisberg, S. B. & McGee, C. D. 2003 Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. *J. Wat. Health* **1**(4).
- Hagedorn, C., Robinson, S. L., Filtz, J. R., Grubbs, S. M., Angier, T. A. & Beneau, R. B. 1999 Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Appl. Environ. Microbiol.* **65**, 5522–5531.
- Harwood, V. J., Whitlock, J. & Withington, V. 2000 Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Appl. Environ. Microbiol.* **66**, 3698–3704.
- Lleo, M. M., Bonato, B., Tafi, M. C., Signoretto, C., Boaretti, M. & Canepari, P. 2001 Resuscitation rate in different enterococcal species in the viable but non-culturable state. *J. Appl. Microbiol.* **91**, 1095–1102.
- McLellan, S. L., Daniels, A. D. & Salmore, A. K. 2003 Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. *Appl. Environ. Microbiol.* **69**, 2587–2594.
- SAS Institute 1990 *SAS/STAT user's guide, version 6, Fourth Edition*. SAS Institute, New York, USA.
- Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R. & Lukasik, J. 2002 Microbial source tracking: current methodology and future directions. *Appl. Environ. Microbiol.* **68**, 5796–5803.
- Simpson, J. M., Santo Domingo, J. W. and Reasoner, D. J. 2002 Microbial source tracking: state of the science. *Environ. Sci. Technol.* **36**, 5280–5288.
- Versalovic, J., Koeuth, T., and Lupski, J. R. 1991 Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.* **24**, 6823–6831.
- Versalovic, J., Schneider, M., de Bruijn, F. J. & Lupski, J. R. 1994 Genomic fingerprinting of bacteria using repetitive sequence based polymerase chain reaction. *Meth. Mol. Cell. Biol.* **5**, 25–40.
- Wheeler, A. L., Hartel, P. G., Godfrey, D. G., Hill, J. L. & Segars, W. I. 2002 Potential of *Enterococcus faecalis* as a human fecal indicator for microbial source tracking. *J. Environ. Qual.* **31**, 1286–1293.
- Whitlock, J. E., Jones, D. T. & Harwood, V. J. 2002 Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. *Wat. Res.* **36**, 4273–4282.
- Wiggins, B. A. 1996 Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. *Appl. Environ. Microbiol.* **62**, 3997–4002.
- Wiggins, B. A., Cash, P. W., Creamer, W. S., Dart, S. E., Garcia, P. P., Gerecke, T. M., Han, J., Henry, B. L., Hoover, K. B., Johnson, E. L., Jones, K. C., McCarthy, J. G., McDonough, J. A., Mercer, S. A., Noto, M. J., Park, H., Phillips, M. S., Purner, S. M., Smith, B. M., Stevens, E. N. & Varner, A. K. 2003 Representativeness testing of multi-watershed libraries using antibiotic resistance analysis. *Appl. Environ. Microbiol.* **69**, 3399–3405.