

1 Casarez, E. A. (*postdoctoral student*), S. D. Pillai, J. Mott, M. Vargas, K. Dean and  
2 G. D. Di Giovanni. 2006. Direct comparison of four bacterial source tracking methods and a  
3 novel use of composite data sets. *Journal of Applied Microbiology*, *In press*.

4

5 **Direct Comparison of Four Bacterial Source Tracking Methods**  
6 **and Use of Composite Data Sets**

7

8 **Elizabeth A. Casarez,<sup>1</sup> Suresh D. Pillai,<sup>2</sup> Joanna B. Mott,<sup>3</sup> Mel Vargas,<sup>4</sup> Kirk E. Dean,<sup>4</sup> and**  
9 **George D. Di Giovanni<sup>1</sup>**

10

11 <sup>1</sup>*Texas Agricultural Experiment Station, Texas A&M University Agricultural Research Center,*  
12 *El Paso, TX, USA,* <sup>2</sup> *Department of Poultry Science, Texas A&M University, College Station,*  
13 *TX, USA,* <sup>3</sup> *Department of Physical and Life Sciences, Texas A&M University-Corpus Christi,*  
14 *TX, USA,* <sup>4</sup> *Parsons Water and Infrastructure, Inc., Austin, TX, USA*

15

16 *Correspondence to: George D. Di Giovanni, Texas Agricultural Experiment Station, Texas A&M*  
17 *University Agricultural Research Center, 1380 A&M Circle, El Paso, Texas 79927, USA, Phone:*  
18 *915-859-9111, Fax: 915-859-1078, E-mail: [gdigiovanni@ag.tamu.edu](mailto:gdigiovanni@ag.tamu.edu)).*

19

20 Running title: BST composite data sets

1 **ABSTRACT**

2

3 **Aims:** Four bacterial source tracking (BST) methods, enterobacterial repetitive intergenic  
4 consensus polymerase chain reaction (ERIC-PCR), automated ribotyping using *HindIII*, Kirby-  
5 Bauer antibiotic resistance analysis (KB-ARA), and pulsed field gel electrophoresis (PFGE)  
6 were directly compared using the same collection of *Escherichia coli* isolates. The data sets from  
7 each BST method and from composite methods were compared for library accuracy and their  
8 ability to identify water isolates.

9 **Methods and Results:** Potential sources of faecal pollution were identified by watershed  
10 sanitary surveys. Domestic sewage and faecal samples from pets, cattle, other avian livestock,  
11 other non-avian livestock, avian wildlife, and non-avian wildlife sources were collected for  
12 isolation of *Escherichia coli* (*E. coli*). A total of 2275 *E. coli* isolates from 813 source samples  
13 were screened using ERIC-PCR to exclude clones and to maximize library diversity, resulting in  
14 883 isolates from 745 samples selected for the library. The selected isolates were further  
15 analyzed using automated ribotyping with *HindIII*, Kirby-Bauer antibiotic resistance analysis  
16 (KB-ARA), and pulsed field gel electrophoresis (PFGE). A total of 555 *E. coli* isolates obtained  
17 from 412 water samples were also analyzed by the four BST methods. A composite data set of  
18 the four BST methods gave the highest rates of correct classification (RCCs) with the fewest  
19 unidentified isolates than any single method alone. RCCs for the four method composite data set  
20 and a seven-way split of source classes ranged from 22% for avian livestock to 83% for domestic  
21 sewage. Two-method composite data sets were also found to be better than individual methods,  
22 having RCCs similar to the four-method composite and identification of the same major sources  
23 of faecal pollution.

1 **Conclusions:** The use of BST composite data sets may be more beneficial than the use of single  
2 methods.

3 **Significance and Impact of the Study:** This is one of the first comprehensive comparisons  
4 using composite data from several BST methods. While the four-method approach provided the  
5 most desirable BST results, the use of two-method composite data sets may yield comparable  
6 BST results while providing for cost, labour, and time savings.

7  
8 **Keywords:** *Escherichia coli*, antibiotic resistance, ERIC-PCR, PFGE, ribotyping, water quality  
9 faecal pollution

1 **INTRODUCTION**

2

3           Concern for water quality and the potential for animal agricultural operations to  
4 contaminate watersheds have led to the need for the identification of sources of faecal  
5 contamination. Bacterial source tracking (BST) has the potential to be used as a tool in  
6 effectively managing water resources and setting total maximum daily loads (TMDLs) as  
7 needed. The premise behind BST is that genetic and biochemical tests can identify bacterial  
8 strains that are host specific so that the original host animal and source of the faecal  
9 contamination can be identified. Often *Escherichia coli* (*E. coli*) or *Enterococcus* spp. are used  
10 as the bacteria targets in source tracking (for example, Parveen *et al.*, 1999, Dombek *et al.*, 2000,  
11 Graves *et al.*, 2002, Griffith *et al.*, 2003, Hartel *et al.*, 2003, Kuntz *et al.*, 2003, Stoeckel *et al.*,  
12 2004, Scott *et al.*, 2005). While there has been some controversy as to the host specificity and  
13 survival of *E. coli* in the environment (Gordon *et al.*, 2002), this indicator organism has the  
14 advantage that it is known to correlate with the presence of faecal contamination and is used for  
15 human health risk assessments. Tracking *E. coli*, therefore, has the advantages of direct  
16 regulatory significance and availability of standardized culturing techniques for water samples,  
17 such as United States Environmental Protection Agency (USEPA) Method 1603 (USEPA,  
18 2005a), as opposed to relying on methods developed for clinical use.

19           There have been many different technical approaches to bacterial source tracking  
20 (reviewed by Scott *et al.*, 2002, Simpson *et al.*, 2002, Meays *et al.*, 2004), but there is currently  
21 no consensus on a single method for field application. Genotypic tools appear to hold the  
22 promise for BST, providing the most conclusive characterization and level of discrimination for  
23 isolates. Of the molecular tools available, ribosomal ribonucleic acid (RNA) genetic

1 fingerprinting (ribotyping), repetitive element polymerase chain reaction (rep-PCR) and pulsed-  
2 field gel electrophoresis (PFGE) are emerging as a few of the versatile and feasible BST  
3 techniques. Antibiotic resistance profiling, a phenotypic characterization method, also has the  
4 potential to identify the human or animal origin of isolates, and variations of this technique have  
5 been applied in several BST studies.

6         Each of these methods has its strengths and weaknesses. Ribotyping has a moderate  
7 ability to resolve different strains of the same bacterial species. An automated ribotyping system  
8 is available, which saves labour costs and requires little training, but the initial investment and  
9 the consumable cost per isolate are expensive. PFGE has very high resolution and can  
10 discriminate between closely related bacterial strains. While this allows higher confidence in the  
11 matches made, typically fewer environmental isolates are identified compared to other BST  
12 techniques. Enterobacterial repetitive intergenic consensus sequence polymerase chain reaction  
13 (ERIC-PCR), a type of rep-PCR, has moderately high ability to resolve different closely related  
14 bacterial strains. Consumable costs for ERIC-PCR are inexpensive and labour costs for sample  
15 processing and data analyses are moderate. Of these four methods, antibiotic resistance analysis  
16 using the Kirby-Bauer method has the lowest ability to discriminate closely related bacterial  
17 strains. It also has the lowest initial and per sample cost and takes the least time and training, but  
18 the statistical analysis of data can be complex and time-consuming. A disadvantage of all of  
19 these techniques is that reference libraries of genetic and phenotypic fingerprints for *E. coli*  
20 isolated from known sources (*e.g.*, domestic sewage, livestock, and wildlife) are needed to  
21 identify the sources of bacteria isolated from environmental water samples. Thus, the  
22 development of an identification library can be a time consuming and expensive component of a  
23 BST study.

1           While there have been comparisons of different BST methods (for example Griffith *et al.*,  
2 2003, Myoda *et al.*, 2003, Stoeckel *et al.*, 2004), few have done a side-by-side comparison using  
3 the same library and water bacterial isolates. The current study compares four BST methods,  
4 ERIC-PCR, automated ribotyping (RiboPrinting), Kirby Bauer antibiotic resistance analysis  
5 (KB-ARA), and PFGE, using the same known source isolates for the library and for the  
6 identification of the same water isolates. Each method measures different genotypic or  
7 phenotypic traits and therefore the methods have potentially different abilities to distinguish  
8 host-specific strains. Importantly, it was hypothesized that a combination of these methods  
9 through the use of composite data sets would be more useful than any individual method.

10           The objectives of the current study were (i) to compare the identification ability of the  
11 four BST methods individually and in combination through the use of composite data sets, and  
12 (ii) to evaluate the use of the developed data sets for the identification of faecal contamination  
13 sources in two Central Texas lakes suspected of being impacted by agricultural operations and  
14 dairy cattle.

15

16

## 17 **MATERIALS AND METHODS**

18

### 19 **Source sample collection**

20

21 To identify the sources of faecal contamination in the Lake Waco and Belton Lake, Central  
22 Texas watersheds, a known source library of *E. coli* isolates from potential faecal pollution  
23 sources was developed. Potential pollution sources were identified through a sanitary survey of

1 the watersheds and data were used to collect a total of 994 known source faecal and sewage  
2 samples from the Lake Waco and Belton Lake watersheds over a 13-month period. Municipal  
3 wastewater treatment plant influent/effluent and household septage samples (collectively referred  
4 to as “domestic sewage”), livestock, wildlife, and pet faecal samples were obtained from a  
5 variety of sources throughout the watersheds. To the extent possible, known source samples  
6 were collected directly from the source animals. An exception was the domestic sewage samples  
7 which were collected from wastewater treatment plants and septic tanks, as opposed to individual  
8 human samples. In some cases, wildlife samples had to be collected indirectly from “found”  
9 faecal samples. The sources of these “found” wildlife faecal samples were identified to the  
10 lowest practical taxonomic level by experienced field biologists. No samples of uncertain  
11 sources were used for library development. Only a single sample was collected from an  
12 individual animal, and multiple animals from each source group were sampled for  
13 representativeness. Fresh animal faecal samples were collected aseptically, using a sterile  
14 spatula or swab, into sterile, screw-cap polypropylene specimen tubes. Domestic sewage  
15 samples were immediately streaked after collection onto USEPA Method 1603 modified  
16 membrane thermotolerant *E. coli* medium (modified mTEC) (USEPA, 2002, USEPA, 2005a)  
17 and incubated overnight at 44.5°C. Faecal samples and parafilm-sealed modified mTEC plates  
18 were kept on ice and shipped overnight to the Texas Agricultural Experiment Station at El Paso  
19 (TAES-El Paso) laboratory for isolation of *E. coli*. Presumptive *E. coli* were isolated from faeces  
20 and modified mTEC plates within three days of receipt at the laboratory. An additional  
21 100 *E. coli* isolates from known wildlife sources obtained in a previous BST study in South  
22 Texas were characterized and included in the library (Mott and Lehman, 2001).

23

1

2 **Water sample collection**

3

4 Monthly water samples were collected over ten months from four stations in Lake Waco, two  
5 stations on the North Bosque River upstream of Lake Waco, four stations in Belton Lake, and  
6 one station on the Leon River upstream of Belton Lake. For most sampling stations, five  
7 independent water samples were collected on each sampling date, 1-2 min and approximately 1  
8 to 3 meters apart. This was done by sampling five points evenly spaced around the perimeter of  
9 the sampling boat. At the two stations near the dams and drinking water intakes of Lake Waco  
10 and Belton Lake, samples were collected in duplicate for a total of 10 samples per event. Water  
11 samples were collected directly from the lake or stream (approximately 0.5 meters below the  
12 surface) into sterile wide-mouthed polypropylene bottles. Care was exercised to avoid the  
13 surface layer of water, which can be enriched with bacteria and may not be representative of the  
14 bulk water. All water samples were placed on ice and transported to the processing laboratory  
15 for analysis within 6 h of collection using modified mTEC medium.

16

17

18 **Cultivation, confirmation, and storage of *E. coli* isolates**

19

20 Presumptive *E. coli* from domestic sewage modified mTEC plates were again streaked onto  
21 modified mTEC medium and were incubated at  $44.5 \pm 0.2^\circ\text{C}$  for 20-24 h. At least two attempts  
22 to isolate *E. coli* were made before considering a source sample negative for *E. coli*.

23 Presumptive *E. coli* colonies on modified mTEC plates from source samples were picked and

1 streaked on nutrient agar with 4-methylumbelliferyl- $\beta$ -D-glucuronide (NA-MUG) to confirm  
2 glucuronidase activity and culture purity.

3 Water sample modified mTEC plates with presumptive *E. coli* colonies were wrapped in  
4 parafilm and shipped on ice overnight to the TAES-El Paso laboratory for *E. coli* isolation,  
5 confirmation, and storage. *E. coli* colonies from the modified mTEC plates were picked and  
6 streaked for purity NA-MUG as described for faecal specimens and domestic sewage samples.

7 Well-isolated colonies of purified *E. coli* cultured on either NA-MUG or brain heart  
8 infusion agar (BHI) were resuspended in tryptone soy broth (TSB) with 20% glycerol in  
9 cryovials and stored at -70 to -80°C for long-term storage. Up to 5 *E. coli* isolates were stored for  
10 each source sample, while up to 12 isolates per water sample were stored.

11

12

### 13 **ERIC-PCR**

14

15 *E. coli* isolates from water samples and faecal samples were fingerprinted using the  
16 enterobacterial repetitive intergenic consensus polymerase chain reaction (ERIC-PCR)  
17 (Versalovic *et al.*, 1991). The ERIC-PCR conditions were modified as previously described (Di  
18 Giovanni *et al.*, 1999). Each 50  $\mu$ l reaction mixture contained 1X PCR buffer with Mg, 200  $\mu$ M  
19 of each dNTP (Amersham Biosciences, Piscataway, New Jersey), 600 nM each ERIC primer  
20 (Invitrogen, Carlsbad, CA), 1.5  $\mu$ g/ $\mu$ l bovine serum albumin (BSA), 2.5 units AmpliTaq Gold  
21 (Applied Biosystems, Foster City, California), and 5  $\mu$ l cell suspension (colonies from overnight  
22 cultures of *E. coli* isolates on BHI plates suspended in 100  $\mu$ l sterile molecular grade water).  
23 PCR amplification was performed using a DNA Engine Dyad Peltier Thermal Cycler (MJ

1 Research, Bio-Rad Laboratories, Hercules, CA) as follows: initial denaturation at 95°C for 7  
2 min; 35 cycles of denaturation at 94°C for 30 s; annealing at 52°C for 1 min; and extension at  
3 72°C for 5 min; followed by a final extension at 72°C for 10 min. Amplification products were  
4 stored at -80°C until analyzed by agarose gel electrophoresis. Ten microliters of loading buffer  
5 was added to each PCR well, and 10 µl of each amplification product was loaded onto a 20 cm  
6 by 25 cm 2% (wt/vol) agarose gel prepared with 1X Tris-borate-EDTA buffer and 30 tooth, 1  
7 mm thick comb. A 100 bp DNA ladder with fragments ranging from 100 to 1,500 bp in 100 bp  
8 increments and an additional band of 2642 bp (DNA Marker XIV; Roche Molecular,  
9 Indianapolis, Indiana) was added to the first and last lanes and after every six samples. A no  
10 template control and a quality control strain, *E. coli* QC101 (ATCC 51739), was run in each PCR  
11 batch and on each gel to evaluate method reproducibility and assure lack of DNA contamination.  
12 Electrophoresis was done at 4°C for 1 h at 100 V, followed by 4 h at 200 V with constant buffer  
13 recirculation. Gels were stained for 20 min in 1X Tris-borate-EDTA buffer containing 0.5 µg of  
14 ethidium bromide per ml. Gel images were captured as tagged image file format (TIFF) files  
15 with a Gel Logic 200 documentation system (Kodak, Rochester, New York).

16

17

## 18 **RiboPrinting**

19

20 *E. coli* isolates were analyzed by automated ribotyping (RiboPrinting) using the RiboPrinter  
21 Microbial Characterization System (DuPont Qualicon, Wilmington, Delaware) according to  
22 manufacturer's instructions. RiboPrinting was performed using the substitute enzyme batch KHB  
23 process (digestion 37°C for 20 min) with *Hind*III restriction endonuclease (Cat. #R0104M, New

1 England Biolabs, Beverly, Massachusetts). *HindIII* was prepared as a 50 U/ $\mu$ l solution by the  
2 addition of 26.5  $\mu$ l *HindIII* and 26.5  $\mu$ l of NEB 10X Buffer 2 in a 500- $\mu$ l microfuge tube (Cat.  
3 #72730-005 Sarstedt, Newton, North Carolina).

4 The RiboPrint patterns were partially processed by the RiboPrinter system software,  
5 automatically reducing the background and noise and normalizing the band positions using the  
6 system's DNA size standards. RiboPrint patterns were further automatically processed when  
7 downloaded into the BioNumerics ver. 4.0 software (Austin, Texas) using the Load Samples  
8 Amplified script (DuPont Qualicon). Reproducibility was determined by replicate analysis of  
9 the *E. coli* QC101 quality control isolate in every fourth batch (1 of 32 isolates) of samples  
10 analyzed in a single day or each day the RiboPrinter was used.

11

12

### 13 **Pulsed-field gel electrophoresis (PFGE)**

14

15 The CDC PulseNet standardized protocol for PFGE using *XbaI* as the restriction endonuclease  
16 (CDC, 2000, Swaminathan *et al.*, 2001) was used to generate the PFGE fingerprints for selected  
17 library and water *E. coli* isolates using the following conditions. The concentration of the cell  
18 suspension was adjusted using optical density measurements on a spectrophotometer; the  
19 optional pre-restriction incubation step was followed; and restricted plug slices were loaded into  
20 the wells of the gel. PFGE was performed on a CHEF Mapper XA Pulsed Field Gel  
21 Electrophoresis System (Bio-Rad Laboratories, Richmond, California) for 22 h at 12°C with  
22 initial switch time of 1 s, final switch time of 90 s, and voltage of 6.0 V/cm. Three lanes of  
23 Lambda ladder PFGE marker (New England Biolabs, Ipswich, Massachusetts) were included as

1 standard references to normalize the patterns. Reproducibility was determined by replicate  
2 analysis of the *E. coli* QC101 quality control isolate in every gel that was run. Gel images were  
3 captured as tagged image file format (TIFF) files with a MultiDoc-It Digital Imaging System  
4 (UVP, Upland, California).

5  
6

### 7 **Kirby-Bauer antibiotic resistance analysis (KB-ARA)**

8

9 The Kirby-Bauer disk diffusion method of antibiotic resistance analysis (KB-ARA) was  
10 performed following clinical laboratory methods as described by the National Committee for  
11 Clinical Laboratory Standards (NCCLS, 2002). Isolates were grown on tryptic soy agar (TSA)  
12 overnight at 35°C, transferred to tryptic soy broth (TSB), and incubated in a 35°C shaker for 2-6  
13 h. Each isolate was then plated onto two Mueller Hinton agar plates, with 10 antibiotic disks  
14 (BD Diagnostic Systems, Sparks, Maryland) per plate. A panel of 20 antibiotics was used in this  
15 study: ampicillin, 10 µg; augmentin, 30 µg; cefazolin, 30 µg; cefotaxime, 30 µg; ceftazidime, 30  
16 µg; ceftriaxone, 30 µg; chloramphenicol, 30 µg; ciprofloxacin, 5 µg; doxycycline, 30 µg;  
17 enrofloxacin, 5 µg; gentamicin, 10 µg; imipenem, 10 µg; kanamycin, 30 µg; nalidixic acid, 30  
18 µg; neomycin, 30 µg; spectinomycin, 100 µg; streptomycin, 10 µg; sulfamethoxazole  
19 trimethoprim, 23.75/1.25 µg; sulfisoxazole, 0.25 mg; and tetracycline, 30 µg. After incubation at  
20 35°C for 16-18 h, the diameter of the zone of inhibition (to the nearest whole mm) of bacterial  
21 growth around each disk was analyzed with an automated plate reader system (BIOMIC; Giles  
22 Scientific, Santa Barbara, California). *Pseudomonas aeruginosa* ATCC 27853, *Staphylococcus*  
23 *aureus* ATCC 25923, and *E. coli* ATCC 25922 quality control strains were run with each batch

1 of samples. Every tenth isolate was run in duplicate to assure reproducible profiles. The image  
2 analysis system included EXPERT software which checked controls and flagged unlikely test  
3 results. Discriminant analysis is commonly used for statistical analysis of KB-ARA data and was  
4 used in this study. The KB-ARA fingerprint of each isolate was compiled into a library of known  
5 sources and analyzed by discriminant analysis using SPSS ver. 12.0. Isolates were divided into  
6 the seven source classes as described below with no allowance for isolates to be left unidentified.  
7 In addition, KB-ARA profiles were also analyzed using BioNumerics and Pearson's product-  
8 moment correlation coefficient (similar to the ERIC-PCR, RiboPrint and PFGE data) by  
9 considering the zone of inhibition measurements as character data with numerical values in a  
10 closed data set.

11

12

### 13 **Gel image processing and preliminary data analysis**

14

15 BioNumerics ver. 4.0 software (Applied Maths, Austin, Texas) was used to analyze the BST data  
16 for this project in three ways: 1) processing gel images, 2) determining the relationships between  
17 the isolates by comparing their ERIC-PCR fingerprint patterns for development of the BST  
18 libraries, and 3) for identification of water isolates based on their ERIC-PCR, RiboPrint, PFGE  
19 and KB-ARA fingerprints.

20 Digital ERIC-PCR and PFGE gel images were imported into BioNumerics version 4.0  
21 and processed using the default settings. Fingerprints were compared using the curve-based  
22 Pearson's product-moment correlation coefficient with optimization and position tolerance

1 settings of 1.56% and 1.00%, respectively. Dendrograms were constructed using the unweighted  
2 pair group method with arithmetic means (UPGMA).

3  
4

### 5 **Library construction**

6

7 ERIC-PCR fingerprinting was used to screen isolates from the same sample in order to identify  
8 clonal isolates and assure as diverse a known source library as possible. One to three *E. coli*  
9 isolates from each known source sample were fingerprinted using ERIC-PCR, then compared to  
10 each other and the previously selected library isolates. Isolates from the same sample having  
11 greater than 80% similarity were considered clonal. This similarity cutoff was based on  
12 preliminary analysis of replicate ERIC-PCR fingerprints for laboratory QC isolates which were  
13 found to be reproducible with approximately 85% similarity. Isolates representing the different  
14 ERIC-PCR types identified for each sample were compared to the *E. coli* isolates previously  
15 selected from other samples for the source library. Isolates having more than an 80% similarity  
16 to an existing library isolate were considered already represented in the library, without regard to  
17 source. Each isolate from a single known source sample that was novel (<80% similarity) as  
18 compared to the library isolates was selected for inclusion in the library. At least one *E. coli*  
19 isolate from each known source sample was included in the library. Also, if an isolate had more  
20 than an 80% similarity to only a single library isolate, then it was also selected for the library,  
21 regardless of the selection of other isolates from the same sample, in an attempt to confirm the  
22 isolation of the same strain of *E. coli* from a separate sample. As a result, library dendrogram  
23 clusters were always composed of isolates from different samples. If all ERIC-PCR types

1 represented by the isolates from a single sample were already present in the library, then an  
2 isolate representing the most abundant (and as such being the most representative) ERIC-PCR  
3 type for that sample was selected for the library. Therefore, abundant/common strains of *E. coli*  
4 isolates from different samples and animals were represented in the library, but clonal isolates  
5 from individual samples were excluded. If isolates were equally abundant in a sample, then the  
6 isolate that best filled in the dendrogram was selected. Building the library was a dynamic  
7 process; isolates were added to the library as their ERIC-PCR patterns were processed. Once  
8 chosen for the library, the selected isolates were then fingerprinted by the other three BST  
9 methods (RiboPrinting, PFGE, and KB-ARA).

10

11

## 12 **Composite data sets and congruence of methods**

13

14 BioNumerics has the unique ability to allow the construction of composite data sets which takes  
15 into account all fingerprint profiles or different combinations for each isolate. To incorporate all  
16 four BST methods, KB-ARA zone of inhibition data were treated as a character set using the  
17 BioNumerics software. BioNumerics was used to calculate a composite data set using the  
18 unweighted averages of the individual BST method similarity matrices, resulting in a new, single  
19 similarity matrix which incorporated attributes of each individual method. The minimum  
20 similarity cutoff used for matching composite data sets was 70% to allow for variation of the  
21 individual methods and to accommodate the diversity of the PFGE fingerprints. This minimum  
22 similarity cutoff was determined by calculating the rates of correct classification (RCC) for the

1 library isolates and discerning where the highest RCCs balanced with the number of water  
2 isolates that would be left unidentified at each minimum similarity cut off.

3 The congruence (concordance) between the groupings of isolates by individual BST  
4 methods and different combinations of composite data sets was determined with BioNumerics  
5 using the Pearson's product moment correlation coefficient. Using the library isolates,  
6 congruence measurements were performed for individual BST methods, all two and three-  
7 method combinations, and the four-method composite data set. In addition, congruence  
8 measurements of selected composite data sets were performed to compare the source class  
9 identification of unknown water isolates.

10

11

## 12 **Library evaluation**

13

14 The library was evaluated by Jackknife analysis, in which isolates were removed from the library  
15 one at a time and treated as unknowns for identification. RCCs were calculated as the  
16 percentage of library isolates correctly identified back to their source class out of the number of  
17 attempts made to identify isolates from that source. Since the library was intentionally  
18 constructed to include unique isolates, RCC values were not penalized for isolates which were  
19 left unidentified. To determine RCCs, matches to host source class were used, as opposed to  
20 matches with individual animal species. For example, an isolate from a wild goose matching  
21 with an isolate from a wild duck was considered a correct match for the avian wildlife source  
22 class. In the rare case of a tie (same percentage of similarity for the best match), the benefit of  
23 the doubt was given and the isolate most similar in host source class was selected as the match.

1           The four-method composite library was further evaluated for sensitivity and specificity as  
2 described in the USEPA Microbial Source Tracking Guide Document (USEPA, 2005b) for both  
3 seven-way and two-way (i.e. human vs. non-human) splits of source classifications. Sensitivity  
4 reflects the percentage of isolates giving a host source-specific fingerprint, and is the same as the  
5 RCC. Specificity measures how well a BST method can discriminate between source categories.  
6 Specificity was determined by performing Jackknife analyses and examining the identifications  
7 of source isolates from other than the respective source being evaluated. The number of isolates  
8 that tested negative (or true negatives) and were correctly identified as not belonging to the test  
9 source class were divided by the sum of this number plus isolates incorrectly identified to the  
10 respective source class (false positives).

11

12

### 13 **Blind QC challenge**

14

15 In addition to the use of the *E. coli* QC101 isolate, BST method analytical accuracy and precision  
16 were quantified through a special QC challenge with blinded safeguards. The funding agency  
17 project manager randomly selected 60 isolates from the list of 980 different *E. coli* from known  
18 sources to prevent any bias in the selection of isolates. From the list of 60 isolates, the list was  
19 narrowed down to 30 isolates by eliminating many duplicate source animal species and ensuring  
20 that the percentage of domestic sewage isolates was similar to the percentage were presented by  
21 this source class in the total library. The list of the 30 selected isolates was then provided to the  
22 TAES-El Paso laboratory and triplicate cultures of each isolate were prepared and sent to the  
23 Parsons project manager. The Parsons project manager randomly selected 10 of the 30 isolates,

1 blind labelled the triplicate cultures of each, and the cultures were shared between the BST  
2 laboratories for subculturing. Each laboratory analyzed the blind isolates using their respective  
3 BST technique. Each attempted to identify the 10 sets of triplicates, identify the triplicates to the  
4 correct library isolate (of the 30 possible), and to the correct source class. After each lab  
5 reported its individual results to the funding agency project manager, the labs shared their raw  
6 data with the TAES-El Paso laboratory for composite data set analysis. The key of the blind  
7 isolates was not provided to the participants until after the TAES-El Paso laboratory reported the  
8 composite data set results to the funding agency project manager.

9

10

#### 11 **Identification of sources of unknown *E. coli* isolates**

12

13 Water isolates were fingerprinted by all four BST methods. To identify unknown isolates using  
14 each molecular technique (ERIC-PCR, RiboPrinting, or PFGE), a one-to-one matching  
15 epidemiological-like approach was used to find the best match of a fingerprint to a single library  
16 isolate. Best matching was accomplished using a custom BioNumerics script provided by  
17 Applied Maths. Pearson's product-moment correlation coefficient was used with the Best  
18 Matches script to identify the single library isolate with the highest similarity to each unknown  
19 water isolate. This is similar to the maximum similarities approach, with the difference that the  
20 match is to a single isolate rather than to a single operational taxonomic unit or cluster. The  
21 minimum similarity cutoff for a match was  $\geq 85\%$  for the ERIC-PCR and RiboPrinting  
22 fingerprints, and  $\geq 70\%$  for the PFGE fingerprints. These similarity cutoffs were based on the  
23 long-term reproducibility of the fingerprints obtained with the *E. coli* QC101 strain using each

1 method. Water isolates matching below these minimum similarity cutoffs were considered  
2 unidentified.

3         Although fingerprint profiles were considered a match to a single entry, identification  
4 was to the host source class (as done for calculation of library RCCs), and not to the individual  
5 animal species represented by the matching library isolate. Host sources were divided into seven  
6 groups, 1) domestic sewage (municipal wastewater treatment plant influent/effluent and septage  
7 samples); 2) pet (dogs, cats, and a rabbit); 3) cattle (beef and dairy); 4) other livestock, avian  
8 (turkey, chicken, state fair animals); 5) other livestock, non-avian (e.g. horse, goat, pig, state fair  
9 animals); 6) wildlife, avian (e.g. wild birds including pigeon, grackle, goose); and 7) wildlife,  
10 non-avian (e.g. deer, raccoon, skunk, armadillo, possum, rabbit, feral hog, javelina). The  
11 division of host sources into these particular classes was based on the anticipated usefulness for  
12 the development of best management practices and input from stakeholders.

13

14

## 15 **RESULTS**

16

### 17 **Source and water isolates**

18

19 From the 1094 collected source samples, a total of 3231 *E. coli* isolates were obtained and  
20 stored. A total of 2275 *E. coli* isolates from source samples were screened using ERIC-PCR, and  
21 980 were subsequently selected for inclusion in the library and fingerprinting with the additional  
22 BST methods. From the 11 water sampling stations a total of 650 water samples were collected.  
23 Of these, 412 samples were positive for *E. coli*, with 631 *E. coli* isolates obtained and stored. All

1 631 water isolates were analyzed using all four BST methods. However, PFGE patterns could  
2 not be generated for approximately 10% of the known library and water isolates, since some  
3 bacterial strains have genomic DNA that do not permit effective restriction endonuclease  
4 digestions. Therefore, method and composite data comparisons were limited to those isolates  
5 successfully fingerprinted by all four BST methods: 883 known source isolates from 745  
6 samples (Table 1) and 555 water isolates from 412 samples.

7  
8

### 9 **Occurrence of unique fingerprints in the library**

10

11 Using the similarity dendrogram and the defined minimum similarity cutoffs ( $\geq 85\%$  for ERIC-  
12 PCR and RiboPrinting,  $\geq 70\%$  for PFGE fingerprints and four-method composite), the number of  
13 unique fingerprints in the 883 *E. coli* library was determined for each BST method and the four-  
14 method composite data set. As expected based on the predicted resolution of each method,  
15 PFGE had the highest number of unique fingerprints while KB-ARA had the least. The number  
16 of unique fingerprints determined using each of the BST methods from highest to lowest were as  
17 follows: PFGE, 614; ERIC-PCR, 346; four-method composite data set, 299; RiboPrinting, 146;  
18 and KB-ARA, 63. For the 299 unique fingerprints identified using the four-method composite  
19 dataset, 168 were represented by a single isolate.

20

21

### 22 **Blind QC challenge**

23

1 Identifications of the blind QC isolates were based on best matching and subjective visual  
2 inspection of the dendrograms for the three molecular methods and four-method composite data  
3 set, while the KB-ARA data were analyzed using discriminant analysis with visual comparisons  
4 of the zone diameter data and subjective judgment.

5 Fingerprints for the blind QC isolates were grouped in their own dendrogram to aid  
6 identification of the triplicates. Identification was relatively straightforward for the ERIC-PCR,  
7 PFGE, and the four-method composite data set, but was more complex for RiboPrinting data.  
8 PFGE fingerprints could not be generated for two QC library isolates. Since identification of  
9 these two isolates was not possible with this method, PFGE was not penalized for its inability to  
10 identify these isolates.

11 Overall, the ERIC-PCR, RiboPrinting, and PFGE methods performed equally well, with  
12 100% identification of replicate isolates (precision) and 70 to 90% accuracy in identification of  
13 replicate isolates to the correct single QC library isolate (method accuracy) and correct source  
14 class (source identification accuracy) (Figure 1). On the other hand, the KB-ARA data analyzed  
15 using discriminant analysis scored only 40% for identification of replicate isolates (precision)  
16 and 50% for method and source identification accuracy.

17 Most important, however, are the four-method composite data set results. Best matching  
18 identification using the composite data set correctly identified 100% of the replicate QC cultures  
19 (precision), and had 100% accuracy for *E. coli* strain and source class identification of the  
20 isolates. Therefore, the four-method composite performed better than any single method.

21

22 **RCCs for each BST method and the four-method composite data set**

23

1 PFGE tended to have the highest RCCs (Table 2), ranging from 0% to 95% for the seven-way  
2 split of sources, but almost half of the library isolates were left unidentified. Therefore, while  
3 there was high confidence in the PFGE matches, fewer identifications could be made, even with  
4 lower similarity cut-offs (data not shown). The four-method composite data library had the next  
5 highest RCCs, ranging from 22% to 83% for the seven-way split of sources, and averaged only  
6 10% of the library isolates left unidentified. In particular, the four-method composite library had  
7 good RCC values for source classes of particular interest: 83% for domestic sewage and 61% for  
8 cattle. Interestingly, RCCs for the KB-ARA data using best matching as opposed to discriminant  
9 analysis were higher for four of seven source classes; including the domestic sewage, cattle, and  
10 wildlife non-avian classes.

11

12

### 13 **Cross-classification of the library source isolates**

14

15 It was expected that some *E. coli* would not be host-specific and could come from several host  
16 source classes, leading to possible source cross-classification (identification). There are two  
17 general ways to approach source cross-classification. One is to determine how isolates from a  
18 specific source class are identified, which is the RCC. The second approach is to evaluate the  
19 true source class isolates identified to a particular source class, which was also done in this study.

20 Cross-validation Jackknife analysis of the four-method composite data library revealed  
21 that, in general, there were low levels of cross-classification (Figure 2). Figure 2 shows that for  
22 each of the seven source classes, the highest bar within each source class belongs to the correct  
23 source (and corresponds to the RCC). RCCs for each of the source classes were three- to seven-

1 fold greater than random chance based on library composition. Only the other livestock avian  
2 source class had cross-classifications in another single source class (i.e. cattle) greater than its  
3 correct classification. Cross-classifying isolates were not excluded from the library since a best  
4 match algorithm was used.

5

6

### 7 **Effects of different minimum similarity cut-offs on the four-method composite data RCCs** 8 **and identification of water isolates**

9

10 Minimum similarity cutoff values from 65% to 80% were evaluated (Table 3). The minimum  
11 similarity cutoff selected for the four-method composite data set was 70% to allow for variation  
12 of the individual methods and to accommodate the diversity of the PFGE fingerprints. This  
13 minimum similarity cutoff also provided a balance between the RCCs and the number of water  
14 isolates that would be left unidentified. In particular, it was a goal of the study that no more than  
15 10% of the water isolates be left unidentified. At the 70% minimum similarity cutoff only 9% of  
16 the water isolates were left unidentified, while at the 75% and 80% minimum similarity cutoffs  
17 24% and 56% of the water isolates were left unidentified, respectively. Increasing the minimum  
18 similarity cutoff above 70% resulted in only a marginal increase in the RCCs for several of the  
19 important source classes, including the cattle and domestic sewage classes. Further, based on  
20 Jackknife analysis, only about 10% of the source library isolates were left unidentified using the  
21 70% minimum similarity cutoff, while this increased to about 30% and 50% of the library  
22 isolates left unidentified at the 75% and 80% minimum similarity cutoffs, respectively.

23

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

**Library quality measures**

The four-method composite library with seven-way split of source classifications was evaluated for sensitivity and specificity. Although there is no consensus, specificity values below 80% are considered of questionable discriminatory power (USEPA, 2005b). In this study, the four-method composite library specificity values for each source class were all above 80% (range 89% to 99%) for the seven-way split of source classifications. Sensitivities and specificities were also determined for a domestic sewage and animal two-way split of source classifications. The sensitivities (RCCs) were 83% and 95% for domestic sewage and animal source classifications, respectively. Specificities were 95% for the domestic sewage source class and 83% for the animal source classification.

**Identification of water isolates using individual BST methods and four-method composite data**

A comparison was made between the percent identification to source class for the 555 unknown water isolates using each BST method and the four-method composite data set (Figure 3). ERIC-PCR, RiboPrinting, PFGE, and KB-ARA (by best match) all identified wildlife, cattle and domestic sewage as the leading sources of contamination. As expected, the four-method composite identified the same major sources of contamination, since it reflects an average of the

1 similarity matrices of the individual methods. The KB-ARA discriminant analysis results also  
2 agreed that wildlife was the leading source of contamination. However, identification of other  
3 sources of contamination based on discriminant analysis of the KB-ARA data differed from the  
4 three molecular methods, as well as the KB-ARA data analyzed using best matching. Best  
5 match analysis of KB-ARA data had higher or similar RCCs than discriminant analysis of the  
6 data for the potential major sources of pollution (i.e. wildlife, cattle, and domestic sewage).

7

8

### 9 **Congruence of BST methods and composite datasets**

10

11 The dendrogram in Figure 4 depicts the relationship between the different techniques, composite  
12 data sets, and the percent similarity of each method or combination to the four-method composite  
13 data set. As expected, the methods do not agree 100% with each other due to the differences in  
14 resolution between the methods and the fact that they each measure different attributes of *E. coli*.

15 The congruence measurement revealed that the two individual techniques most similar to the  
16 four-method composite data set were ERIC-PCR and RiboPrinting, with 77.9% and 60.8%  
17 similarity, respectively. This was also not surprising, since for the methods tested, PFGE and  
18 KB-ARA are at the extremes of the spectrum for their ability to resolve differences between  
19 bacterial isolates. The ERIC-PCR RiboPrinting composite data set (ERIC-RP) was found to be  
20 the closest two-method combination (90.7% similar) to the four-method composite data set.

21

22

### 23 **RCCs for two-method composite data sets**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

It may not be feasible in terms of both cost and time considerations to perform all four BST methods for future BST studies. Therefore, several different two-method composite data sets were selected for comparison to the four-method composite data set. The ERIC-RP composite data set was selected since it was found to have the highest congruence to the four-method composite data set. The ERIC-ARA composite data set was selected since these individual methods were the most cost effective and amenable to technology transfer. Lastly, the PFGE-ARA composite data set was selected since these methods had the highest and lowest strain resolution, respectively. As expected, the four-method composite data set still had the highest RCCs for most of the source classes (Table 4). Surprisingly, the RCCs of the three different two-method composite data sets were comparable to each other, and in general the RCCs for the ERIC-ARA and ERIC-RP composite data sets were higher than the RCCs of the ERIC-PCR, KB-ARA, or RiboPrinting methods alone. The RCCs of the ARA-PFGE composite data set were not as high as the RCCs of the PFGE method alone, but the ARA-PFGE composite dataset had significantly fewer unidentified library isolates.

**Identification of water isolates using selected two-method and the four-method composite data sets**

Based on the four-method composite data set, approximately 25% of the *E. coli* water isolates were identified to the avian wildlife source class, followed by 18% identified to the non-avian wildlife source class (Figure 5). Thus, it was estimated that wildlife contributed over 40% of the

1 *E. coli* from all water sampling sites combined. The remaining source allocations for the *E. coli*  
2 water isolates were 31% livestock (17% cattle, 3% other avian livestock, and 11% other non-  
3 avian livestock); 14% domestic sewage; and 3% pets. The source of 49 of the 555 (9%) *E. coli*  
4 water isolates could not be identified with acceptable confidence with the four-method composite  
5 data set.

6 To explore the use of two-method composite data sets, the source class identifications of  
7 the 555 water isolates using ERIC-RP, ERIC-ARA and ARA-PFGE composite data sets were  
8 also determined (Figure 5). Similar to the four-method composite data set, each of the two-  
9 method composite data sets identified wildlife, livestock, and domestic sewage as the three  
10 leading sources of *E. coli*. In addition, identification of the *E. coli* water isolates to the cattle  
11 source class were similar for the two method composite data sets compared to the four-method  
12 composite data set, ranging from 14% to 22%. However, 24% of the water isolates could not be  
13 identified using the ERIC-RP composite data set.

14

15

## 16 **DISCUSSION**

17

18 To our knowledge, this is the first study to directly compare BST methods and composite data  
19 sets by using the same collection of known source and water *E. coli* isolates. The methods used  
20 in this study covered the spectrum in cost, ease of use, and discriminatory ability. This allowed  
21 us to not only evaluate the practical application of these methods for the identification of human  
22 and animal sources of faecal pollution, but also to perform a comprehensive comparison of the  
23 methods and composite data sets for further consideration and use in future studies. Although

1 this study was performed in Texas and focused on the use of *E. coli*, the methodology used in  
2 this study for comparison of BST techniques and use of composite data sets are applicable to  
3 other locales and BST targets (e.g. *Enterococcus* spp.).

4 The approach used to construct the library in this study was different than most previous  
5 BST studies. Historically, microbiologists have known that the culture medium used to isolate  
6 bacteria has an effect on the types and strain diversity of the organisms recovered. For example,  
7 the types of *E. coli* isolated from source faecal specimens using clinical media may be different  
8 from the types of *E. coli* isolated from water using regulatory testing media. In this study, this  
9 potential problem was minimized by using the same medium, modified membrane  
10 thermotolerant *E. coli* agar (modified mTEC), for the isolation of *E. coli* from known source and  
11 water samples. From previous BST studies, it is known that multiple *E. coli* isolates from the  
12 same source sample are frequently clonal, and their inclusion in the same library may inflate  
13 rates of correct classification determined by Jackknife analysis (Johnson *et al.*, 2004, USEPA,  
14 2005b). Therefore, in our study, we used high numbers of source samples from individual  
15 animals and selected only one to three *E. coli* isolates per sample for inclusion in the library.  
16 Further, ERIC-PCR was used to screen the isolates from each sample to exclude clonal isolates  
17 and to maximize the diversity of *E. coli* strains selected. Therefore, the developed library of 883  
18 *E. coli* isolates is in essence much larger, since it is the result of screening 2275 isolates.  
19 Adequate library size is in part dictated by the resolution of the BST technique being used. For  
20 example, with a high resolution technique such as PFGE, a much larger library would be needed  
21 than for a lower resolution technique such as KB-ARA. In this study, the 883 isolates selected  
22 for the library represented 614 unique fingerprints for PFGE, compared to 63 for KB-ARA. For  
23 ERIC-PCR, the 883 *E. coli* isolates represented 346 unique fingerprints. Our ERIC-PCR results

1 are similar to the rep-PCR results of Johnson *et al.* (Johnson *et al.*, 2004) who reported screening  
2 2466 *E. coli* isolates from 982 individual samples, resulting in 657 unique fingerprints. The  
3 authors also reported that approximately 59% of their 657 unique fingerprints were present only  
4 once in the database. In our study, the four-method composite data set of the 883 isolates  
5 represented 299 unique fingerprints, of which 56% were present only once in the library. While a  
6 diverse library may pose some challenges for evaluation of library internal accuracy (i.e. RCCs),  
7 it is advantageous for the identification of water isolates. In theory, a larger library would be  
8 needed for the four-method composite to represent the diversity of *E. coli* strains in source  
9 samples. However, in practice, 91% of the 555 water isolates obtained from 412 different  
10 samples matched a pattern in the developed library, suggesting that the current library size was  
11 adequate for its purpose.

12         BST results can be affected by the statistical methods used for data analysis and  
13 identification of unknown isolates (Albert *et al.*, 2003, Ritter *et al.*, 2003, Hassan *et al.*, 2005). In  
14 this study, one-to-one best matching with a minimum similarity cutoff was used for the  
15 identification of water isolates and for performing library Jackknife analyses. Best matching is  
16 similar to the maximum similarities approach in which an unknown isolate is identified to a  
17 single user-defined library group or operational taxonomic unit that contains the member with  
18 the highest similarity to the unknown. The difference between best matching and maximum  
19 similarities is that with best matching the identification is to a single library isolate, rather than to  
20 a user-defined group of isolates. Best matching is more amenable to dynamic libraries, avoiding  
21 the need to classify each isolate into a user-defined group. It is also more flexible in that it allows  
22 the user to easily sort identifications based on any characteristic of the library isolate (e.g.,  
23 specific animal species, sampling location, sampling date, etc.), and facilitates manual

1 verification of matches. The effect of the minimum similarity cutoff value on RCCs and  
2 identification of water and library isolates was also evaluated in this study. While increasing the  
3 minimum similarity cutoff value for the four-method composite library from 70% to 80%  
4 resulted in a modest increase in RCCs, more than half of the library and water isolates were left  
5 unidentified, making method comparisons difficult.

6         Our blind QC challenge was similar to the replicate challenge done by Stoeckel *et al.*  
7 (2004) where the ability to identify replicates of library isolates was tested. Their results  
8 showed that PFGE identified 100% of their 26 QC replicates correctly, followed by 48% by  
9 REP-PCR, 23% by antibiotic resistance analysis, and 13% by ribotyping using *HindIII*. The  
10 higher rates of correct identification of QC challenge isolates observed in our study may have  
11 partially been due to the reduced list of 30 possible QC isolates compared to the entire library as  
12 in the Stoeckel *et al.* study, although they only identified to source class and not to specific  
13 isolate as we did. It should be noted that our QC challenge design was more suitable for the  
14 genetic analyses than for the KB-ARA method using discriminant analysis. The KB-ARA  
15 profiles of the randomly selected isolates for the QC challenge were not distinct enough to easily  
16 distinguish or group the isolates into the three replicates. KB-ARA is generally used to classify  
17 isolates into broad source class categories, and the use of individual isolates for the QC challenge  
18 made discriminant analysis statistically inappropriate. This further highlights the limited ability  
19 of KB-ARA to resolve differences between highly similar bacterial strains. Although the KB-  
20 ARA data from the QC isolates were not suitable for discriminant analysis, they were  
21 successfully incorporated into the four-method composite data set for best match identification.  
22 In a previous study examining the effects of different statistical analyses on BST results,  
23 maximum similarity analysis of antibiotic resistance data was found to have a 90% RCC for the

1 cattle source class, compared to an RCC of only 51% using discriminant analysis of the same  
2 data (Ritter *et al.*, 2003). More important for our study was the finding that the four-method  
3 composite data set performed better than any of the individual BST methods and correctly  
4 identified 100% of the QC challenge isolates.

5 A composite data set of ARA and rep-PCR fingerprints from a small group of  
6 *Enterococcus faecalis* isolates was recently reported to have higher confidence source  
7 identifications than either method alone (Genthner *et al.*, 2005). In our study, the advantage of  
8 the four-method composite data set over individual methods was observed not only in the QC  
9 challenge, but also in library RCCs. PFGE tended to have the highest RCCs, but almost half of  
10 the library isolates were left unidentified after Jackknife analyses, and only 20% (112) of the 555  
11 water isolates could be identified using PFGE. This indicates that an extremely large PFGE  
12 library would be required for the identification of water isolates. The next highest RCCs were for  
13 the four-method composite data set which had only 10% of the library isolates left unidentified  
14 after Jackknife analyses and was able to provide an identification for 91% of the 555 water  
15 isolates. In addition, the four-method composite library had RCCs of 83% and 61% for the  
16 domestic sewage and cattle source classes, respectively, which were of particular interest for  
17 these watersheds. Overall, the four-method composite data set had RCCs for each of the source  
18 classes that were three- to seven-fold greater than random chance based on library composition.

19 There are over 150 waterbodies in Texas alone that are impaired by high faecal bacteria  
20 levels, and it is clearly not possible for BST studies of the current magnitude to be done for each.  
21 The four-method composite library generated the most desirable BST results in this study.  
22 However, as few as two methods in combination may be useful based on congruence  
23 measurements, library internal accuracy (i.e. RCCs), and comparison of water isolate

1 identifications. In particular, the combinations of ERIC-PCR and RiboPrinting (ERIC-RP) or  
2 ERIC-PCR and Kirby-Bauer antibiotic resistance analysis (ERIC-ARA) appear promising.  
3 These two-method composite data sets were found to have 90.7% and 87.2% congruence,  
4 respectively, to the four-method composite data set. More importantly, based on the  
5 identification of water isolates, they identified the same leading sources of faecal pollution as the  
6 four-method composite library. ERIC-ARA has the lowest cost for consumables and has high  
7 sample throughput, but requires a considerable amount of hands-on sample and data processing.  
8 Due to the high cost of RiboPrinting consumables and instrumentation, ERIC-RP has a higher  
9 cost than ERIC-ARA. However, ERIC-RP has the advantage of automated sample processing  
10 and data preprocessing that the RiboPrinter system provides. Further research is needed to  
11 determine if a regional library built from different projects using the same protocols may be  
12 useful for the identification of water isolates from other watersheds. Regional ribotype libraries  
13 have shown some success in being able to distinguish *E. coli* of human or animal origin (Scott *et*  
14 *al.*, 2003) or for certain animal source groups (Hartel *et al.*, 2002) from different geographic  
15 locations.

16 Regulatory agencies continue to have high hopes and expectations for BST in aiding  
17 them to address water quality issues. Ideally, the regulatory agencies would prefer identification  
18 of pollution sources to the level of individual animal species. Performing a three-way split of  
19 pollution sources into domestic sewage, livestock and wildlife source classes would likely be  
20 more scientifically justified. The division of host sources into the seven classes in this study was  
21 a compromise between the capabilities of the BST techniques and their practical application.  
22 Since this study was initiated there have been significant developments in library-independent  
23 BST methods, including bacterial genetic markers specific to different animal sources and

1 humans (for example, Bernhard and Field, 2000, Dick *et al.*, 2005, Scott *et al.*, 2005, Hamilton *et*  
2 *al.*, 2006). Library-independent methods are cost-effective, rapid, and potentially more specific  
3 than library-independent methods. Concerns with many of the recently developed library-  
4 independent approaches include uncertainties regarding geographical stability of markers and the  
5 difficulty of interpreting results in relation to regulatory water quality standards and microbial  
6 risk, since some target microorganisms are not regulated. For future studies an assessment phase  
7 using a “toolbox” approach is recommended. The assessment phase should include targeted  
8 monitoring of suspected pollution sources, use of library-independent methods to identify the  
9 presence of domestic sewage pollution, and screening of water isolates from the new watershed  
10 against the previously developed library to determine the need for collection of local source  
11 samples and expansion of the library. Currently, we are cross-validating the library generated in  
12 this study with a library generated for another Texas watershed in an attempt to explore issues of  
13 geographical and temporal stability of BST libraries, refine library isolate selection, and  
14 determine accuracy of water isolate identification.

15  
16

## 17 **ACKNOWLEDGEMENTS**

18

19 The authors wish to thank Anthony Sisk, Adriana Galindo, Patricia Garrido, Walter Q.  
20 Betancourt, Jeanon Smith, Melissa N. Manuel, Pamela Brown, J. Eric Wilson, Russell  
21 McDowell, Bette Harrison and Sara Coley for sample processing. This project was funded by the  
22 U.S. Environmental Protection Agency under a CWA §319 grant and by the Texas State Soil and  
23 Water Conservation Board, Texas Farm Bureau, Brazos River Authority, City of Waco, and the

1 Texas Agricultural Experiment Station. Special thanks to Ned Meister, Kevin Wagner, T.J.  
2 Helton and Dr. Alan Jones. Thanks to the Kyle Kingsley, Lem Delrosario, and Brian West of  
3 Applied Maths for BioNumerics technical assistance.  
4  
5  
6

## 7 **REFERENCES**

- 8  
9 Albert, J. M., Munakata-Marr, J., Tenorio, L. and Siegrist, R. L. (2003) Statistical evaluation of  
10 bacterial source tracking data obtained by rep-PCR DNA fingerprinting of *Escherichia*  
11 *coli*. *Environ Sci Technol*, **37**, 4554-4560.
- 12 Bernhard, A. E. and Field, K. G. (2000) A PCR assay To discriminate human and ruminant feces  
13 on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA.  
14 *Appl Environ Microbiol*, **66**, 4571-4574.
- 15 CDC (2000) *Standardized molecular subtyping of foodborne bacterial pathogens by pulsed-field*  
16 *gel electrophoresis: a manual*, National Center for Infectious Diseases, Atlanta.
- 17 Di Giovanni, G. D., Watrud, L. S., Seidler, R. J. and Widmer, F. (1999) Comparison of parental  
18 and transgenic alfalfa rhizosphere bacterial communities using Biolog GN metabolic  
19 fingerprinting and enterobacterial repetitive intergenic consensus sequence-PCR (ERIC-  
20 PCR). *Microb Ecol*, **37**, 129-139.
- 21 Dick, L. K., Bernhard, A. E., Brodeur, T. J., Santo Domingo, J. W., Simpson, J. M., Walters, S.  
22 P. and Field, K. G. (2005) Host distributions of uncultivated fecal Bacteroidales bacteria

1 reveal genetic markers for fecal source identification. *Appl Environ Microbiol*, **71**, 3184-  
2 3191.

3 Dombek, P. E., Johnson, L. K., Zimmerley, S. T. and Sadowsky, M. J. (2000) Use of repetitive  
4 DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and  
5 animal sources. *Appl Environ Microbiol*, **66**, 2572-2577.

6 Genthner, F. J., James, J. B., Yates, D. F. and Friedman, S. D. (2005) Use of composite data sets  
7 for source-tracking enterococci in the water column and shoreline interstitial waters on  
8 Pensacola Beach, Florida. *Mar Pollut Bull*, **50**, 724-732.

9 Gordon, D. M., Bauer, S. and Johnson, J. R. (2002) The genetic structure of *Escherichia coli*  
10 populations in primary and secondary habitats. *Microbiology*, **148**, 1513-1522.

11 Graves, A. K., Hagedorn, C., Teetor, A., Mahal, M., Booth, A. M. and Reneau, R. B., Jr. (2002)  
12 Antibiotic resistance profiles to determine sources of fecal contamination in a rural  
13 Virginia watershed. *J Environ Qual*, **31**, 1300-1308.

14 Griffith, J. F., Weisberg, S. B. and McGee, C. D. (2003) Evaluation of microbial source tracking  
15 methods using mixed fecal sources in aqueous test samples. *J Water Health*, **1**, 141-151.

16 Hamilton, M. J., Yan, T. and Sadowsky, M. J. (2006) Development of goose- and duck-specific  
17 DNA markers to determine sources of *Escherichia coli* in waterways. *Appl Environ*  
18 *Microbiol*, **72**, 4012-4019.

19 Hartel, P. G., Summer, J. D., Hill, J. L., Collins, J. V., Entry, J. A. and Segars, W. I. (2002)  
20 Geographic variability of *Escherichia coli* ribotypes from animals in Idaho and Georgia.  
21 *J Environ Qual*, **31**, 1273-1278.

22 Hartel, P. G., Summer, J. D. and Segars, W. I. (2003) Deer diet affects ribotype diversity of  
23 *Escherichia coli* for bacterial source tracking. *Water Res*, **37**, 3263-3268.

1 Hassan, W. M., Wang, S. Y. and Ellender, R. D. (2005) Methods to increase fidelity of repetitive  
2 extragenic palindromic PCR fingerprint-based bacterial source tracking efforts. *Appl*  
3 *Environ Microbiol*, **71**, 512-518.

4 Johnson, L. K., Brown, M. B., Carruthers, E. A., Ferguson, J. A., Dombek, P. E. and Sadowsky,  
5 M. J. (2004) Sample size, library composition, and genotypic diversity among natural  
6 populations of *Escherichia coli* from different animals influence accuracy of determining  
7 sources of fecal pollution. *Appl Environ Microbiol*, **70**, 4478-4485.

8 Kuntz, R. L., Hartel, P. G., Godfrey, D. G., McDonald, J. L., Gates, K. W. and Segars, W. I.  
9 (2003) Targeted sampling protocol as prelude to bacterial source tracking with  
10 *Enterococcus fecalis*. *J Environ Qual*, **32**, 2311-2318.

11 Meays, C. L., Broersma, K., Nordin, R. and Mazumder, A. (2004) Source tracking fecal bacteria  
12 in water: a critical review of current methods. *J Environ Manage*, **73**, 71-79.

13 Mott, J. B. and Lehman, R. L. (2001) DNA Fingerprinting to Identify Sources of Bacteria in  
14 Coastal Waters of Texas, Final Report - Phase II, Coastal Coordination Council, National  
15 Oceanic and Atmospheric Administration Award No. NA870Z0251.

16 Myoda, S. P., Carson, C. A., Fuhrmann, J. J., Hahm, B. K., Hartel, P. G., Yampara-Lquise, H.,  
17 Johnson, L., Kuntz, R. L., Nakatsu, C. H., Sadowsky, M. J. and Samadpour, M. (2003)  
18 Comparison of genotypic-based microbial source tracking methods requiring a host  
19 origin database. *J Water Health*, **1**, 167-180.

20 NCCLS (2002) Performance standards for antimicrobial disk and dilution susceptibility tests for  
21 bacteria isolated from animals. *Approved Standard-2nd edition M31-A2*, **22**.

1 Parveen, S., Portier, K. M., Robinson, K., Edmiston, L. and Tamplin, M. L. (1999) Discriminant  
2 analysis of ribotype profiles of *Escherichia coli* for differentiating human and nonhuman  
3 sources of fecal pollution. *Appl Environ Microbiol*, **65**, 3142-3147.

4 Ritter, K. J., Carruthers, E., Carson, C. A., Ellender, R. D., Harwood, V. J., Kingsley, K.,  
5 Nakatsu, C., Sadowsky, M., Shear, B., West, B., Whitlock, J. E., Wiggins, B. A. and  
6 Wilbur, J. D. (2003) Assessment of statistical methods used in library-based approaches  
7 to microbial source tracking. *J Water Health*, **1**, 209-223.

8 Scott, T. M., Jenkins, T. M., Lukasik, J. and Rose, J. B. (2005) Potential use of a host associated  
9 molecular marker in *Enterococcus faecium* as an index of human fecal pollution. *Environ*  
10 *Sci Technol*, **39**, 283-287.

11 Scott, T. M., Parveen, S., Portier, K. M., Rose, J. B., Tamplin, M. L., Farrah, S. R., Koo, A. and  
12 Lukasik, J. (2003) Geographical variation in ribotype profiles of *Escherichia coli* isolates  
13 from humans, swine, poultry, beef, and dairy cattle in Florida. *Appl Environ Microbiol*,  
14 **69**, 1089-1092.

15 Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R. and Lukasik, J. (2002) Microbial source  
16 tracking: Current methodology and future directions. *Appl Environ Microbiol*, **68**, 5796-  
17 5803.

18 Simpson, J. M., Santo Domingo, J. W. and Reasoner, D. J. (2002) Microbial source tracking:  
19 state of the science. *Environ Sci Technol*, **36**, 5279-5288.

20 Stoeckel, D. M., Mathes, M. V., Hyer, K. E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T.  
21 L., Fenger, T. W., Samadpour, M., Strickler, K. M. and Wiggins, B. A. (2004)  
22 Comparison of seven protocols to identify fecal contamination sources using *Escherichia*  
23 *coli*. *Environ Sci Technol*, **38**, 6109-6117.

1 Swaminathan, B., Barrett, T. J., Hunter, S. B. and Tauxe, R. V. (2001) PulseNet: the molecular  
2 subtyping network for foodborne bacterial disease surveillance, United States. *Emerg*  
3 *Infect Dis*, **7**, 382-389.

4 USEPA (2002) *Method 1603: Escherichia coli (E. coli) in water by membrane filtration using*  
5 *modified membrane-thermotolerant Escherichia coli agar (Modified mTEC)*, Office of  
6 Research and Development, Government Printing Office, Washington, DC.

7 USEPA (2005a) *Method 1603: Escherichia coli (E. coli) in water by membrane filtration using*  
8 *modified membrane-thermotolerant Escherichia coli agar (Modified mTEC)*, Office of  
9 Research and Development, Government Printing Office, Washington, DC.

10 USEPA (2005b) *Microbial Source Tracking Guide Document*, Office of Research and  
11 Development, Cincinnati, OH.

12 Versalovic, J., Koeuth, T. and Lupski, J. R. (1991) Distribution of repetitive DNA sequences in  
13 eubacteria and application to fingerprinting of bacterial genomes. *Nucl Acids Res*, **19**,  
14 6823-6831.

**Table 1.** Summary of known source *E. coli* isolates used for library construction

Source Samples	Desired Number of Samples	Number of Samples Collected	Number of <i>E. coli</i> -positive Samples	Number of <i>E. coli</i> Isolated and Stored From Samples	Number of <i>E. coli</i> Screened by ERIC-PCR	Number of <i>E. coli</i> Identified for Library by ERIC-PCR	Number of <i>E. coli</i> -positive Samples Used for Library	Number of <i>E. coli</i> Isolates in Library (Analyzed by All Four Methods)
Domestic sewage	240	294	186	803	624	229	184	226
Pet	85	56	35	140	95	44	33	42
Cattle	150	173	150	657	440	170	130	147
Other livestock avian	20	23	21	92	59	28	19	25
Other livestock non-avian	108	115	97	413	284	112	79	89
Wildlife avian	191	195	121	559	371	163	111	145
Wildlife non-avian*	306	238	203	567	402	234	189	209
TOTALS	1100	1094	813	3231	2275	980	745	883

\*Includes 100 South Texas wildlife isolates from a previous BST study (Mott and Lehman 2001)

**Table 2.** Jackknife analysis rates of correct classification (%) for individual and four-method composite BST methods and the 883 isolate library

Source class	Random*	PFGE	ERIC-PCR	RiboPrinting	KB-ARA using best matching	KB-ARA using discriminant analysis	Four-method composite data set
Domestic sewage	26	95 (35) <sup>†</sup>	64 (29)	60 (2)	60 (3)	43 (0)	83 (15)
Pet	5	54 (69)	19 (38)	17 (0)	17 (0)	27 (0)	33 (14)
Cattle	17	80 (60)	46 (13)	43 (4)	41 (0)	27 (0)	61 (3)
Other Livestock avian	3	0 (60)	10 (20)	0 (0)	8 (4)	36 (0)	22 (8)
Other livestock non-avian	10	55 (55)	30 (20)	16 (3)	24 (2)	10 (0)	40 (8)
Wildlife Avian	16	74 (52)	37 (27)	40 (6)	35 (2)	41 (0)	48 (11)
Wildlife Non-avian	24	84 (49)	55 (17)	47 (5)	60 (0)	44 (0)	66 (11)

\*Random is the percentage of isolates from each source class represented in the library of 883 source isolates.

<sup>†</sup>The number in parentheses is the percentage of isolates for that source class left unidentified after Jackknife analyses (<85% similarity for ERIC, RiboPrinting and KB-ARA best match, <70% similarity for PFGE and the composite data set). There is not an unidentified classification or a minimum similarity in discriminant analysis.

**Table 3.** Effects of different minimum similarity cutoffs on the rates of correct classification and isolate identification using the four-method composite data set

Similarity cutoff	Percent unidentified water isolates (555)*	Domestic sewage (226)		Pet (42)		Cattle (147)		Other livestock, avian (25)		Other livestock, non-avian (89)		Wildlife, avian (145)		Wildlife, non-avian (209)	
		% RCC	% no ID <sup>†</sup>	% RCC	% no ID	% RCC	% no ID	% RCC	% no ID	% RCC	% no ID	% RCC	% no ID	% RCC	% no ID
65%	3	83	6	30	5	60	1	21	4	39	1	49	6	65	7
70%	9	83	15	33	14	61	3	22	8	40	8	48	11	66	11
75%	24	84	30	36	33	64	15	29	44	43	22	52	29	70	22
80%	56	91	48	44	57	66	46	11	64	56	54	62	48	73	41

\*Number of isolates in parentheses.

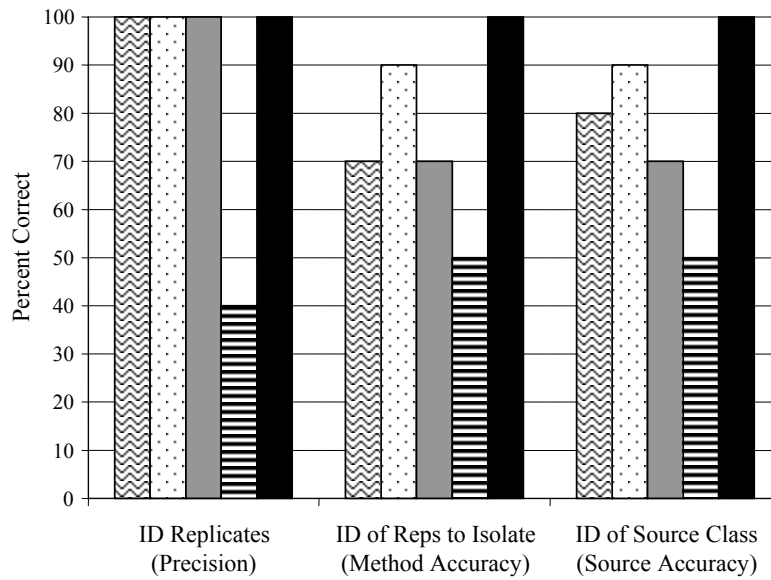
<sup>†</sup>% no ID, percentage of library isolates in the source class that were below the similarity cutoff and were therefore left unidentified after Jackknife analysis.

**Table 4.** Jackknife analysis percent rates of correct classification (RCC) for selected two-method and the four-method composites and the 883 isolate library

	Random*	ARA-PFGE	ERIC-ARA	ERIC-RP	Four-method composite data set
Domestic sewage	26	73 (5) <sup>†</sup>	71 (31)	72 (37)	83 (15)
Pet	5	20 (2)	25 (24)	23 (38)	33 (14)
Cattle	17	51 (1)	50 (7)	54 (17)	61 ( 3)
Other Livestock avian	3	10 (16)	19 (16)	0 (24)	22 ( 8)
Other livestock non-avian	10	40 (8)	24 (15)	19 (19)	40 ( 8)
Wildlife Avian	16	51 (8)	43 (23)	43 (30)	48 (11)
Wildlife Non-avian	24	63 (5)	62 (12)	62 (22)	66 (11)

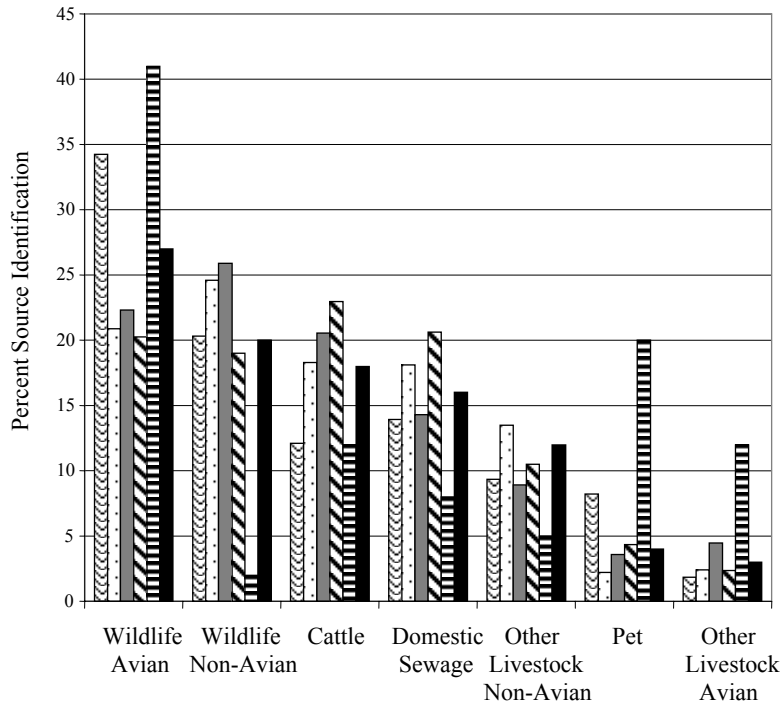
\*Random is the percentage of isolates from each source class represented in the library of 883 source isolates, i.e. library composition.







<sup>†</sup>The number in parentheses is the percentage of isolates for that source class left unidentified after Jackknife analyses (<85% similarity for ERIC-RP and ERIC-ARA, <70% similarity for ARA- PFGE and the 4-method composite data set).

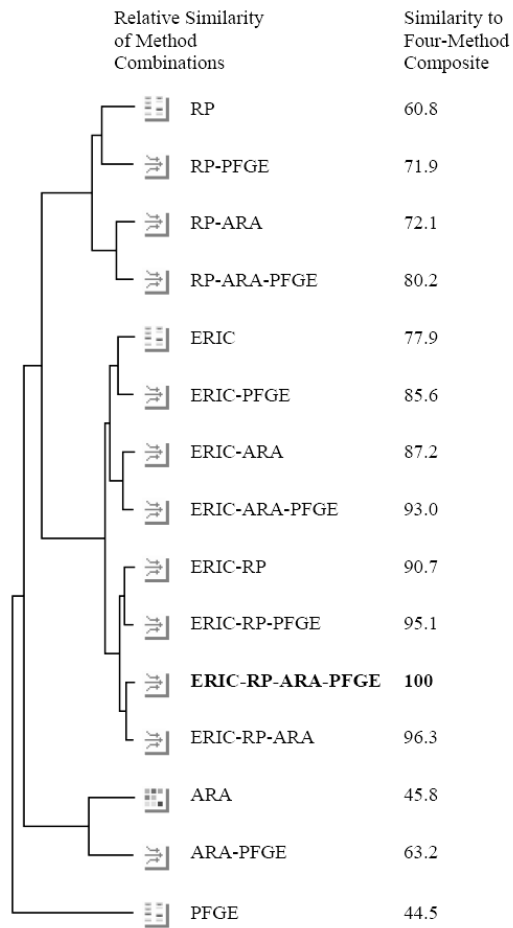


**Fig. 1.** Blind quality control study results. Comparison of ERIC-PCR (▩), RiboPrinting (▨), PFGE (■), KB-ARA (▧), and the 4-Method Composite Data Set (■).

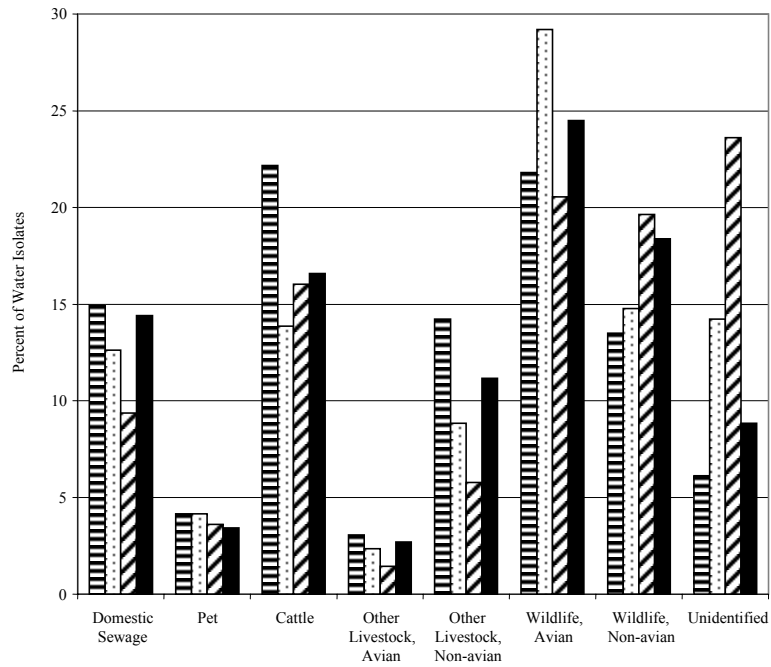




**Fig. 3.** Identification of water isolates from the Lakes Waco and Belton watersheds using the individual methods and four-method composite (555 water isolates vs. 883 known source isolates). Percent source identifications are based on the number of isolates identified to each source class out of the number of total identification attempts (shown in parentheses for each method). ERIC-PCR (438) (  ), RiboPrinting (541) (  ), PFGE (112) (  ), KB-ARA by Best Match (553) (  ), KB-ARA by Discriminant Analysis (555), (  ), 4-Method Composite (506) (  ). Isolates that were unidentified because they did not match a library isolate at the minimum similarity are not shown.



**Fig. 4.** Congruence of individual BST methods and composite data sets. Labels are as follows: ERIC, ERIC-PCR; RP, RiboPrinting; ARA, KB-ARA; ERIC-RP-ARA-PFGE, four-method composite data set. Other composite data sets are hyphenated.



**Fig. 5.** Comparison of source class identifications for the 555 water isolates using composite data sets of ARA-PFGE (▨), ERIC-ARA (⋄), ERIC-RP (▧), and the 4-method composite (■).