Identification of the sources of *Escherichia coli* in a watershed using carbon-utilization patterns and composite data sets

Samir H. Moussa and Rene D. Massengale

ABSTRACT

The field of bacterial source tracking (BST) has been rapidly evolving to meet the demands of water pollution analysis, specifically the contamination of waterways and drinking water reservoirs by point source and nonpoint source pollution. The goal of the current study was to create a BST library based on carbon-utilization patterns (CUP) for predicting sources of *E. coli* in a watershed, to compare this library to an antibiotic-resistance analysis (ARA) library previously published for the same isolates, and to determine the efficacy of using a composite dataset which combines data from both datasets into a single library for predicting the source of unknown isolates. This was accomplished by generating a CUP dataset and a composite ARA-CUP dataset for the *E. coli* isolates from known fecal sources within a watershed. These libraries were then used to predict the sources of *E. coli* isolates collected from 13 water sites in the same watershed and compared in regard to predictive accuracy. The dominant sources of *E. coli* in the South Bosque watershed were cattle as identified by all three methods. The 6-source composite library had higher average rates of correct classification (96.7%), specificity (99.2%), positive-predictive value (99.1%), and negative-predictive value (96.8%) than either the ARA or CUP 6 source libraries (ARCC 80.1% and 86.7% respectively). The current study is the first field study to compare two phenotypic methods, Antibiotic Resistance Analysis (ARA) and Carbon Utilization Profiling (CUP). This study is also the first to combine both of these methods to create a composite "toolbox" type approach.

Key words | bacterial source tracking, biolog, carbon utilization patterns, *Escherichia coli*

Samir H. Moussa
Department of Biology,
Texas A & M University,
3258 Tamu,
College Station,
Texas 77843-3258,
USA

Rene D. Massengale (corresponding author)
Department of Biology,
Baylor University,
Molecular Biosciences Center,
101 Bagby St. #97388,
Waco TX 76798,
USA
Tel.:254-710-2136
Fax: 254-710-2969
E-mail: *Rene_Massengale@baylor.edu*

INTRODUCTION

Contamination of waterways and drinking water reservoirs by point source and nonpoint source pollution is a major environmental problem which is most often caused by fecal contamination from agricultural sources. Fecal contamination increases nutrient, sodium, and phosphorous levels which can lead to lake eutrophication, algal blooms, and taste and odor problems in drinking water. Elevated levels of bacteria, viruses, and protozoa in the contaminated streams and waterways are often associated with increased rates of illness and disease in the surrounding communities which use the waterways for recreation or as a drinking water source. Levels of fecal contamination in water are quantified by measuring the number of indicator pathogens such as *Escherichia coli* or enterococci in water samples. However, even though these methods are useful in determining which waterbodies exceed safety limits set by the Environmental Protection Agency (EPA) for either recreational or drinking water use, they do not give an indication of the source of the contamination.

Bacterial source tracking is the scientific field that has emerged to address the issue of identifying sources of fecal contamination in water. Bacterial source tracking (BST) is

doi: 10.2166/wh.2008.021

based on the idea that fecal indicator organisms originating from different sources will have unique characteristics which can be used to classify or associate each organism with its original source group. There are a number of BST methods in use today, most of which require the creation of a source library which is a collection of bacterial isolates from fecal samples of known origin. Bacterial isolates from water samples are compared to the library, and the source of the water isolates are then determined based on similar characteristics. A number of BST methods have been developed including both phenotypic and genotypic methods such as antibiotic resistance analysis (ARA) (Wiggins 1996; Parveen *et al.* 1997), carbon-utilization analysis (CUP) (Hagedorn *et al.* 2003), ribotyping (Parveen *et al.* 1999; Carson *et al.* 2001), repetitive DNA PCR (repPCR) (Dombek *et al.* 2000), and pulsed-field gel electrophoresis (Kariuki *et al.* 1999). These methods are not new, but have only recently been applied to the field of BST.

Library-dependent BST involves the collection of indicator bacterial isolates (i.e. *E. coli*, enterococci, or fecal coliforms) and the generation of a database or "library" of specific characteristics about the collected isolates (i.e. antibiotic resistance patterns or carbon utilization patterns (CUP)). A classification rule is created for the library based on statistical analysis of the library characteristics (i.e. discriminant analysis (DA), jackknife analysis (JA), cluster analysis). This classification rule is then used to predict the sources of indicator bacteria from unknown sources into the nearest source category in the library (Wiggins 1996; Wiggins *et al.* 1999). Then the predictive accuracy and representativeness of the library must be tested. BST libraries are often initially compared based on their Average Rates of Correct Classification (ARCCs), which is the average of the rates of correct classification (RCC) for all source categories included in the library. The ARCC is generated when the library isolates are self-crossed as both the calibration data set and test data set. ARCC values are not a complete measure of the predictiveness of a library because the validity of these values is affected by the size of the library and the representative diversity of the isolates included in the library (Harwood *et al.* 2000; Wiggins *et al.* 2003). This means that high ARCC values do not always imply a robust and representative library. Therefore, additional measures of predictive accuracy have been proposed for library-based BST such as analysis of the library's Average Frequency of Misclassification (AFM), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Specificity (Whitlock *et al.* 2002; Harwood *et al.* 2003; USEPA 2005). Finally, pulled-isolate (jackknife) and pulled-sample analysis have been demonstrated as important measures of potential bias in libraries as well (Harwood *et al.* 2000; Wiggins *et al.* 2003).

ARA is perhaps the best-analyzed BST method which uses multiple concentrations of antibiotics to compare isolates. Isolates from different source groups have unique antibiotic-resistance patterns which can be used to group them together or match isolates from unknown sources to the library. It has been used in a number of geographical areas in the United States and has potential for supporting the development of Total Maximum Daily Load (TMDL) values for waterbodies. BST using ARA was first proposed by Wiggins (1996), and a later study by Wiggins *et al.* (2003) indicated that ARA libraries could be combined and successfully used to classify water isolates into source categories. This has produced ARCC values of 37–91% in a number of studies depending on the number of categories used for classification (the ARCC value decreased with increasing number of categories) (Wiggins *et al.* 1999). This method was recently used by Moussa & Massengale to predict the sources of *E. coli* in a Texas watershed and produced ARCC values ranging from 74–85% depending on the statistical method used for analysis (Moussa & Massengale (In Press)).

Another BST approach using the CUP method was recently introduced by Hagedorn *et al.* (2003) with more consistent and higher ARCC values overall than ARA. The CUP method uses the Biolog Microlog II Microbial Identification System and Biolog GN2 microplates which are coated with 95 unique carbon sources coupled to a tetrazolium dye. This system allows the simultaneous identification of an isolate and generation of data for library comparison. Hagedorn *et al.* (2003) demonstrated that bacterial isolates from the same source categories have similar characteristics which allow them to be grouped together based on carbon-source utilization. DA indicated that 30 of the 95 carbon sources contributed to the differences between groups, and these were used to

compare library and water isolates for prediction of the source of the water isolates (Hagedorn *et al.* 2003). The results from his study were promising, with ARCC values at 92.7% for a human vs. nonhuman comparison and ARCC values ranging from 65% to 85% for an analysis of eight sources. Additionally, he found that as the number of animal categories being compared increased, the ARCC decreased (Hagedorn *et al.* 2003).

Several studies have been conducted recently which compare various BST methods (Harwood *et al.* 2003; Myoda *et al.* 2003; Carson *et al.* 2003). These studies have suggested that each BST method has its benefits and limitations and that the creation of libraries based on combined methods could possibly increase the predictive potential and accuracy of library-based BST. However, only one published study to date has actually tested such a "toolbox" approach. Genthner *et al.* (2005) used a composite dataset of ARA data and repPCR fingerprints to source-track enterococci in the shoreline marine waters on Pensacola Beach, Florida. This study suggested that use of the composite ARA-repPCR dataset compared to the ARA dataset alone improved the confidence and predictive accuracy of the library and source predictions. However, the conclusions of this study were limited by the use of a very small library (less than 300) in the analysis (Genthner *et al.* 2005).

The goal of the current study was to create a CUP dataset for a collection of known-source *E. coli* isolates, to compare this dataset to an ARA dataset previously published for the same isolates, and to determine the efficacy of using a composite dataset which combines data from both datasets into a single library for predicting the source of unknown isolates. This was accomplished by generating a CUP dataset and a composite ARA-CUP dataset for the *E. coli* isolates from known fecal sources within a watershed. These libraries were then used to predict the sources of *E. coli* isolates collected from 13 water sites in the same watershed and compared in regard to predictive accuracy. The current study is the first field study to compare two phenotypic methods, Antibiotic Resistance Analysis (ARA) and Carbon Utilization Profiling (CUP). This study is also the first to combine both of these methods to create a composite "toolbox" type approach.

MATERIALS AND METHODS

Fecal sample collection, water sample collection, and isolation of *E. coli*

The bacterial isolates used in the current study were collected and identified as part of a recent ARA study as previously described (Moussa & Massengale (In Press)). These isolates had been stored at −80°C in 10% glycerol between studies; and for the current study, the individual cultures were thawed, streaked onto tryptic soy agar (TSA) plates, and incubated at 37°C for 24 hours. The set of isolates included *E. coli* isolated from fecal samples from beef cattle, horse, goat, sheep, and dog; *E. coli* from sewage samples (designated as "human" isolates); and *E. coli* from water samples collected in the South Bosque watershed in central Texas. In total, 160 beef cattle isolates, 62 dog isolates, 32 goat isolates, 32 horse isolates, 230 human isolates, and 80 sheep isolates were analyzed to construct the known library of isolates for the previous ARA analysis; and these same isolates were analyzed by CUP in the current study. Additionally, 276 *E. coli* water isolates collected from the South Bosque watershed were analyzed by CUP in the current study as described for the previous ARA study (Moussa & Massengale (In Press)).

Carbon utilization analysis (CUP) of library and water isolates

CUP analysis was performed for each library and water *E. coli* isolate by transferring each isolate to Biolog Universal Growth (BUG) agar containing defibrinated sheep's blood and incubating at 35°C for 24 hours according to the Biolog protocol (Holmes *et al.* 1994; Biolog, Hayward, CA). Each cultivated isolate was then transferred to Biolog GN/GP inoculation fluid at 61% ± 2% turbidity and used to inoculate a Biolog GN2 microplate (150 μl/well). These plates were incubated at 35°C for 16–24 hours and the resulting color change and turbidity was measured using a plate reader. Each well of the GN2 plate is coated with a carbon source coupled to a tetrazolium dye. When the carbon source is utilized by the bacteria, it oxidizes the dye and causes a purple color change in the well.

Statistical analysis

Three library datasets were used in the current study: 1) ARA dataset from Moussa and Massengale (In Press), 2) CUP dataset for the same *E. coli* isolates generated in the current study, and 3) a composite dataset that combined the ARA data from the previous study and the CUP data from the current study. Data were analyzed using both SAS JMP software (SAS Institute, Cary, NC) and BioNumerics software (Applied Maths, Austin, TX). Discriminant analyses in SAS JMP were performed to determine how well each animal-source isolate was classified into its original category for CUP, and the Toolbox Approach. DA is a multivariate method of predicting some level of a one-way classification based on known values of the responses. The technique is based on how close a set of measurement variables are to the multivariate means of the levels being predicted.

These discriminant analyses were performed in several combinations including an analysis comparing all animal sources, also called a 6-way classification (i.e. human vs. beef vs. dog vs. goat vs. horse vs. sheep); a 3-way classification (i.e. human vs. beef vs. other animals); and a 2-way classification (i.e. human vs. nonhuman). These combinations are important because they address different source-tracking objectives. For example, if the exact animal source which is contributing bacteria to a waterbody needs to be identified then a 6-way classification (or even more categories) would be appropriate. A 2-way classification would be appropriate when the exact animal does not need to be identified; rather the investigator only needs to know whether bacteria were contributed by humans or nonhumans. The DA results including the RCC and ARCC values were compared for the three libraries (ARA, CUP, and ARA + CUP). The group distances generated by DA were then compared by ANOVA (analysis of variance) to determine the significance of the distances between groups. When group variances were not equal, the Welch ANOVA was used for comparison (Welch 1951; Brown & Forsythe 1974).

Stepwise discriminant analyses in SAS JMP were performed to identify the carbon sources which contributed to the distance between groups. Then discriminant analyses were run on this subset of carbon-source data to construct and the resulting classification frequency tables were used to generate Rates of Correct Classification (RCC) and Average Rate of Correct Classification (ARCC) values (Hagedorn *et al.* 2003). The multivariate analysis of variance (MANOVA) function was utilized in order to create a graphical representation of where each isolate was graphed in Cartesian space based on the largest elements of discrimination in the comparison (Dombek *et al.* 2000). The Average Frequency of Misclassification (AFM) was calculated for each library by first calculating the False Positive (FP) frequency for each category within the library which is the number of isolates incorrectly classified into a source divided by the number of isolates in the library which are not from that source (Harwood *et al.* 2003). Then this value is averaged and multiplied by 100 to generate the AFM which is expressed as a percent.

Cluster analysis in BioNumerics was used to create a dendrogram characterizing the similarity of the isolates to each other. A Jaccard similarity coefficient was used to analyze the ARA data due to their binomial nature while a Pearson correlation was used to analyze the continuous CUP and Toolbox Approach data (Ritter *et al.* 2003). JA was subsequently used in order to determine how well the similarity values were able to predict the bacterial isolate source group. This was done by first manually assigning the bacterial isolates to their correct source group. This method takes out one bacterial isolate fingerprint and compares it to the other fingerprints in order to identify it as originating from a particular animal source. Maximum similarities were used with ties split evenly among the groups in the comparison (Ritter *et al.* 2003).

Bootstrap analysis in BioNumerics was also used in order to identify any isolates that may be clones of each other (Ritter *et al.* 2003). This analysis removed thirty isolates at random and recalculated similarity values. One thousand iterations were selected, meaning that 1000 groups of 30 isolates would be removed at random and a similarity dendrogram generated for each. Contingency table analysis in SAS JMP was used to compare the RCC of the three methods to each other (ARA, CUP, Toolbox). Linear regression was finally used to correlate ARCC values from DA and JA.

The predicted accuracy of each of the methods (DA and JA) was determined by calculating the percentage of true-positives (TP), the Positive Predictive Value (PPV), the Percentage False-Positive (FP), specificity, and Negative

Predictive Value (NPV) according to the method of Harwood *et al.* (2003) and the recommended EPA guidelines (USEPA 2005). Percent TP was calculated by dividing the number of isolates for which the source was correctly classified (CC) by the total number of isolates in that source category (AC) (TP = CC/AC × 100). FP was calculated as described above. Specificity was calculated as the true negatives (TN) divided by FP + TN (Specificity = TN/(FP + TN) × 100). The PPV was calculated as TP/(TP + FP) × 100. Finally, the NPV was calculated as the TN/(false negative + TN) × 100).

RESULTS

E. coli isolates collected from fecal and water samples from the South Bosque watershed for a previous ARA BST study were used to create the known-source library for the current project (Table 1, Moussa & Massengale (In Press)). Each isolate was then grown on Biolog University Growth (BUG) agar and inoculated into a Biolog GN2 plate which was incubated and read with the Biolog plate reader according to the manufacturer's protocol. The data were analyzed using a stepwise DA, and carbon sources which did not contribute to the differences among the animal groups were removed from comparison. This analysis revealed that there were 77 carbon wells which contributed to the differences among the animal groups (Table 2). These 77 wells were the only ones used for all analyses involving the CUP method and Toolbox method.

Table 1 | Description of datasets used in the current study

Dataset	Number of library isolates	Number of water isolates	Reference
ARA dataset	596	276	Moussa & Massengale (In Press)
CUP dataset[a]	596	276	Current study
Composite dataset[a]	596	276	Current study

[a]All 3 datasets were generated using the same library of *E. coli* isolates originally isolated in the previous study by Moussa & Massengale (In Press).

Table 2 | Top 77 carbon sources significant for discrimination among sources

Carbon sources	
Cyclodextrin	Itaconic Acid
Dextrin	Ketobutyric Acid
Glycogen	Ketoglutaric Acid
Tween 40	Ketovaleric Acid
Tween 80	Lactic Acid
Acetyl-D-Galactosamine	Propionic Acid
Adonitol	Quinic Acid
Arabinose	Saccharic Acid
Arabitol	Sebacic Acid
Cellobiose	Succinamic Acid
Fructose	Glucuronamide
Fucose	Alaninamide
Galactose	D-Alanine
Gentiobiose	L-Alanine
Glucose	Alanyl-Glycine
Inositol	Asparagine
Lactose	Aspartic Acid
Lactulose	Glutamic Acid
Maltose	Glycyl-L-Aspartic Acid
Mannitol	Glycyl-L-Glutamic Acid
Mannose	Proline
Melibiose	D-Serine
Methyl-D-Glucoside	L-Serine
Psicose	Threonine
Rhamnose	Carnitine
Sucrose	Aminobutyric Acid
Trehalose	Urocanic Acid
Xylitol	Inosine
Pyruvic Acid Methyl Ester	Uridine
Acetic Acid	Phenylethylamine
Citric Acid	Putrescine
Formic Acid	Butanediol
Galactonic Acid Lactone	Glycerol
Galacturonic Acid	Glycerol Phosphate
Gluconic Acid	Glucose-1-Phosphate
Glucosaminic Acid	Glucose-6-Phosphate
Glucuronic Acid	
Hydroxybutyric Acid	
Hydroxybutyric Acid	
Hydroxybutyric Acid	
Hydroxyphenyl-acetic Acid	

DA of CUP data was used to construct RCC tables and determine ARCC values. When isolates from the six sources were analyzed, based on their utilization of the top 77 carbon wells, the ARCC was 86.7% (Table 3). The best-classified category was goat with an RCC of 93.8% and an FP rate of 0.4%. The worst classified category was dog with an RCC of 72.6% and a FP rate of 3.2%. Pairwise analysis of the group distances indicated that there were significant distances between library groups with all comparisons exceeding the critical F value at $p < 0.0001$. The AFM for the entire 6-way CUP library was 2.7% ± 2.2% S.D. The ARCC for a three-way classification (beef cattle vs. human vs. other) was 86.6% (Table 3) with RCCs ranging from 85.7% (human) to 87.9% (other animals). A 2-way comparison (human vs. nonhuman) resulted in an ARCC of 89.7%.

DA is a resubstitution analysis statistical method which compares each isolate to the library without pulling that isolate out of the library, thus the isolate is being compared against itself. Therefore, there is a potential for the similarity probability to be overestimated by DA. Therefore, JA was also used to assess the library because this method pulls each isolate from the library before comparison (it is a pulled-isolated comparison) (Table 3). The ARCC values of the 2-way classification were higher than either the 3-way or 6-way classifications (88.4% vs. 84.9% and 78.6% respectively). ARCC values for the CUP library generated by DA were higher than JA in every comparison; however, no significant differences were found when RCC and ARCC values based on DA or JA were compared by contingency table analysis. DA and JA results for the CUP correlated well as indicated by regression analysis of RCC values with an R^2 value of 0.52, ($F(1,9) = 1049$ $p < 0.0001$). This indicates that the DA classification was not significantly overestimated. When the two methods (DA vs. JA) were compared in terms of Positive Predictive Value (PPV), Negative Predictive Value (NPV), and specificity, DA produced better results than JA in each case. However, these higher results were not statistically significant within this library.

Table 3 | Rates of correct classification by source for CUP library using DA and JA of 596 *E. coli* isolates

		% Correctly classified	
Source	**No. isolates**	**DA**	**JA**
Two-way classification			
Human	230	84.8	83.9
Non-human	366	94.5	92.9
Totals	596	89.7 (ARCC)	88.4 (ARCC)
Three-way classification			
Human	230	85.7	83.9
Beef cattle	160	86.3	86.9
Other animals	206	87.9	84
Totals	596	86.6 (ARCC)	84.9 (ARCC)
Six-way classification			
Human	230	89.6	83.9
Beef cattle	160	90.6	86.9
Dog	62	72.6	67.7
Sheep	80	86.3	80
Horse	32	87.5	71.9
Goat	32	93.8	81.3
Totals	596	86.7 (ARCC)	78.6 (ARCC)

Composite "toolbox" dataset analysis

A composite dataset was then created by combining the ARA dataset from a previous study by Moussa and Massengale (In Press) and the CUP dataset from the current study in a "toolbox" approach. Both datasets were derived from the same *E. coli* isolate library. ARA and CUP data for the same *E. coli* isolates were combined into a single dataset and then subjected to DA and JA to determine the RCCs and ARCCs of the composite library. This created a set of 109 data points (32 ARA results plus 77 CUP data results) to be compared for each isolate. When isolates from the six categories were analyzed by DA based on their resistance of antibiotics and utilization of carbon sources, the ARCC was 96.7% (Table 4). The best-classified category was goat with an RCC of 100% and an FP rate of 0%. The least well-classified category was beef cattle with an RCC of 95% and an FP rate of 1.8%. When isolates from the six categories were analyzed by JA based on their resistance of antibiotics and utilization of carbon sources, the ARCC was 87.1% (Table 4). The best-classified category was sheep with an

Table 4 | Rates of correct classification by source for Composite Library (ARA + CUP) using DA and JA of 596 *E. coli* isolates

Source	No. isolates	% Correctly classified DA	JA
Two-way classification			
Human	230	95.2	83.9
Non-human	366	98.6	96.7
Totals	596	96.9 (ARCC)	90.3 (ARCC)
Three-way classification			
Human	230	96.5	83.9
Beef cattle	160	92.5	88.8
Other animals	206	95.1	93.2
Totals	596	94.7 (ARCC)	88.6 (ARCC)
Six-way classification			
Human	230	97	83.9
Beef cattle	160	95	88.8
Dog	62	95.2	79
Sheep	80	96.3	98.8
Horse	32	96.9	90.6
Goat	32	100	81.3
Totals	596	96.7 (ARCC)	87.1 (ARCC)

RCC of 98.8%. The least well-classified category was dog with an RCC of 79%.

Each library was then used to predict the sources of *E. coli* isolates derived from water samples in the South Bosque watershed. Both DA and JA of the 6-way CUP library indicated that beef cattle were the dominant source of *E. coli* in the South Bosque water samples (Table 5). A major watershed management question which is frequently asked in agricultural areas is whether fecal contamination is coming from human, cattle, or other sources. To address this question, this 6-way library can be used to predict contamination sources, or library groups can be collapsed into larger categories for comparison for greater statistical power of the prediction if needed; therefore, a three-way classification was used for analysis: beef cattle, human, and other animals (dog, goat, horse, sheep). The three-way comparison of the CUP library (with JA) confirmed that cattle were the dominant source of *E. coli* in the watershed

Table 5 | Predicting sources of *E. coli* isolates from South Bosque water samples based on the CUP 6-category Library generated from DA* and JA**

Source	Discriminant analysis	Jackknife analysis
Human	15.5%	25.3%
Beef cattle	50.2%	45.7%
Dog	18.8%	11.2%
Sheep	8.6%	16.0%
Horse	4.5%	0.7%
Goat	2.4%	1.1%

*100% of unknown water isolates classified to the 6 categories
**100% of unknown water isolates classified to the 6 categories

with cattle and human accounting for 45.7% and 25.3% of water isolates respectively. When the composite library was used to classify water isolates into source categories, beef cattle were again identified as the dominant source of *E. coli* in the water samples with 34.1–51.2% of water isolates identified as beef cattle depending on statistical method (Table 6). Horse isolates were the least prevalent in the South Bosque watershed with 3.7% of the isolates identifying as horse.

The classification of *E. coli* water isolates using each library (ARA, CUP, and Toolbox) was compared (Figure 1). Moussa and Massengale (In Press) had previously reported that classification of these same 276 water *E. coli* isolates with the ARA library resulted in 8.3% human, 40.6% beef cattle, 10.1% dog, 10.5% sheep, 8.7% horse, and 21.7% goat. When those ARA classification results were compared to the current Toolbox library classification by regression analysis, the prediction of water isolate sources correlated well (Figure 2). The classification of water isolates in the CUP library correlated even better with the Toolbox library which is to be expected since two out of every three data points considered for discriminant analysis were from the CUP dataset (Figure 3).

The ARCC and average frequency of misclassification (AFM) for each of the three methods was calculated (Figure 4). The AFM was calculated by dividing the number of isolates incorrectly classified into a certain category, such

Table 6 | Predicting sources of *E. coli* isolates from South Bosque water samples based on the Composite 6-category Library generated from DA* and JA**

Source	Discriminant analysis	Jackknife analysis
Human	13.4%	18.9%
Beef cattle	51.2%	34.1%
Dog	15.9%	14.2%
Sheep	7.7%	19.6%
Horse	3.7%	3.7%
Goat	8.1%	8.1%

*100% of unknown water isolates classified to the 6 categories
**98.6% of unknown water isolates classified to the 6 categories

as beef, by the total number of isolates, excluding beef. The AFM for the Toolbox library was lower than that of the ARA or CUP libraries ($F = 3.5$, $p = 0.08$). Additionally, the RCCs from the different animal groups were compared and found to be significantly different from each other ($X^2 = 7845$, $p < 0.00001$). When the RCCs were compared between libraries, the Toolbox library RCCs were significantly higher than either the ARA or CUP libraries ($F = 18.2$, $p < 0.002$). Bootstrap analysis in Bionumerics showed that less than 1% of the isolates in all three analyses were found to be clonal.

The predicted accuracy of each of the methods was determined by calculating the percentage of true-positives

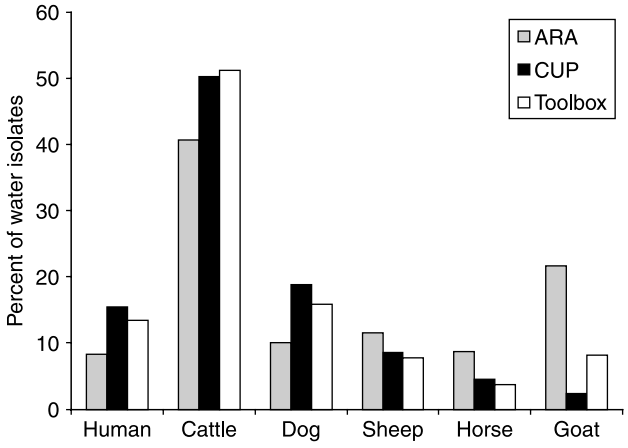

Figure 1 | Comparison of sources of water *E. coli* isolates by method. Each bar represents the percent of water isolates classified into that category.

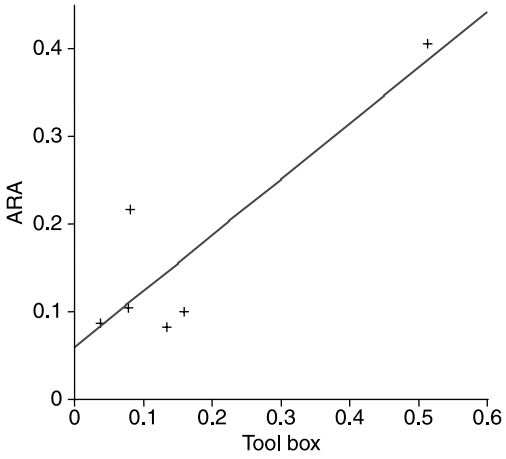

Figure 2 | Correlation of water isolate source prediction by ARA and toolbox libraries. Each cross represents the ratio of water isolates classified into one of the six library categories. *E. coli* classification with the ARA library correlated well with the toolbox library classification results ($R^2 = 0.77$; $F = 13$; $p = 0.022$).

(TP), the Positive Predictive Value (PPV), the Percentage False-Positive (FP), specificity, and Negative Predictive Value (NPV) (Table 7). TP is the percentage of isolates which were identified as originating from a certain animal category when they truly were from that group. This corresponds to the ARCC for each library and is functionally the sensitivity of the library to predicting isolate sources. The PPV is the percentage that isolates that were true-positive are correctly classified as such. The FP is the

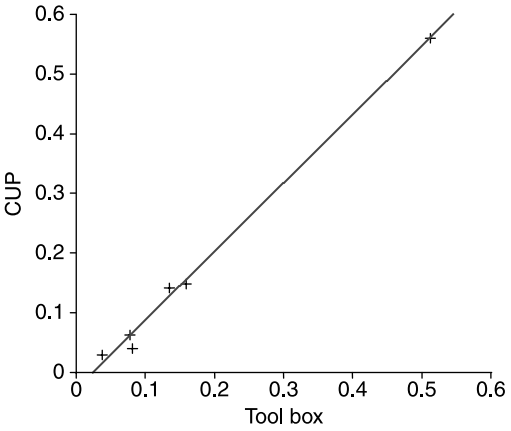

Figure 3 | Correlation of water isolate source prediction by CUP and toolbox libraries. Each cross represents the ratio of water isolates classified into one of the six library categories. *E. coli* classification with the CUP library correlated well with the toolbox library classification results ($R^2 = 0.99$; $F = 719$; $p = 0.0001$).

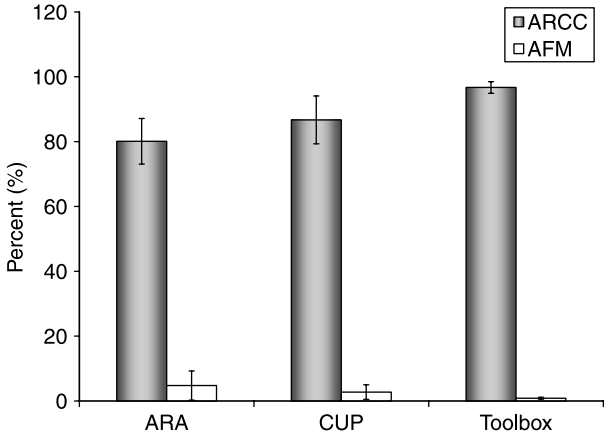

Figure 4 | Average rate of correct classification (ARCC) and average frequency of misclassification (AFM) for ARA, CUP and toolbox approach.

percentage of isolates which were identified as originating from a certain animal category when those isolates actually originated from another. Specificity is the percentage of isolates which do not classify into a certain animal category when they truly are not from that animal category. Finally, the NPV is the percentage that an isolate identified as not originating from a particular group is true. In general, PPV, NPV, and specificity values above 85–90% are considered acceptable (USEPA 2005). The closer these values are to 100%, the more robust the library theoretically is in terms of its utility in successfully and accurately classifying isolates into source groups. Overall, all of the libraries classified isolates well with sensitivity rates of 80% or greater and specificities of 95% or greater as indicated by jackknife cross-validation assays and pulled-sample analysis. The composite "toolbox" library had the highest results in each category except the FP measure in which it had the lowest result of all three assays.

Table 7 | Average predictive accuracy of ARA, CUP, and Toolbox approach for 6-way Libraries

Method	Percentage TP	PPV	Percentage FP	Specificity	NPV
ARA	80.1%	94.5%	4.8%	95.2%	82.9%
CUP	86.7%	97.0%	2.7%	97.3%	88.3%
Toolbox	96.7%	99.1%	0.8%	99.2%	96.8%

DISCUSSION

E. coli isolates from the six animal categories were analyzed by CUP using the top 77 carbon sources which contributed to the discrimination between CUP patterns from different source groups. DA of the CUP data produced an ARCC of 86.7% for the six-way classification and an ARCC of 89.7% for the 2-way classification. Finally, the CUP data were combined with ARA data from a previous study, to determine whether a composite statistical analysis (called the Toolbox Approach in the current study) would result in higher ARCC values as compared to ARA and CUP individually. DA of the data produced an ARCC of 96.7% when the six animal sources were analyzed by this combined method and an ARCC of 96.9% for the 2-way analysis (human vs. nonhuman). The rates of correct classification obtained from the composite dataset in the Toolbox approach are significantly better than those published in most recent studies. This study indicated that combining datasets into a composite set for analysis provided better discrimination between source groups in the library.

Few BST studies utilizing CUP have been published to date. The current results are directly in line with another study by Harwood *et al.* (2003) which generated an *E. coli* CUP library and enterococci CUP library with ARCC values of 86.6% and 84.8% respectively. Another study was conducted by Hagedorn *et al.* in which the library compared eight animal categories (Hagedorn *et al.* 2003). The ARCC for the 6-way classification in the current study was higher than the ARCC of 75.9% from the 8-way classification in the Hagedorn study (Hagedorn *et al.* 2003). In contrast, the ARCC of the 2-way comparison in the current CUP analysis was slightly lower than the two-way ARCC of 92.7% from Hagedorn's study. These differences may be due to a variety of factors including 1) differences in the carbon sources included for analysis, 2) differences in the number of carbon sources utilized, and 3) difference in target organisms (enterococci in the Hagedorn study vs. *E. coli* in the current study).

The ARA library generated previously and the CUP and composite Toolbox Library from the current study can all be compared in regard to library accuracy and representativeness. The ARCC values of the 6-way

classifications and 2-way classifications from all three libraries are much higher than the random chance of assignment of 16.7% and 50%, respectively. When the probability of random chance is calculated as proportional to the number of isolates, the random chance of assignment increases for those groups which contain a large amount of isolates (human, beef cattle) and decreased for those groups which contain a smaller amount of isolates (dog, goat, horse, and sheep). Even with these proportional probabilities, the ARCC values for the 6-way classification and 2-way classification remained large. This shows that the sample size of 596 isolates from 6 animal categories was large enough to minimize random chance classification. Finally, although the ARCC were similar between the three libraries, the Contingency Table Analysis revealed that the individual Rates of Correct Classification (RCC) of the animal groups in all three libraries and comparisons were significantly different. This indicates that some individual groups were better classified in one library than the others.

In order to further compare the three methods, several factors such as the percent false positives (FP) were calculated. This factor reveals the approximate percent of isolates from each of the three libraries (ARA, CUP, and Toolbox Approach) which incorrectly identified an isolate as originating from an animal source. Average FP values for the 6-category libraries were under 5%: ARA FP rate was 4.8%, CUP FP rate was 2.7%, and Toolbox Approach FP rate was 0.8%. This further supports the conclusion that the Toolbox library was superior to the ARA and CUP libraries individually.

JA, in addition to DA, was also used to determine how well the library classified isolates from different animal sources. ARCC values produced by the Jackknife analyses (maximum similarities, spread ties equally) were consistently less than those produced by DA. DA may have outperformed the JA on constructing the known library since source patterns, as found in an MDS plot, had central tendencies, revealing that a statistical method based on averages such as DA was more appropriate for phenotypic methods, as previously suggested by Ritter *et al.* (2003). However, JA is useful because it allows library cross-validation while removing each isolate from the library before comparison.

E. coli isolates from water were also analyzed using the ARA, CUP, and Toolbox Approach to determine the effectiveness of these libraries in estimating *E. coli* sources. All three libraries predicted that the major source of contamination in the South Bosque water samples was from cattle. Even though each library produced consistent classifications, the Toolbox library performed significantly better than the other two libraries in terms of sensitivity (capacity to detect true positives), specificity, PPV, and NPV. This indicates that the Toolbox library was more predictive of the sources of water *E. coli* isolates than either the ARA or CUP libraries alone. All three library-based methods produced excellent classification rates and were used to classify 95–100% of water isolates into source categories. This is significantly higher than the unknown classification rates in other studies which to date have only been able to use libraries to classify 10–50% of unknown isolates in general.

Future BST studies should explore the effectiveness of increasing the number of *E. coli* isolates collected from water in improving the representativeness of the isolates used to predict contamination sources. While 276 *E. coli* isolates from water were analyzed by the three methods in the current study, more study isolates would serve to assure that all possible major types of *E. coli* strain present in the watershed are represented in the water sample isolates. Theoretically, the library size requirement may actually be smaller for composite libraries if they do indeed improve PPV, NPV, and AFM rates. Future studies should focus on the effectiveness of the Toolbox Approach in assessing larger libraries and watersheds in addition to determining the minimum library size required for such a composite library.

REFERENCES

Brown, M. B. & Forsythe, A. B. 1974 The small sample behavior of some statistics which test the equality of several means. *Technometrics* **16**, 129–132.

Carson, C., Shear, B., Ellersieck, M. & Asfaw, A. 2001 Identification of fecal *Escherichia coli* from humans and animals by ribotyping. *Appl. Env. Microbiol.* **67**, 1503–1507.

Carson, C., Shear, B., Ellersieck, M. & Schneu, J. 2003 Comparison of ribotyping and repetetive extragenic palindromic-PCR for

identification of fecal *Escherichia coli* from humans and animals. *Appl. Environ. Microbiol.* **69**, 1836–1839.

Dombek, P., Johnson, L., Zimmerley, S. & Sadowsky, M. 2000 Use of repetitive DNA sequences and the PCR to differentiate *E. coli* isolates from human and animal sources. *Appl. Environ. Microbiol.* **66**, 2572–2577.

Genthner, F., James, J., Yates, D. & Friedman, S. 2005 Use of composite data sets for source-tracking enterococci in the water column and shoreline interstitial waters on Pensacola Beach, Florida. *Marine Poll. Bull.* **50**, 724–732.

Hagedorn, C., Crozier, J., Mentz, K., Booth, A., Graves, A., Nelson, N. & Reneau, R. Jr. 2003 Carbon source utilization profiles as a method to identify sources of faecal pollution in water. *J. Appl. Microbiol.* **94**, 792–799.

Harwood, V., Whitlock, J. & Withington, V. 2000 Classification of antibiotic resistance patterns of indicator bacteria by DA: use in predicting the source of fecal contamination in subtropical waters. *Appl. Environ. Microbiol.* **66**, 3698–3704.

Harwood, V., Wiggins, B., Hagedorn, C., Ellender, R., Gooch, J., Kern, J., Samadpour, M., Chapman, A., Robinson, B. & Thompson, B. 2003 Phenotypic library-based microbial source tracking methods: efficacy in the California collaborative study. *J. Wat. Health* **1**, 153–166.

Holmes, B., Costa, M., Ganner, S. & Stevens, O. 1994 Evaluation of the biolog system for the identification of some gram-negative bacteria of clinical importance. *J. Clin. Microbiol.* **32**, 1970–1975.

Kariuki, S., Gilks, C., Kimari, J., Obanda, A., Muyodi, J., Waiyaki, P. & Hart, C. 1999 Genotype analysis of *Escherichia coli* strains isolated from children and chickens living in close contact. *Appl. Environ. Microbiol.* **65**, 472–476.

Moussa, S. & Massengale, R. (In press) Identification of the sources of *Escherichia coli* in a watershed using antibiotic resistance analysis: a comparison of discriminant and jackknife analyses. *J. Environmental Detection*.

Myoda, S., Carson, C., Fuhrmann, J., Hahm, B., Hartel, P., Kuntz, R., Nakatsu, C., Sadowsky, M., Samadpour, M. & Yampara-Iquise, H. 2003 Comparing genotypic bacterial source tracking methods that require a host origin database. *J. Wat. Health* **1**, 167–180.

Parveen, S., Murphree, R., Edmiston, L., Kaspar, C., Portier, K. & Tamplin, M. 1997 Association of multiple-antibiotic-resistance profiles with point and nonpoint sources of *Escherichia coli* in Appalachicola Bay. *Appl. Environ. Microbiol.* **63**, 2607–2612.

Parveen, S., Portier, K., Robinson, K., Edmiston, L. & Tamplin, M. 1999 Discriminant analysis of ribotype profiles of Escherichia coli for differentiating human and nonhuman sources of fecal pollution. *Appl. Env. Microbiol.* **65**, 3142–3147.

Ritter, K., Carruthers, E., Carson, C., Ellender, R., Harwood, V., Kingsley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B., Whitlock, J., Wiggins, B. & Wilbur, J. 2003 Assessment of statistical methods used in library-based approaches to microbial source tracking. *J. Wat. Health* **1**, 209–223.

United States Environmental Protection Agency (US EPA) 2005 Microbial Source Tracking Guide Document. Document # EPA/600-R-05-064.

Welch, B. L. 1951 On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330–336.

Whitlock, J., Jones, D. & Harwood, V. 2002 Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. *Water Res.* **36**, 4273–4282.

Wiggins, B. 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., Andrews, R., Conway, R., Corr, C., Dobratz, E., Dougherty, D., Eppard, J., Knupp, S., Limjoco, M., Mettenburg, J., Rinehardt, J., Sonsino, J., Torrijos, R. & Zimmerman, M. 1999 Use of antibiotic resistance analysis to identify nonpoint sources of fecal pollution. *Appl. Environ. Microbiol.* **65**, 3483–3486.

Wiggins, B., Cash, P., Creamer, W., Dart, S., Garcia, P., Gerecke, T., Han, J., Henry, B., Hoover, K., Johnson, E., Jones, K., McCarthy, J., McDonough, J., Mercer, S., Noto, M., Park, H., Phillips, M., Purner, S., Smith, B., Stevens, E. & Varner, E. 2003 Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Appl. Environ. Microbiol.* **69**, 3399–3405.

First received 15 February 2006; accepted in revised form 27 February 2007. Available online January 2008