International Journal of
Environmental Research and Public Health
ISSN 1660-4601
www.ijerph.org
© 2005 by MDPI

Application of Oligonucleotide Microarrays for Bacterial Source Tracking of Environmental *Enterococcus sp.* Isolates.

Karl J. Indest[1*], Kelley Betts[3], and John S. Furey[2]

[1]U.S. Army Engineer Research and Development Center, Waterways Experiment Station, 3909 Halls Ferry Road, Vicksburg, MS 39180, USA, [2]CSC, 3530 Manor Drive, Vicksburg, MS 39180, [3]Analytical Services, Inc., 555 Sparkman Drive, Huntsville, Alabama 35816, USA
*Correspondence to Dr. Karl J. Indest. E-mail: indestk@wes.army.mil

Received: 15 November 2004 / Accepted: 06 February 2005 / Published: 30 April 2005

Abstract: In an effort towards adapting new and defensible methods for assessing and managing the risk posed by microbial pollution, we evaluated the utility of oligonucleotide microarrays for bacterial source tracking (BST) of environmental *Enterococcus sp.* isolates derived from various host sources. Current bacterial source tracking approaches rely on various phenotypic and genotypic methods to identify sources of bacterial contamination resulting from point or non-point pollution. For this study *Enterococcus sp.* isolates originating from deer, bovine, gull, and human sources were examined using microarrays. Isolates were subjected to Box PCR amplification and the resulting amplification products labeled with Cy5. Fluorescent-labeled templates were hybridized to in-house constructed nonamer oligonucleotide microarrays consisting of 198 probes. Microarray hybridization profiles were obtained using the ArrayPro image analysis software. Principal Components Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were compared for their ability to visually cluster microarray hybridization profiles based on the environmental source from which the *Enterococcus sp.* isolates originated. The PCA was visually superior at separating origin-specific clusters, even for as few as 3 factors. A Soft Independent Modeling (SIM) classification confirmed the PCA, resulting in zero misclassifications using 5 factors for each class. The implication of these results for the application of random oligonucleotide microarrays for BST is that, given the reproducibility issues, factor-based variable selection such as in PCA and SIM greatly outperforms dendrogram-based similarity measures such as in HCA and K-Nearest Neighbor KNN.

Keywords: bacterial sourcing tracking, *Enterococcus*, oligonucleotides microarrays, principal components analysis.

Introduction

As the number of beach closings and advisories continue to rise, so does the public's concern regarding microbial pollution in recreational waters. In a survey of more than 230 U.S. coastal and Great Lake communities, there were at least a total of 13,410 days of beach closings or advisories during 2001 [1]. The majority of beach closings and advisories were based on the presence of elevated levels of fecal contamination as measured by fecal bacterial indicators, such as *Escherichia coli* and *Enterococci*. Under section 303(d) of the 1972 Clean Water Act, states, territories, and authorized tribes are required to develop pollutant-specific lists of impaired waters and may be required to establish a total maximum daily load (TMDL) for those impaired waters [2]. TMDLs specify the maximum amount of a pollutant that a water body can receive and still meet water quality standards. Fecal coliforms are frequently listed as impairment on many states 303(d) list of associated water-quality impairments [3]. While TMDLs have historically focused on chemical impairments, more attention is now being focused on microbial impairments. Recently, the EPA published an extensive protocol for developing pathogen TDMLs [2]. Currently, there are several regional pilot projects underway aimed at establishing fecal coliform TMDLs for impacted watersheds [4].

Reducing the loads of fecal contamination can be problematic because often the pollution sources are not known or have non-point sources. Non-point sources of microbial fecal pollution are mobilized by rain/snow events and can include urban litter, agricultural runoff, failing sewer lines, malfunctioning septic systems, and domestic and wildlife excrement. Implementation of best management practices (BMPs) for TMDL compliance is dependent upon accurately identifying the source(s) of the impairment. Source tracking of non-point sources of microbial pollution, specifically indicator bacteria, has been generically referred to as bacterial source tracking (BST) [5] or microbial source tracking (MST) [6,7] and can be accomplished using a collection of multidisciplinary bacterial sub-typing methods. In addition to determining the origin of fecal contamination, BST methods can differentiate between human and non-human sources of microbial pollution [6,7], which can aid in generating more accurate risk assessments for managing the risk posed by microbial pollution.

BST methods can be divided into two general groups, 1) phenotypic or biochemical-based methods, and 2) genotypic or molecular-based methods [7]. Of the phenotypic methods, multiple antibiotic resistance (MAR) analysis has been reported the most and has been shown to be successful in 1) discriminating human and animal sources of *E. coli* or fecal streptococci [8, 9, 10] and, 2) further discriminating animal sources by animal type [11]. This method involves isolating and culturing target indicator organisms from various sources and locations to create a reference library. These isolates are subsequently replica plated on selective media containing multiple antibiotics at a range of concentrations. Antibiotic susceptibilities are characterized, subjected to discriminant analysis and compared to a reference antibiotic susceptibility library to determine identity. Reliability of the method is determined by analyzing isolates as both standards and as unknowns. The number of isolates assigned to the correct categories divided by the total number of isolates is referred to as the average rate of correct classification (ARCC) [12]. ARCC values for this method range from 62% to 94% when individuals are compared. Despite the success of this method in simple watersheds [11], some researchers have indicated that MAR lacks the sensitivity, reproducibility, and host specificity that is needed for BST [13].

In contrast to the limited number of phenotypic sub-typing methods, numerous genotypic methods have been described including ribotyping [14, 15, 16], length heterogeneity polymerase chain reaction (LH-PCR), terminal restriction fragment length polymorphism (T-RFLP) PCR [17, 18], repetitive PCR (rep-PCR) [19], denaturing gradient gel electrophoresis (DGGE) [20, 21], pulsed-field gel electrophoresis (PFGE) [22, 23, 24], and amplified fragment length polymorphism (AFLP) [25]. Most of these molecular methods rely on PCR to interrogate a fraction of the target organisms' available genetic information. PCR amplification products are subsequently resolved by gel-electrophoresis and the resulting banding pattern may be compared to a reference library to determine the identity of the organism. ARCC values can approach 100% when using some of these methods, such as rep-PCR [19]. Despite the success of genotypic methods, there is an ongoing need in BST for increased resolving power to discriminate between closely related microorganisms. Newer technologies, like DNA microarrays, which have been employed for various environmental microbiology applications [26], could potentially increase the resolving power of BST analysis [27]. For example, DNA microarrays interrogate DNA samples at the DNA sequence level. In contrast, gel-based methods rely on DNA fragment sizing; a method in which co-migration of heterogeneous DNA sequence populations of similar sized fragments is possible. Unlike gel-based methods, which rely on size fractionation of banding patterns that are subject to positional variation, DNA microarray profiles are comprised of physically immobilized, addressable spots. In addition, the resolving power of the microarray can be further improved by increasing the amount of oligonucleotide elements on the micro array.

The methods and data analysis algorithms for the application of DNA microarrays towards BST are just starting to be developed. Recently, oligonucleotide microarrays were evaluated for their ability to differentiate 25 closely related *Salmonella* isolates [27]. Previously, the same authors used a similar microarray approach to discriminate closely related *Xanthomonas* pathovars [28]. In this study, we aim to build upon these findings and further the development of oligonucleotide microarrays for use in BST. Here we report the application of a microarray, consisting of 198 oligonucleotide elements, to discriminate 17 unique environmental isolates of *Enterococcus sp.* based on the host source of the bacteria.

Materials and Methods

Bacterial Isolates

A collection of 51 *Enterococcus sp.* isolates originating from bovine, deer, gull, and human sources were provided by Dr. Shiao Wang (University of Southern Mississippi; Hattiesburg, MS). Details of the isolation and characterization of these strains have been described in detail elsewhere [29]. Isolates were routinely propagated in brain heart infusion liquid media (Becton Dickenson, San Jose, CA). High molecular weight genomic DNA for PCR analysis was obtained from each isolate using Qiagen's DNeasy Tissue Kit (Qiagen, Valencia, CA).

PCR Amplification and Labelling

PCR primer BOX A 1R 5' CTA CGG CAA GGC GAC GCT GAC G 3', was custom synthesized by Qiagen and targeted repetitive extragenic palindromic BOX sequences [19]. Primer BOX A 1R was used to amplify select portions of the *Enterococcus sp.* isolate genomes to be used as target DNAs for microarray analysis. All PCR reactions and their subsequent microarray analysis were carried out in triplicate. Final reaction conditions were as follows: 10mM Tris, pH 8.3, 50mM KCl_2, 4.5mM $MgCl_2$, 0.001 (w/v) gelatin, 0.2mM dNTP's, 2μM BOX A 1R primer, and 5U *Taq* polymerase (Promega, Madison, WI) in a final reaction volume of 100μl. A total of 100ng of genomic DNA was used as template for each reaction. Amplification was carried out in a MJ Research Tetrad thermocycler (MJ Research, Inc., Waltham, MA) programmed as follows: initial step at 95°C for 2 min followed by 35 cycles of: 94°C for 3 sec, 92°C for 30 sec, 50°C for 60 sec, 65°C for 8 min and finally cooling to 4°C at the end of the last cycle. Ten microliter portions from each reaction were electrophoresed through a 1.0% agarose gel in 1x TAE (40mM Tris-Acetate, 1mM EDTA) running buffer and stained with Sybergold (Molecular Probes, Inc., Eugene, OR) for visualization to confirm amplification. The remaining portions of each amplification reaction were ethanol precipitated with sodium acetate [30] and the resulting air-dried DNA pellets were re-suspended in 20µl Millipore water.

PCR products were aminoallyl(aa)-labeled as described previously [31]. Briefly, 3.3µl (3µg/µl) of random hexamers (Invitrogen, Carlsbad, CA) were added to each of the re-suspended PCR products and the final volume brought up to 39µl. The sample was heated to 100°C for five minutes and immediately placed in an ice bath. Twenty units of DNA polymerase I Klenow fragment (New England BioLabs, Beverly MA), 5µl of EcoPol (Klenow) buffer (New England Biolabs), and 2µl of 3mM dNTP/aa-labeling mix [100mM each dNTP, 50 mM aa-dUTP (Ambion, Austin TX)] were added to the reaction and the reaction was incubated at 37°C overnight. The reaction was stopped by adding 5µl of 0.5M EDTA. Unincorporated aa-dUTPs and free amines were removed from each reaction using the QIA quick PCR purification (Qiagen) kit with the following modifications: PE wash buffer was replaced with a 5 mM KPO_4, 80% ethanol solution and elution buffer was replaced with a 4mM KPO_4 solution. Purified PCR templates were dried down in a vacuum centrifuge and resuspended in 4.5µl of 0.1M Na_2CO_3 buffer, pH 9.3. DNA samples were labeled with a Cy5 dye by adding 4.5µl of a Cy5 mono-Reactive Dye Pack solution (Amersham Biosciences, Piscataway, NJ) and allowing the reaction to proceed in the dark at room temperature for two hours. The reaction was stopped by the addition of 35µl of 100mM NaOAc. Free dye was removed from the samples by using the QIA quick PCR purification kit (Qiagen) according to the manufacture's instructions. DNA samples were dried down and immediately processed for microarray analysis.

Microarray Oligonucleotide Probes and Fabrication

One hundred ninety eight 9mers (Table 1), with an amine-modification at the 5' end, (Sigmagenosys, Woodlands, TX) were randomly selected from a list of 102,403 9mer sequences that conform to criteria described previously [28]. Briefly, 9mer sequences had GC contents between 44-55%, could not have: 1) four nucleotide (or higher) repeats, 2) inverted repeats three nucleotides (or higher), 3) dual-terminal inverted repeats of 3 nucleotides (or higher), and 4) single-terminal inverted repeats of three nucleotides or higher. In addition to these criteria, all 9mer sequence combinations that occurred in *Enterococcus sp.* rRNA genes present in GeneBank as of 5/03 were eliminated. A Cy3-labeled control oligonucleotide, 5 'TTG GCA GAA GCT ATG AAA CGA TAT GGG 3', with an amine-modification at the 5' end, was used as a positional reference and hybridization control.

Microarrays were fabricated on aldehyde-coated glass microscope slides (Telechem International, Inc., Sunnyvale, CA) using the BioRad VersArray ChipWriter (BioRad, Hercules, CA) equipped with SMP3 Stealth microspotting pins (Telechem Internation, Inc.). Prior to fabrication, amine-modified oligonucleotides were transferred to a 384-well plate (Whatman, Clifton, NJ) and diluted to a concentration of 80 µM in 50% dimethyl sulfoxide (DMSO). Probes were printed in duplicate, using a 2-pin configuration, at a relative humidity of 60%. The resulting grid pattern and corresponding oligonucleotide probe location is illustrated in Fig. 1. After printing, slides were baked for 45 minutes at 80°C, briefly washed with 0.2% SDS, and subsequently rinsed with reagent grade water. Free aldehyde groups were chemically blocked by soaking printed slides in a fresh $NaBH_4$ solution [0.75g $NaBH_4$ (Sigma, MO), 225 ml phosphate buffered saline (pH 7.0), 66.5ml 100% ethanol] for five minutes. Following chemical blocking, printed slides were momentarily dipped 3 times in 0.2% SDS, washed for one minute in reagent grade water, and individually spun dried in 50ml Falcon conical tubes (Fisher Scientific, MO) at 700rpm for 10 minutes in a tabletop centrifuge. Microarray substrates were stored at room temperature in a desiccator.

Microarray Hybridization

Prior to hybridization, printed slides were pre-hybridized in 0.1% SDS, 4X SSC (1X SSC, 0.15M NaCl, 0.015M trisodium citrate, pH 7.0), and 10mg/ml bovine serum albumin (BSA) in 50ml Falcon conical tubes at 40°C with slight agitation for 2 hours. Pre-hybridized slides were rinsed 5 times in reagent grade distilled water and chilled to 4°C on a solid metal platform. Cy5 aminoallyl-labelled DNA targets were resuspended in 15µl of 4X SSC, heated at 95°C for 5 min, and immediately placed on ice. The Cy3 labelled oligonucleotide, 5'CCC ATA TCG TTT CAT AGC TTC TGC CA 3', was also included in the hybridization reaction (final concentration 0.6µM) as a control to hybridize with the control oligonucleotide attached to the microarray. Chilled hybridization reactions were pipetted on prechilled printed microarray slides, covered with array cover slips (PGC Scientifics, Gaitherburg, MD), and incubated overnight at 4°C as described previously [28]. Hybridized microarrays were gently rinsed in 4°C 4X SSC 5 times for 1 minute intervals followed by a final 30 second rinse in reagent grade water. Microarray slides were spun dried in 50 ml conical tubes as described above prior to scanning slides.

Image Analysis and Statistics

Processed microarray slides were scanned at 532nm and 635 nm using the VersArray Chipreader system (BioRad, Hercules, CA) configured at a 5µm resolution. Spot intensity data from the resulting 16-bit TIF images were initially extracted using the ArrayPro Analyzer software (Media Cybernetics, Silver Spring, MD). Background signal was determined locally for each spot using the "local corners" option. Individual spot intensities, minus local backgrounds, were normalized to total spot intensity for all of the spots on each micro array. The mean-normalized datasets were transformed by taking the logarithm of these values. An empirical data reduction process was employed (see Results)

C	C	1	1	33	33	65	65	97	97	129	129	161	161	2	2	34	34	66	66
98	98	130	130	162	162	3	3	35	35	67	67	99	99	131	131	163	163	4	4
36	36	68	68	100	100	132	132	164	164	5	5	37	37	69	69	101	101	133	133
165	165	6	6	38	38	70	70	102	102	134	134	166	166	7	7	39	39	71	71
103	103	135	135	167	167	8	8	40	40	72	72	104	104	136	136	168	168	9	9
41	41	73	73	105	105	137	137	169	169	10	10	42	42	74	74	106	106	138	138
170	170	11	11	43	43	75	75	107	107	139	139	171	171	12	12	44	44	76	76
108	108	140	140	172	172	13	13	45	45	77	77	109	109	141	141	173	173	14	14
46	46	78	78	110	110	142	142	174	174	15	15	47	47	79	79	111	111	143	143
175	175	16	16	48	48	80	80	112	112	144	144	176	176	193	193	195	195	197	197

C	C	17	17	49	49	81	81	113	113	145	145	177	177	18	18	50	50	82	82
114	114	146	146	178	178	19	19	51	51	83	83	115	115	147	147	179	179	20	20
52	52	84	84	116	116	148	148	180	180	21	21	53	53	85	85	117	117	149	149
181	181	22	22	54	54	86	86	118	118	150	150	182	182	23	23	55	55	87	87
119	119	151	151	183	183	24	24	56	56	88	88	120	120	152	152	184	184	25	25
57	57	89	89	121	121	153	153	185	185	26	26	58	58	90	90	122	122	154	154
186	186	27	27	59	59	91	91	123	123	155	155	187	187	28	28	60	60	92	92
124	124	156	156	188	188	29	29	61	61	93	93	125	125	157	157	189	189	30	30
62	62	94	94	126	126	158	158	190	190	31	31	63	63	95	95	127	127	159	159
191	191	32	32	64	64	96	96	128	128	160	160	192	192	194	194	196	196	198	198

Figure 1: Configuration of the printed microarray spots and the physical location of the corresponding oligonucleotide probes as referenced in Table 1. Control oligonucleotide designated by C.

to identify which of the 198 probe spots had the most information (example: spots that were always “on” or “off” for all isolates would have no information for this dataset) and which of the spots that were too variable within the replicates of the same isolates. Principal Components Analyses (PCA) and cluster and classification analyses were run on the remaining dataset using Pirouette (Infometrix, Inc., Bothell, WA).

Results

Oligonucleotide Microarray Bacterial Source Tracking

Oligonucleotide microarrays were evaluated for their ability to resolve BOX PCR amplification products derived from environmental sources of *Enterococcus sp.* isolates originating from deer, bovine, gull, and human. Purified genomic DNA from *Enterococcus sp.* isolates was subjected to BOX PCR amplification and the resulting amplification products were visualized by agarose gel electrophoresis. The results of a typical experiment can be seen in Fig. 2, which represents the subset of samples originating from deer. Agar gel electrophoresis confirms amplification as well as consistency of the BOX PCR reaction. PCR products were fluorescently labelled with aminoallyl dUTP and Cy5 then resolved by hybridization to in house fabricated 9mer oligonucleotide microarrays (see Material & Methods). The results of a representative microarray experiment can be seen in Fig. 3, in which replicate BOX PCR reactions from *Enterococcus sp.* deer isolate 49.1.1 were hybridized to replicate oligonucleotide micro arrays. A histogram of fluorescent spot intensities indicates that these randomly selected nonamer intensities follow a lognormal distribution (data not shown). Of the 17 environmental isolates analysed, not all replicate microarrays were usable. For six of these isolates (4 human and 2 deer) a single microarray hybridization replicate, consisting of duplicate microarray spots, was available for analysis. For the remaining 11 isolates and their replicates, spots that exhibited extreme variability in normalized spot intensities among replicates within a specific source were identified and subsequently eliminated from

Table 1: Microarray Oligonucleotide Probes

ID No.	*Sequences (5'-3')*	*ID No.*	*Sequences (5'-3')*	*ID No.*	*Sequences (5'-3')*	*ID No.*	*Sequences (5'-3')*
1	AAATACCCG	51	GGATAGCGA	101	TGGCTACGT	151	GAGGAGATA
2	CAAATACCC	52	TATTGGTCG	102	GGGCTGAAT	152	TAGCGAGTG
3	AATTGCCCT	53	AAGCAGCAG	103	TGGCTCGAA	153	GAGTTTCAG
4	GGGCCATTT	54	CAGACACGA	104	GGCCCATAT	154	TAGTCGTCT
5	GACGAGCTT	55	AAAGTGCCC	105	TGCGTACAT	155	GAGAGAAAC
6	AGCAGATAG	56	CAATCGTTC	106	GGCTCAAGA	156	TAGCATAGG
7	CTTTCCAGG	57	AATCCGTAG	107	TGCCCAAGA	157	GACTCTACG
8	ATGACAGAC	58	CAAGAGGGT	108	GGTTCTGTA	158	TACGGTTCT
9	TGAGAGGCT	59	AATGGAACC	109	TGCAGAACG	159	GACAGTTCA
10	GGTAGTGCT	60	CATATCCTC	110	GGTTTGTGT	160	TACCCAGTT
11	CATTGTCCG	61	AACTTGCCG	111	TGCAAGTTC	161	GATGATACC
12	ATCTCTTGC	62	CATCTTGAC	112	GGTAGTTTC	163	GAAGGAAAG
13	CTACCAAGG	63	AAGACAGTG	113	TGTCTATCG	164	TAAGCCGCA
14	AACACTACC	64	CACTACGCA	114	GGACCTAAC	165	GAACTAAGC
15	CCATAATCC	65	AAGGGATGA	115	TGAGGATAG	166	AGGCTGTTC
16	GAACTGGCA	66	CACGAATCC	116	GGAAATCTG	167	CGGTCAGAT
17	CAAATCTGG	67	ATATCACGG	117	TGATGAGAC	168	AGGTAGGAA
18	GCGATGTTG	68	CAGATGACC	118	GCGATATTC	169	CGCCTATGT
19	AGAGAAGCC	69	ATAGTCCAG	119	GCCAATGTT	170	AGCCGTACA
20	TCAGCGCAT	70	CAGCAGATG	120	TCGCCCTTA	171	CGTGTTCTC
21	GCAACCAAA	71	ATTCACACC	121	TCGTTATGG	172	AGCTATGCG
22	CTTGATTCC	72	CAGGTGTGT	122	GCTTCCGTT	173	CGTGGTTAT
23	TACCCACTG	73	ATTGGTGGG	123	TCCGAGACT	174	AGCAAGTGT
24	TTACACCGC	74	CTATACGCA	124	GCTTGTGAT	175	CGTCTAACC
25	CTGCGATCA	75	ATCAGGGAA	125	TCCGTCAAG	176	AGTCTCAAG
26	GAGCTGTCA	76	CTACACGCA	126	GCTACCTTC	177	CGAAGTTTG
27	TGGGCGTTT	77	ATCGAGCCT	127	TCCTTGGTT	179	CGACTGGAA
28	GGGCGTTTA	78	CTTATAGGG	128	GCACTCTAA	180	AGTTACCCT
29	CATCTGTCG	79	ATGTCAAGG	129	TCCATCGTG	181	CGAAACAGG
30	AAGTAGCCC	80	CTTCCATAC	130	GCATGTAGG	182	AGAGTTCGA
31	AATATGCGG	81	ATGCCGGTT	131	TCTCGTACC	183	CCGTGGAAA
32	GTACGGAGT	82	CTTGGAACC	132	GTGGGCATT	184	CACAACTCT
33	TCTGCTATG	83	ATGGACACC	133	GCAAAGCCT	185	AACGAAACG
34	CAAATGTCC	84	CTCATAGGT	134	TCTTCCTAC	186	AACACGCTT
35	AAATCTCGC	85	ATGGGTACG	135	GTGCTGGAT	187	CATGAGGTT
36	AATTTCGGC	86	CTCTGTTCC	136	TCTACCCAC	188	CATAGCGAA
37	ACTCTCCCT	87	ACATGACAG	137	GTGTAGAAC	189	CAAACGAGG
38	CCAAGTTCT	88	CTCCTTTGC	138	TCAGCATAG	190	AAACACGTC
39	GAAAGAGCA	89	ACACACCAT	139	GTGAGGTTC	191	CTGCTACGA
40	CCCTTTCCA	90	CTCGATCAC	140	TCAACCTTC	192	ACTAGCGGT
41	ACCTATGCG	91	ACAGTCTCA	141	GTCACGTTA	193	ACAACACTC
42	TTGGGTTCG	92	CTGAGTACA	142	TTGAGCTGA	194	CTATGTCGG
43	ATACCGATG	93	ACTAAGCGC	143	GTTTCGTGT	195	CTATCAACC
44	TGCTTCACA	94	CTGTAGACC	144	TTCGCACTC	196	AACGATACC
45	ACGCTACGA	95	ACTTAGCCA	145	GTTAGGGTG	197	CAAACGGGA
46	TACTGTCGG	96	CTGCTACAC	146	TTCTAGCGC	198	CTGTCACTG
47	GCTGCTACA	97	ACTTCGTCG	147	GTATCGCTA		
48	TCCAACTAG	98	CTGCTGTGT	148	TTAGCGTGC		
49	CCGCAAAGT	99	ACTCTCTCT	149	GTAACTGTC		
50	GATTAGCGC	100	CTGGCTTCT	150	TTACCTGGC		

5’ amine-modified 9mer oligonuceotide microarray probes and corresponding I.D. numbers analysis for all isolates. Normalized spot intensities with above median standard deviations > or = 0.7 within source-specific datasets (i.e. bovine, deer, etc.), were eliminated leaving 45 of the 200 probes. The remaining 45 probes were then used for analyzing all 17 isolates.

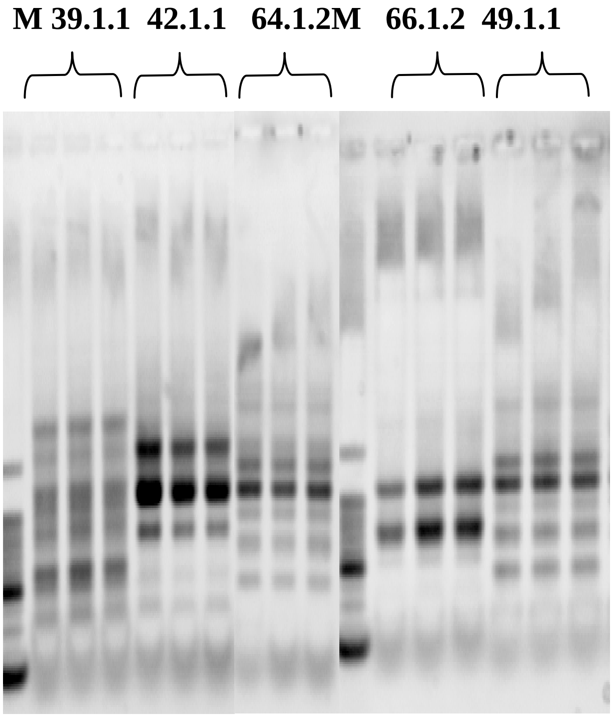

Figure 2: BOX-PCR agarose gel fingerprints run in triplicate from *Enterococcus sp.* isolates originating from deer. A *HindIII* digested Lambda marker was included in the gel run as a size standard

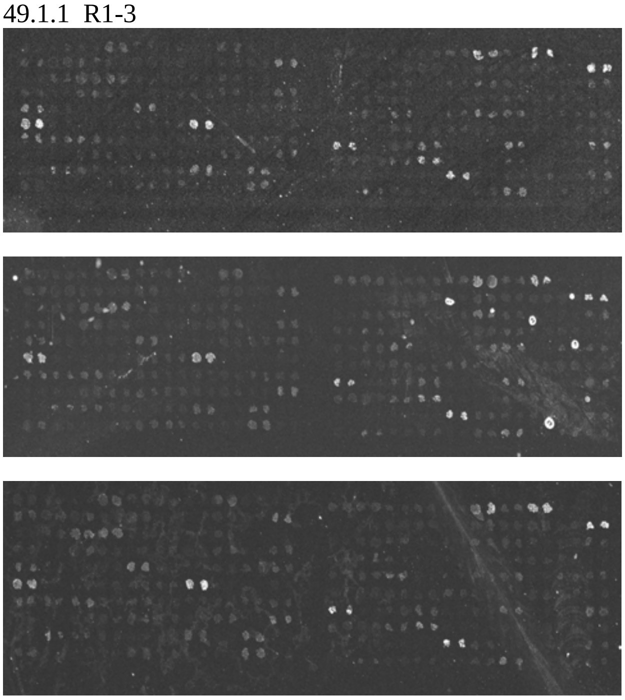

Figure 3: Oligonucleotide microarray replicate hybridization profiles resulting from hybridization with BOX-PCR amplification products from deer isolate 49.1.1.

PCA and HCA Analysis

The dendrogram of a complete Euclidean distance Hierarchical Cluster Analysis (HCA) did not project good origin-specific clustering of the isolates. In particular, the bovine-origin replicates were spread among several clusters (example part of dendrogram Fig. 4). A K-Nearest Neighbour classification confirmed the HCA, misclassifying 8% of the deer, 16% of the human, and 50% of the gull isolates as bovine isolates. The PCA was visually superior at separating origin-specific clusters, even for as few as 3 factors (Fig. 5). A Soft Independent Modelling (SIM) classification confirmed the PCA, resulting in zero misclassifications using 5 factors for each class. Numerical descriptions of the SIM classification model for bovine-origin *Enterococcus sp.* are presented in Table II. These factors describe the multidimensional subspace within the PCA projection in which the various microarray source profiles exist. Factor numbers indicate the relative linear weights of each probe in each factor. For instance probes 2 and 16 have the highest weights for the most important factor, Factor 1, which accounts for 30% of the variability. Thus for this set of isolates, SIM classifications based on 5 factors for each class and 5 linear combinations of the 45 probes sufficed to distinguish the origins of *Enterococcus sp.* isolates.

Discussion

In an effort towards adapting new defensible methods for assessing and managing the risk posed by microbial pollution, we evaluated the utility of oligonucleotide microarrays for bacterial source tracking. Specifically, we evaluated the ability of oligonucleotide microarrays to visually discriminate 17 unique environmental isolates of *Enterococcus sp.* based on host origin, i.e. gull, bovine, deer, and human. As observed in an earlier study by Kingsley et al. [28], many of the microarray oligonucleotide probes exhibited high variations in fluorescent spot intensities within a series of replicates. A strong down selection for reproducible spot intensities within replicates produced a set of 45 probes, and this reduced set proved useful for classifying isolates by source. It should be reiterated that this data reduction was performed in order to improve reproducibility, and had the side effect of improving the classification fit. This is the opposite of the familiar problem of model over fitting, in which the addition of extra variables improves classification at the expense of robustness and reproducibility.

Following data reduction, a number of multivariate statistical analysis procedures are available for evaluating the relationships among microarray hybridization profiles. Previously, PCA was successfully used to visualize relationships among microarray hybridization profiles derived from closely related *Xanthomonas* pathovars [28]. In this study, PCA and HCA were compared for their ability to visually cluster microarray hybridization profiles based on the environmental source from which the *Enterococcus sp.* isolate originated. Classification of *Enterococcus sp.* isolates by source using a Soft Independent Modelling of

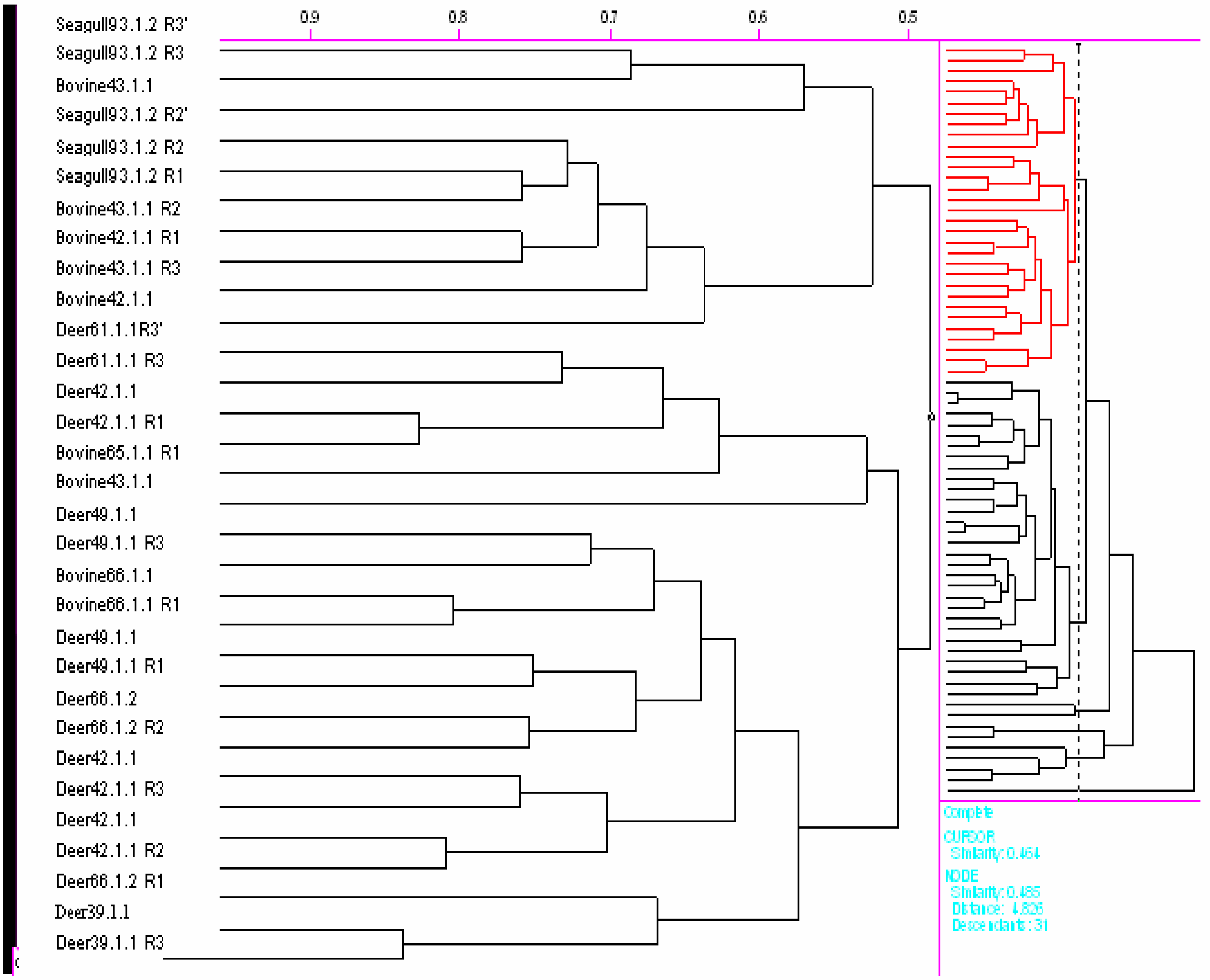

Figure 4: Hierarchical Cluster Analysis of normalized microarray spot intensities of replicates of 17 environmental isolates of *Enterococcus sp*. The dendrogram does not show good clustering by host origin at reasonable similarities. The bovine-origin replicates were most spread.

class analogies consisting of 5 factors was more accurate than classification based on K-Nearest Neighbour calculations. This difference is apparent when comparing the PCA, which is a visualization of some of the SIM calculations, to the HCA, which is a visualization of some of the KNN calculations. The implication of these results for the application of random oligonucleotide microarrays for BST is that, given the reproducibility issues, factor-based variable selection such as in PCA and SIM greatly outperforms dendrogram-based similarity measures such as in HCA and KNN. Given any sample based strictly on the microarray intensity values, the SIM model outputs the best fitting class for that sample, with zero misclassifications for the dataset.

Further optimization of source classifications may result from the application of information theory to detect patterns in microarray profiles. In particular, bacterial source tracking may benefit from several measures of classification utility, such as those based on mutual information that have been developed as part of information theory [32]. However, successful application of information theory for microarray analysis will be dependant upon accurately understanding, capturing, and modelling sources of variation in the microarray experimental process. Some of these sources of variation, such as PCR amplification and microarray fabrication have been described previously [27]. Once improved microarray experimental protocols and statistical methods have been developed, it will be possible to incorporate microarray technology into the growing toolbox of technologies that is rapidly defining bacterial source tracking. While there is currently no one best method that accomplishes the ambitious goal of source tracking as demonstrated in the latest study by Stoeckel et al. [33], it is likely that a combination of methods will lead to effective source tracking.

Acknowledgements: This work was funded by the Long-term Effects of Dredging Operations (LEDO) program at the U. S. Army Engineer Research and Development Center (ERDC). Permission was granted by the Chief of Engineers to publish this information.

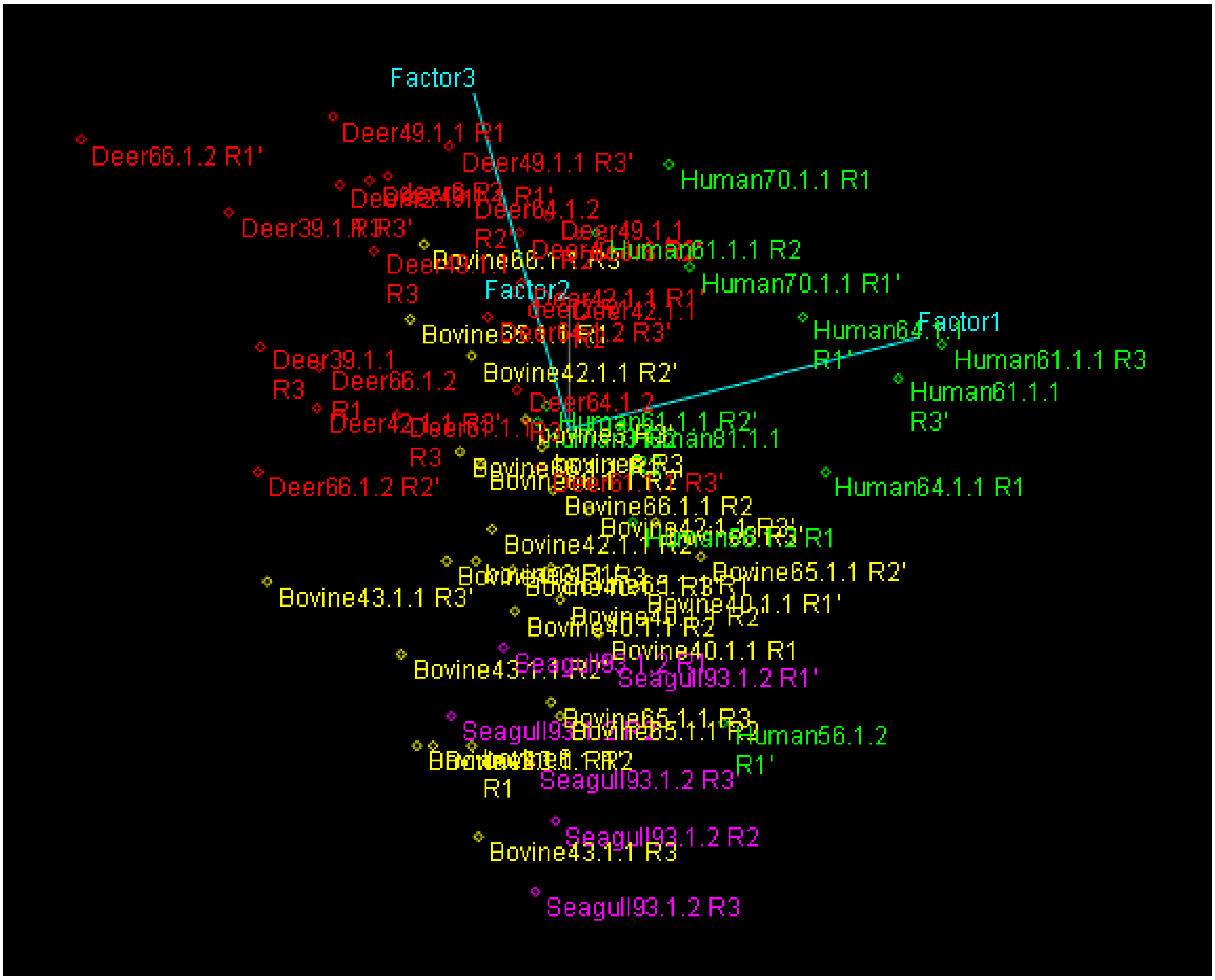

Figure 5: Principal Components Analysis of normalized microarray spot intensities of replicates of 17 environmental isolates of *Enterococcus sp.*, colored by host origin: deer is red, bovine is yellow, human is green, gull is purple. For this 3D view only the first 3 components can be plotted, but clustering is evident.

Table 2: The 5-factor oligonucleotide microarray SIM classification model for bovine-origin *Enterococcus*

probe I.D.#	*Factor1*	*Factor2*	*Factor3*	*Factor4*	*Factor5*
10	-0.0693	-0.0124	-0.1724	0.3365	0.1906
101	-0.2427	-0.0110	0.1314	-0.0107	-0.0573
103	-0.0893	-0.0051	0.1625	0.0274	0.0169
109	0.0946	-0.1203	0.0935	-0.2996	0.2562
110	0.1979	0.1221	0.0276	0.1051	-0.0005
116	0.1914	0.0207	0.1926	0.0159	0.1193
120	-0.0518	0.0044	0.2388	0.1803	0.1916
129	-0.0098	-0.1774	-0.2056	0.1891	0.0237
135	0.2337	-0.0282	-0.0125	0.1349	0.1671
139	-0.0159	0.1796	-0.0037	-0.0411	0.2386
143	-0.0029	-0.1737	-0.0814	-0.1052	-0.2598
148	-0.1357	0.0204	0.2510	0.1294	0.0086
151	0.0471	-0.0152	-0.0363	0.3958	0.1300
152	-0.0195	-0.2508	0.1577	0.0328	-0.1484
156	-0.1935	-0.0510	0.1517	0.0772	-0.0967
16	0.2880	-0.0279	0.0481	0.0090	0.0807
163	-0.1135	-0.1457	0.1707	0.2306	-0.0090
164	-0.0643	0.1756	0.1276	-0.2373	0.1324
173	-0.0276	-0.0829	-0.1689	-0.1325	0.2091
179	0.1752	0.1325	0.2503	0.0267	0.0454
183	-0.1061	-0.0186	0.1716	0.3182	0.1273
188	0.0372	-0.1136	0.2749	-0.0896	-0.1253
197	0.1127	0.1055	0.1382	0.0709	-0.2228
2	0.3161	-0.1013	-0.0028	0.0365	-0.0345
23	0.0120	0.0386	0.1942	-0.2457	0.0996
24	0.0043	-0.1837	0.1502	-0.0565	-0.1489
27	0.0415	0.2624	0.0418	0.1381	-0.0770
3	-0.1492	-0.1004	0.1142	0.0727	-0.1953
31	-0.1750	-0.0222	0.1739	-0.0174	-0.0754
39	0.0871	-0.2652	0.0902	0.1966	0.0277
42	-0.0778	-0.0590	0.1760	-0.0517	0.2568
43	0.0425	-0.2962	-0.0494	0.1284	0.0829
51	0.2070	-0.0761	0.0827	0.1584	-0.0281
52	0.1503	0.2894	-0.0144	0.0820	-0.1347
54	-0.1305	0.1739	0.0558	0.0927	0.2616
61	-0.0134	0.1594	0.3046	-0.0007	0.0549
63	0.2623	-0.1179	0.0312	0.0517	0.1014
65	0.2654	0.1653	0.0571	-0.0874	-0.0411
67	0.1681	-0.2348	-0.0132	-0.1448	0.1331
7	0.2416	-0.2058	0.0358	-0.0196	-0.0125
72	-0.0466	0.1237	-0.3268	0.0042	0.1066
74	0.1058	0.2964	0.0585	0.1051	-0.1957
76	-0.0709	-0.1606	0.0171	-0.1116	-0.1814
85	0.2414	0.0085	0.1413	-0.0517	-0.0178
97	-0.1241	-0.0890	0.1126	-0.1321	0.3621

References

1. Natural Resource Defence Council (NRDC): Testing the Waters 2001. *Natural Resources Defense Council, New York, NY.* http://www.nrdc.org/water/oceans/ttw/titinx.asp, **2002**.
2. USEPA: Protocol for Developing Pathogen TMDLs. U. S. Environmental Protection Agency, *Washington DC, EPA 841-R-00-00*, **2001b**.
3. For list of impairment by state visit http://www.epa.gov/OWOW/tmdl/
4. Mostaghimi, S.; Brannan, K. M.; Dillaha, T. A.: Fecal Colifrom TMDL Development: Case Study and Ramifications. *Water Resources Update,* **2002**, *122 (March 2002),* 27-33.
5. Term first used by Hagedorn and Wiggins at http://www.bsi.vt.edu/biol_4684/BST/BST.html
6. Scott, T. M.; Rose, J. B.; Jenkins, T. M.; Farrah, S. R.; Lukasik, J.: Microbial Source Tracking: Current Methodology and Future Directions. *Appl. Environ. Microbiol.*, **2002**, *68*, 5796-5803.
7. Simpson, J. M.; Santo Domingo, J. W.; Reasoner, D. J.: Microbial Source Tracking: State of the Science. *Environ. Sci. Technol.*, *36*, 5279-5288.
8. Harwood, V. J.; Whitlock, J.; Withington, V.: Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Appl. Environ. Microbiol.*, **2000**, *66,* 3698-3704.
9. Wiggins, B. A.: 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.*, **1996**, *62*, 3997-4002.
10. Wiggins, B. A.; Andrews, R. W.; Conway, R. A.; Corr, C. L.; Dobratz, E. J.; Dougherty, D. P.; Eppard, J. R.; Knupp, S. R.; Limjoco, M. C.; Mettenburg, J. M.; Rinehardt, J. M.; Sonsino, J.; Torrijos, R. L.; Zimmerman, M. E.: Use of antibiotic resistance analysis to identify nonpoint sources of fecal pollution. *Appl. Environ. Microbiol.*, **1999**, *65*, 3483-3486.
11. Hagedorn, C. S.; Robinson, S. L.; Filtz, J. R.; Grubbs, S. M.; Angier, T. A.; Reneau, R. B.: Using antibiotic resistance patterns in the fecal streptococci to determine sources of fecal pollution in a rural Virginia watershed. *Appl. Environ. Microbiol.*, **1999**, *65*, 5522-5531.
12. Hardwood, V. J.: Lessons learned and questions unanswered from 5 years of Bacterial Source Tracking. *Marriott Hotel and Conference Center, Irvine, CA. February 5,* **2002**.
13. Samadpour, M.: Microbial Source Tracking: Principles and Practice. U. S. EPA Workshop on Microbial Source Tracking. *Marriott Hotel and Conference Center, Irvine, CA,. February 5,* **2002**.
14. Carson, C. A.; Shear, B. L.; Ellersieck, M. R.; Asfaw, A.: Identification of fecal *Escherichia coli* from humans and animals by ribotyping. *Appl. Environ. Microbiol.,* **2001**, *67*, 1503-1507.
15. Hartel, P. G.; Summer, J. D.; Hill, J. L.; Collins, J. V.; Entry, J. A.; Segars, W. I.: Geographic variability of *Escherichia coli* ribotypes from animals in Idaho and Georgia. *J. Environ. Qual.*, **2002**, *31,* 1273-1278.
16. Parveen, S.; Portier, K. M.; Robinson, K.; Edmiston, L.; Tamplin, M. L.: Discriminant analysis of ribotype profiles of *Escherichia coli* for differentiating human and nonhuman sources of fecal pollution. *Appl. Environ. Microbiol.*, **1999**, *65*, 3142-3147.
17. Bernhard, A. E.; Field, K. G.: A PCR assay to discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA. *Appl. Environ. Microbiol.* **2000**, *66*, 4571-4574.
18. Bernhard, A. E.; Field, K. G.: Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. *Appl. Environ. Microbiol.*, **2000**, *66*, 1587-1594.
19. Dombek, P. E.; Johnson, L. K.; Zimmerley, S. T.; Sadowsky, M. J.: Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Appl. Environ. Microbiol.*, **2000**, *66,* 2572-2577.
20. Buchan, A.; Alber, M.; Hodson, R. E.: Strain-specific differentiation of environmental Escherichia coli isolates via denaturing gradient gel electrophoresis (DGGE) analysis of the 16S-23S intergenic spacer region. *FEMS. Microbiol. Ecol.*, **2001**, *35*, 313-321.
21. Farnleitner, A. H.; Kreuzinger, N.; Kavka, G. G.; Grillenberger, S.; Rath, J.; Mach, R. L.: Simultaneous detection and differentiation of *Escherichia coli* populations from environmental freshwaters by means of sequence variations in a fragment of the beta-D-glucuronidase gene. *Appl. Environ. Microbiol.*, **2000**, *66*, 1340-1346.
22. Kariuki, S.; Gilks, C.; Kimari, J.; Obanda, A.; Muyodi, J.; Waiyaki, P.; Hart, C. A.: Genotype analysis of *Escherichia coli* strains isolated from children and chickens living in close contact. *Appl. Environ. Microbiol.*, **1999**, *65*, 472-476.
23. Dicuonzo, G.; Gherardi; G., Lorino, G.; Angeletti, S.; Battistoni, F.; Bertuccini, L.; Creti, R.; Di Rosa, R.; Venditti, M.; Baldassarri, L.: Antibiotic resistance and genotypic characterization by PFGE of clinical and environmental isolates of enterococci. *FEMS. Microbiol. Lett.*, **2001**, *201*, 205-211.
24. Parveen, S.; Hodge, N. C.; Stall, R. E.; Farrah, S. R.; Tamplin, M. L.: Phenotypic and genotypic characterization of human and nonhuman *Escherichia coli*. *Water Res.*, **2001**, *35*, 379-386.
25. Guan, S.; Xu, R.; Chen, S.; Odumeru, J.;Gyles, C.: Development of a procedure for discriminating among *Escherichia coli* isolates from animal and human sources. *Appl. Environ. Microbiol.*, **2002**, *68*, 2690-2698.
26. Sharkey, F. H.; Banat, I.; Marchant, R.: Detection and Quantification of gene expression in

environmental bacteriology. *Appl. Environ. Microbiol.*, **2004**, *70*, 3795-3806.

27. Willse, A.; Straub, T. M.; Wunschel, S. C.; Small, J. A.; Call, D. R.;Daly, D. S.; Handler, D. P.: Quantitative oligonucleotide microarray fingerprinting of *Salmonella enterica* isolates. *Nucleic Acids Res.*, **2004**, *32*, 1848-1856.
28. Kingsley, M. T.; Straub, T. M.; Call, D. R.; Daly, D. S.; Wunschel, S. C.; Chandler, D. P.: Fingerprinting closely related *xanthomonas* pathovars with random nonamer oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **2002**, *68*, 6361-6370.
29. Hassan, W. M; Wang, S. Y.; Ellender, R. D.: Methods to increase the fidelity of rep-PCR fingerprint-based bacterial source tracking efforts. *Appl. Environ. Microbiol.*, In Press.
30. Sambrook, J.; Fritsch, E. F.; Maniatis, T.: Molecular cloning: a laboratory manual, 2nd ed. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.*
31. Gilbert, J.; Hasseman, J.; Cline, R.: Microbial genomic DNA aminoallyl labelling for microarrays. The Institute for Genomic Research Standard Operating Procedure #M009, ver 0.3. http://pfgrc.tigr.org/protocols.shtml
32. Butte, A. J.; Kohane, I. S.: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **2000**, 418-429.
33. Stoeckel, D. M.; Mathes, M. V.; Hyer, K. E.; Hagedorm, C.; Kator, H.; Lukasik, J.; O'Brien, T. L.; Fenger, T. W.; Samadpour, M.; Stickler, K. M.; Wiggens, B. A.: Comparison of seven protocols to identify fecal contamination sources using *Escherichia coli*. *Environ. Sci. Technol.*, **2004**, *38*, 6109-6117.