Expansion and Evaluation of the Texas Bacterial Source Tracking Program (FY18-FY19)

Texas Water Resources Institute TR-530
May 2020

Texas Water Resources Institute
make every drop count

Anna Gitter, Lucas Gregory, Brian Hux, John Boswell, Terry Gentry, Carlos Monserrat, Elizabeth Casarez, Kristina Mena
Texas Water Resources Institute

Expansion and Evaluation of the Texas Bacterial Source Tracking Program (FY18-FY19)

Texas' Bacterial Source Tracking Program

State Nonpoint Source Grant Program

TSSWCB Project 18-50

Prepared for:

Texas State Soil and Water Conservation Board

Prepared by:

Anna Gitter [1], Lucas F. Gregory [1], Brian Hux [2], John Boswell [2], Terry J. Gentry [2], Carlos Monserrat [3], Elizabeth A. Casarez [3], Kristina D. Mena [3]

[1] Texas Water Resources Institute

[2] Texas A&M Agrilife Research – Department of Soil and Crop Sciences

[3] University of Texas Health Science Center at Houston School of Public Health, El Paso Campus

Texas Water Resources Institute TR-530

May 2020

Table of Contents

List of Tables

List of Figures

List of Acronyms

AgriLife SCSC	Texas A&M AgriLife Research, Department of Soil and Crop Sciences
ARCC	Average rate of correct classification
BMP	Best management practice
BST	Bacterial source tracking
CFU	Colony forming units
ddPCR	droplet digital PCR
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EPA	Environmental Protection Agency
ERIC-PCR	Enterobacterial repetitive intergenic consensus sequence
PCR MDS	Multi-dimensional scaling
MAF70	Mission and Aransas 70 isolate local source library
MPN	Most probable number
mTEC	Membrane Thermotolerant *Escherichia coli*
NA-MUG	Nutrient agar- 4-methylumbelliferyl- β -D-glucuronide
OSSF	On-site sewage facility
PCR	Polymerase chain reaction
QC	Quality Control
qPCR	Quantitative PCR
RARCC	Random average rate of correct classification based on library composition
RCC	Rate of correct classification
RP	RiboPrinting
rRNA	Ribosomal ribonucleic acid
SARA	San Antonio River Authority
SV	Self-validated
SVU	Singleton
TCEQ	Texas Commission on Environmental Quality
TSSWCB	Texas State Soil and Water Conservation Board
TWRI	Texas Water Resources Institute
UHCL	University of Houston-Clear Lake
UTSPH EP	University of Texas Health Science Center at Houston School of Public Health, El Paso Campus, Environmental Microbiology Laboratory
WWTF	Wastewater treatment facility

Executive Summary

The 2018 Texas Integrated Report of Surface Water Quality (TCEQ 2018) identified 250 water bodies as being impaired due to excessive bacteria in Texas. To identify bacterial sources and help address these impairments, Texas established the Bacterial Source Tracking (BST) Program in 2006. To support the maintenance, expansion and use of the Texas BST Library and other BST tools, the Texas Water Resources Institute (TWRI), University of Texas Health Science Center at Houston School of Public Health, El Paso Campus, Environmental Microbiology Laboratory and the Texas A&M AgriLife Research, Department of Soil and Crop Sciences collaborated with the Texas State Soil and Water Conservation Board in fiscal years 2018 and 2019 to:

(1) Expand the Texas *Escherichia coli* (*E. coli*) BST Library through known source sample collection in the Mission and Aransas rivers' watersheds.
(2) Support BST efforts in the Mission and Aransas rivers' watersheds.
(3) Evaluate and refine the Texas *E. coli* BST Library by assessing geographic and temporal stability, composition, average rates of correct classification, diversity of source isolates of the updated library, and working to develop/refine source-specific polymerase chain reaction (PCR) markers.
(4) Provide outreach regarding BST.

Major findings from the project include:

- The Texas *E. coli* BST Library was expanded and refined, with the current version now containing 1,912 isolates from 1,653 known source fecal samples retrieved from 4,301 individual known source samples in over 20 watersheds. An additional 30 isolates from the Mission and Aransas rivers were added to the BST Library.
- BST analysis in the Mission and Aransas watersheds indicate that wildlife (non-avian and avian) are the leading contributors of *E. coli* in the two individual watersheds, followed by domestic animals and humans.
- Analysis of the Texas *E. coli* BST Library and quantitative PCR (qPCR) markers identified: 1) the need for continued evaluation of geographic impacts on source identification as the statewide library continues to expand and 2) potential application of new human-specific qPCR markers for future BST projects in Texas.
- Outreach of the BST Program resulted in:
    - Three conferences and five meetings where BST Program results were shared with the public.
    - The Texas BST Program website was updated as part of the TWRI's overall website redesign.
    - The BST Program website that resulted in 385 visits.

Introduction

Bacteria impairments make up the majority of impairments of water bodies across the state. The *2018 Texas Integrated Report and 303(d) List* indicates that of the 1,071 water bodies assessed, 574 are impaired. Of those 574 impairments, 237 are impaired for bacteria or roughly 39% of total impairments. Identifying and assessing sources of these bacteria is critical to target best management practices (BMPs), develop bacterial total maximum daily loads or watershed protection plans and assess risks from contact recreation.

BST is a valuable tool that can identify and rule-out significant sources of *E. coli* (fecal) pollution in a watershed. The premise behind BST is that genetic and phenotypic tests can identify bacterial strains that are host-specific, which allow the original host species and source of the fecal contamination to be identified. Numerous BST methods are available that use deoxyribonucleic acid (DNA) fingerprints and bacterial markers to identify fecal pollution sources. Based on a multi-year study initiated in 2002, the State of Texas selected the two-method approach using Enterobacterial repetitive intergenic consensus sequence (ERIC-PCR) and RiboPrinting (ERIC-RP), as this approach was found to be the most accurate and cost-effective. *E. coli* is used as the target bacterium because it provides a direct link with water quality standards.

For more than a decade, the Texas BST Program has successfully identified sources of *E. coli* in dozens of watersheds across Texas. Comprehensive BST has been completed by UTSPH EP and the Texas A&M AgriLife Research, Department of Soil and Crop Sciences (AgriLife SCSC) for the following watersheds: (1) Lake Waco and Belton Lake, (2) San Antonio area, (3) Lake Granbury, (4) Buck Creek, (5) Leon and Lampasas rivers, (6) Little Brazos River tributaries, (7) Big Cypress Creek, (8) Leona River, (9) Attoyac Bayou, (10) Arroyo Colorado, (11) Navasota River, (12) Big Elm Creek, (13) Plum Creek and (14) the Trinity River in Tarrant Regional Water District's service area. A Texas *E. coli* BST Library has been developed based on known source isolates from these and other (i.e. Upper Trinity River and Upper Oyster Creek) watersheds.

The Texas *E. coli* BST Library is dynamic, with new isolates being added with each successive BST project. To support maintenance, expansion and use of the library and other BST tools, Texas Water Resources Institute (TWRI), University of Texas Health Science Center at Houston School of Public Health, El Paso Campus, Environmental Microbiology Laboratory (UTSPH EP) and AgriLife SCSC collaborated to:

(1) further evaluate and refine the Texas *E. coli* BST library by assessing geographic and temporal stability, composition, average rates of correct classification (ARCC), diversity of source isolates of the updated library, and working to develop/refine source-specific Polymerase chain reaction (PCR) markers;
(2) support BST efforts in high priority watersheds; and
(3) provide outreach regarding BST.

Expansion of the Texas *E. coli* BST Library

The Texas *E. coli* BST Library is a key component of the Texas BST Program, successfully identifying sources of *E. coli* in more than a dozen watersheds across Texas over the past decade. The Texas *E. coli* BST Library is dynamic, with new isolates being added with each successive BST project. In an effort to expand the Texas *E. coli* BST Library and support BST analyses in the Aransas and Mission rivers' watersheds, a goal of collecting approximately 50 known source

fecal samples, from which 75 *E. coli* isolates would be fingerprinted for potential addition to the library, was established. A target list of species for fecal sample collection was developed, including a numeric goal for each group (Table 1). Over the course of the project, multiple attempts were made to gather known source samples. Specific arrangements were made to meet with landowners and collect both livestock and wildlife samples. Human wastewater treatment facility (WWTF) samples were collected from both the inlets and outlets of functioning WWTFs in the watersheds. On-site sewage facility (OSSF) samples were collected from septic pump trucks operating in the watershed areas. Lastly, road kill was also utilized as a source of wildlife samples when opportunities presented themselves.

Table 1. Target known-source fecal sample collection for the Aransas and Mission watersheds.*

Fecal Source	**# of Samples**	**Notes**
Human	8-10	WWTFs and OSSFs- 1 sample per WWTF if possible
Livestock	10	Cattle, goats, sheep, horses, etc.
Wildlife	18-20	Avian wildlife, rats, deer, raccoons, swallows, possums, skunks, etc.
Other	8-10	Feral hogs and any other opportunities that present themselves
Pets	5	Focus on dogs and cats
Total	**~50**	

*Targeted 25 samples per watershed, but ultimately grouped analyzed samples together in the Texas *E. coli* BST Library to represent one contiguous area given the proximity of the watersheds.
WWTF, wastewater treatment facility; OSSF, on-site sewage facility

Known-source sampling in the Mission and Aransas watersheds resulted in a total of 71 unique samples being collected between December 2018 and November 2019. Samples collected were held on ice until being transported to AgriLife SCSC for processing within 96 hours of collection. A portion of the known-source samples, upon receipt by AgriLife SCSC, were shipped to UTSPH-EP within 24 hours for processing. Table 2 describes the number of samples collected per source, the number of samples testing positive for *E. coli*, screened, validated, archived and added to the Texas *E. coli* BST library.

Of the 71 fecal known-source samples processed, 63 had culturable *E. coli* as determined using Environmental Protection Agency (EPA) Method 1603 (modified membrane thermotolerant *Escherichia coli* (m TEC) and NA-MUG positive). A total of 151 isolates from these samples were collected and archived. The samples were split between the AgriLife SCSC and UTSPH EP labs for ERIC-RP.

AgriLife SCSC archived 51 *E. coli* isolates from 37 known-source samples (up to 3 isolates per sample). All 51 *E. coli* isolates were screened with ERIC-PCR, and 34 known-source isolates from 34 unique known-source fecal samples were DNA fingerprinted using RiboPrinting (ERIC-RP). There were some fecal samples (cow, cat, dog, coyote, deer and mouse) that did not yield culturable *E. coli* via processing using EPA Method 1603.

UTSPH EP archived 100 *E. coli* isolates from 26 known-source samples (up to 5 isolates per sample). A total of 75 *E. coli* isolates (up to 3 per sample) were fingerprinted with ERIC-PCR. After screening for clonality, UTSPH EP analyzed 36 of these isolates with RiboPrinting (RP).

Collectively between AgriLife SCSC and UTSPH EP, 70 *E. coli* isolates from 60 known-source samples were fingerprinted by ERIC-RP.

The 70 known-source isolates from the Mission-Aransas local library were screened using the traditional self-validation step (a stringent 7-way split of source classes and an 80% similarity cutoff), resulting in 44 self-validated (SV) isolates from 42 samples.

These 44 SV isolates from the local library were combined with the similarly screened isolates from all previous watershed studies in order to perform serial Jackknife analyses to create the Texas *E. coli* BST library ver. 03-20, which contains 1,912 isolates, including 30 from the Mission and Aransas watersheds.

The composition, ARCC and diversity of this new version of the library are detailed in Table 3.

During the project period, the Texas *E. coli* BST Library was used to identify fecal pollution source contributions in the Aransas River and Mission River watersheds as part of this project and other watersheds as part of multiple projects funded by the Texas Commission on Environmental Quality (TCEQ) and the San Antonio River Authority (SARA).

Table 2. Number of samples collected per source, the number of samples testing positive for *E. coli*, screened, validated, archived and added to the Texas *E. coli* BST Library.

Source	Samples Collected	Samples (+) for E.coli	Isolates archived	Isolates screened by ERIC	Isolates RP in local library	Self-validated (isolate/sample)	TXSV 03-20 (isolate/sample)
Human	**11**	**11**	**24**	**24**	**11**	**8/8**	**2/2**
Sewage	8	8	16	16	8	7/7	2/2
Septic	3	3	8	8	3	1/1	0/0
Cattle	**12**	**10**	**37**	**37**	**13**	**5/4**	**3/2**
Other non-avian livestock	**4**	**4**	**9**	**9**	**4**	**4/4**	**3/3**
Horse	4	4	9	9	4	4/4	3/3
Other avian livestock	**2**	**2**	**2**	**2**	**1**	**1/1**	**1/1**
Chicken	2	2	2	2	1	1/1	1/1
Pets	**10**	**8**	**22**	**22**	**9**	**6/6**	**5/5**
Cat	4	3	15	15	4	3/3	2/2
Dog	6	5	7	7	5	3/3	3/3
Avian Wildlife	**4**	**4**	**5**	**5**	**4**	**1/1**	**0/0**
Crow	1	1	1	1	1	0/0	0/0
Dove	2	2	2	2	2	1/1	0/0
Wild Turkey	1	1	2	2	1	0/0	0/0
Non-Avian Wildlife	**28**	**24**	**52**	**51**	**28**	**19/18**	**16/15**
Bobcat	1	1	3	2	2	1/1	1/1
Coyote	1	0	0	0	0	0/0	0/0
Deer	3	2	2	2	2	2/2	2/2
Hog, Feral	10	10	23	23	13	8/7	7/6
Mouse	3	1	1	1	1	1/1	0/0
Possum	2	2	2	2	1	1/1	1/1
Raccoon	7	7	20	20	8	5/5	5/5
Skunk	1	1	1	1	1	1/1	0/0
Total	**71**	**63**	**151**	**150**	**70**	**44/42**	**30/28**

Escherichia coli, *E. coli*; bacteria source tracking, BST; Enterobacterial repetitive intergenic consensus sequence, ERIC; RiboPrinting, RP

BST Library Refinement

UTSPH EP and AgriLife SCSC collaborated to evaluate the geographical and temporal stability, composition, ARCC (accuracy) and diversity of source specific isolates, while continuing to further develop and refine the Texas *E. coli* BST library with new known-source isolates.

Table 3. Texas *E. coli* BST Library (ver. 03-20, cross-library validation) composition and rates of correct classification (RCCs) by Jackknife analysis of ERIC-RP composite data sets using an 80% similarity cutoff and 3- and 7-way splits.

Source Class	Number of Isolates	Number of Samples	Library Composition and Expected Random Rate of Correct Classification*	Calculated Rate of Correct Classification (RCC)	RCC to Random Ratio***	Left Unidentified (unique patterns)
HUMAN	**426**	**362**	**22%**	**100**	**4.5**	**22%**
DOMESTIC ANIMALS	**561**	**503**	**29%**	**100**	**3.4**	**19%**
Pets	89	80	5%	83	16.6	42%
Cattle	248	216	13%	93	7.2	10%
Avian Livestock	98	86	5%	86	17.2	28%
Other Non-Avian Livestock	126	121	7%	91	13.0	15%
WILDLIFE	**925**	**788**	**48%**	**100**	**2.1**	**17%**
Avian Wildlife	273	251	14%	79	5.6	19%
Non-Avian Wildlife	652	537	34%	92	2.7	16%
%Overall	**1912**	**1653**		**ARCC** = 3-way 100% 7-way 91%**		**19%**

*RARCC, expected random average rate of correct classification based on library composition
**ARCC = average rate of correct classification: the proportion of all identification attempts which were correctly identified to source class for the entire library, which is similar to the mean of the RCCs for all source classes when the number of isolates in each source class is similar
***An RCC/Random Ratio greater than 1.0 indicates that the rate of correct classification is better than random. For example, the rate of correct classification for human is 4.5-fold greater than random chance based on library composition.

Escherichia coli, E. coli; bacteria source tracking, BST; Enterobacterial repetitive intergenic consensus sequence RiboPrinting, ERIC-RP

To increase its accuracy and utility, the updated Texas *E. coli* BST Library with pooled SV local watershed libraries as described in Table 3 (2,299 isolates) was refined through cross-validation. To attempt to remove cosmopolitan (non-specific*) E. coli* source isolates, repetitive Jackknife analyses of the combined SV libraries were performed to remove isolates that cross-identified between human, domestic animals and wildlife with the goal of 100% ARCC using a 3-way split of source classes. In the first round of serial Jackknife analysis, 343 isolates were removed leaving 1,956 isolates. Four additional rounds of Jackknife analysis were performed, resulting in 1,912 isolates with a 100% ARCC using a 3-way split of source classes and a 91% ARCC using a 7-way split. A total of 19% of the isolates were singletons (SVUs) (i.e., unique fingerprints; Table 3). The Texas *E. coli* BST Library ver. 03-20 contains 1,912 isolates obtained from 1,653

individual fecal samples. Library composition is based on 3- and 7-way source class splits (Figures 1 and 2 respectively).

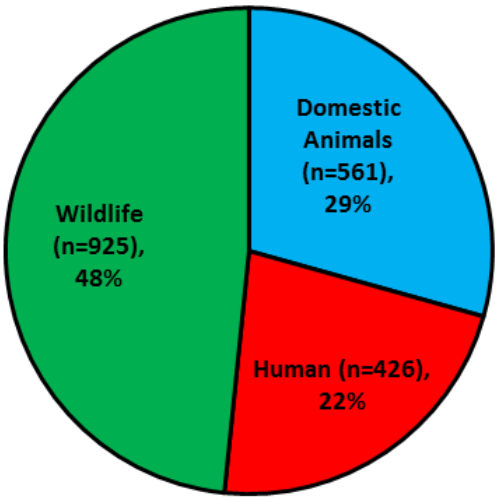

Figure 1. Texas *E. coli* BST Library (ver. 03-20) composition by 3-way split of source classes (1,912 isolates from 1,653 different fecal source samples).

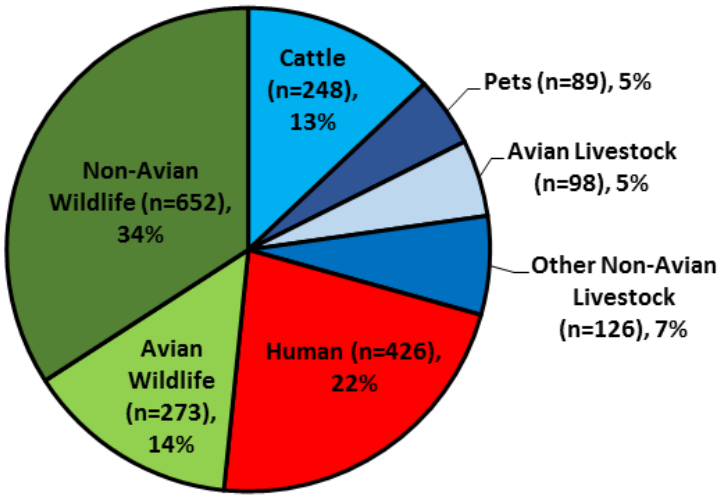

Figure 2. Texas *E. coli* BST Library (ver. 03-20) composition by 7-way split of source classes (1,912 isolates from 1,653 different fecal source samples).

Trinity River – Galveston Bay

In addition to the Aransas River and Mission River watersheds known-source isolates that were added to generate the Texas *E. coli* BST Library ver. 03-20, 76 isolates were also fingerprinted and evaluated through the Bacterial Source Tracking (BST) on Tributaries of Trinity and Galveston Bays project funded by TCEQ (Contract Number: 582-18-80240). Of the 91 total known source fecal samples collected by TWRI from the watershed, AgriLife SCSC successfully isolated *E. coli* from 77 individual samples. A total of 76 of these isolates (one isolate per known source sample) were screened using ERIC-RP and included in the local watershed library. UTSPH-EP did subsequent library evaluation and used Jackknife analysis of the ERIC-RP to identify isolates that correctly classified using a 7-way split of source classes (i.e., human, pets, cattle, other non-avian livestock, avian livestock, avian wildlife and non-avian wildlife). Isolates with unique fingerprints (left unidentified using an 80% similarity cutoff) were also included to create the local SV library. In total, 46 isolates were SV in the local library.

The 46 local SV source isolates from the watershed were then added to the current library of Texas *E. coli* BST SV source isolates from 13 previous watershed projects across Texas. A series of Jackknife analyses were run on the combined libraries, removing all isolates that cross-identified between human, domestic animals and wildlife. After each removal, the Jackknife was run again with the goal of 100% ARCC using a 3-way split of source classes. After four iterations of cross-watershed validation, the resulting Texas *E. coli* BST Library (ver. 1-20) contained 1,886 isolates from 1,645 samples, resulting in a 100% ARCC with a 3-way split of source classes and a 91% ARCC using the 7-way split of source classes. A total of 19% of the isolates were identified as SVU (unique fingerprints left unidentified using an 80% similarity cutoff) and were kept in the library in order to reflect the diversity of patterns potentially seen in unknown water samples. After cross-watershed validation, 33 isolates (43% of the local library samples) were included in the Texas *E. coli* BST Library (ver. 1-20). The 33 isolates were comprised of individual fecal samples from cattle (1), goat (1), domestic cat (1), sewage (4), septic (6), feral hogs, armadillo, opossum, raccoon, squirrel, deer (19) and seagull (1). The 76 isolates were included with the new Aransas and Mission isolates when evaluating and generating the newest version of the library as described in the previous section (Table 2).

Utilization of the Texas *E. coli* BST Library

Aransas River

TWRI collected Aransas River water samples monthly at TCEQ Station ID 12947 from December 2018 through November 2019. Samples were delivered to AgriLife SCSC for initial processing using EPA Method 1603 and subsequent shipment of isolates to UTSPH EP. Other than the first month of sampling, *E. coli* counts were low, ranging from 1 to 33 colony forming units (CFU)/100 mL (Table 4). Up to five presumptive *E. coli* cultures were isolated from each sample confirmed as *E. coli* (modified mTEC and NA-MUG positive), and up to three isolates per sample were selected for ERIC-PCR and RP. UTSPH EP performed BST analysis to support watershed planning efforts in the Aransas River watershed. A total of 120 *E. coli* isolates were fingerprinted using ERIC-RP and compared against Texas *E. coli* BST Library v. 03-20 for source determination. Overall results for the Aransas River isolates are shown in Figures 3 and 4.

Using a 3-way split, 71% of the isolates were classified as originating from wildlife, 12% from livestock and domesticated animals and 6% from humans. Using the more detailed 7-way split, 48% of the isolates were categorized as originating from non-avian wildlife, 23% from avian wildlife, 8% from cattle, 6% from humans and 4% for other non-avian livestock. The source could not be identified for 12% of the isolates.

Table 4. Monthly counts of *E. coli* isolates from water samples for Aransas River (TCEQ Station ID 12947) between December 2018 and November 2019.

Sampling Months	*E. coli* (CFU/100 mL)
December '18	576
January '19	2
February '19	12
March '19	33
April '19	3
May '19	1
June '19	3
July '19	4
August '19	2
September '19	4
October '19	3
November '19	7
Geometric Mean	6.1

CFU, colony forming unit; *Escherichia coli*, *E. coli*; Texas Commission on Environmental Quality, TCEQ

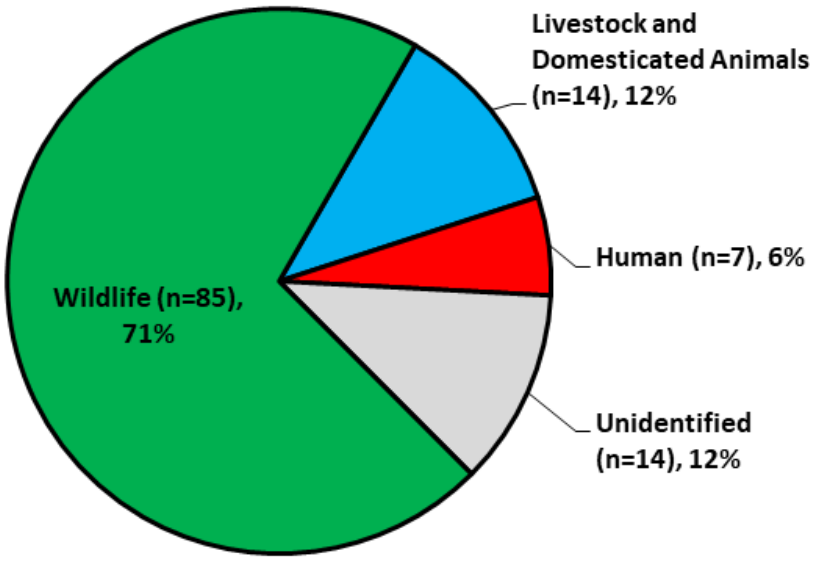

Figure 3. Source classification of *E. coli* isolates (combined n=120) from Aransas River (TCEQ Station ID 12947) using a 3-way split.

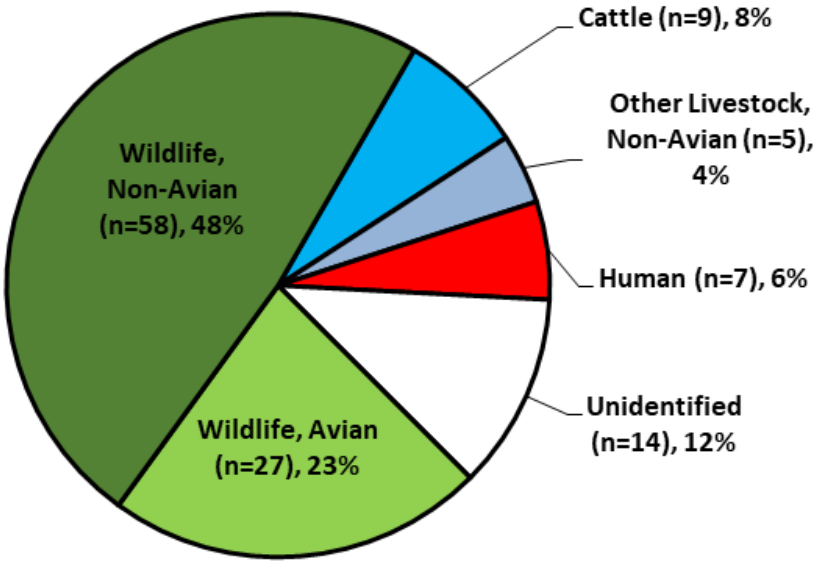

Figure 4. Source classification of *E. coli* isolates (combined n=120) from Aransas River (TCEQ Station ID 12947) using a 7-way split.

As with many previous watershed projects in mostly rural watersheds, the dominant sources at the Aransas River site appear to have been wildlife, with large contributions from both non-avian and avian wildlife species. Human and livestock contributions appear to have represented relatively smaller portions of the sources.

Mission River

TWRI collected Mission River water samples monthly at TCEQ Station ID 12943 from December 2018 through November 2019 (Table 5). Samples were delivered to AgriLife SCSC for processing using EPA Method 1603. Other than the first month of sampling, *E. coli* counts were low, ranging from 3 to 48 CFU/100 mL (Table 5). Up to three isolates from each sample were isolated, confirmed as *E. coli* (modified mTEC and NA-MUG positive) and archived. All isolates were fingerprinted using ERIC-PCR and RP. A total of 121 *E. coli* isolates were fingerprinted using ERIC-RP and compared against Texas *E. coli* BST Library v. 03-20 for source determination. Overall results for the Mission River isolates are shown in Figures 5 and 6.

Using a 3-way split, 66% of the isolates were classified as originating from wildlife, 8% from livestock and domesticated animals and 3% from humans. Using the more detailed 7-way split, 49% from non-avian wildlife, 17% of the isolates were classified as originating from avian wildlife, 3% from cattle, 3% from humans, 2% for other non-avian livestock, 2% for other avian livestock and 2% from pets. The source could not be identified for 22% of the isolates.

Table 5. Monthly counts of *E. coli* isolates from water samples for Mission River (TCEQ Station ID 12943) between December 2018 and November 2019.

Sampling Months	*E. coli* (CFU/100 mL)
December '18	141
January '19	3
February '19	6
March '19	8
April '19	48
May '19	6
June '19	9
July '19	2
August '19	6
September '19	17
October '19	6
November '19	9
Geometric Mean	9.6

CFU, colony forming unit; *Escherichia coli*, *E. coli*; Texas Commission on Environmental Quality, TCEQ

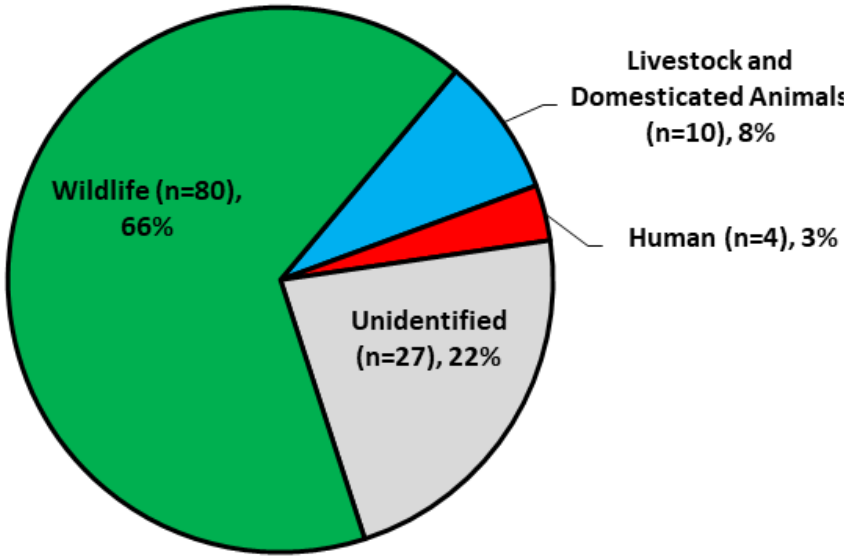

Figure 5. Source classification of *E. coli* isolates (combined n=121) from Mission River (TCEQ Station ID 12943) using a 3-way split.

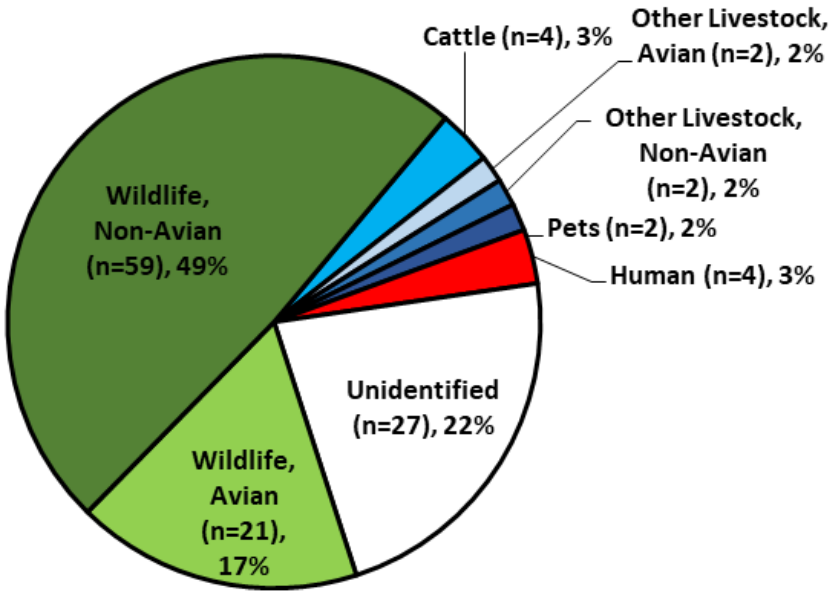

Figure 6. Source classification of *E. coli* isolates (combined n=121) from Mission River (TCEQ Station ID 12943) using a 7-way split.

The dominant sources at the Mission River site appears to have been wildlife, with large contributions from both non-avian and avian wildlife species. Human and livestock contributions appear to have represented relatively smaller portions of the sources. The Mission River BST results were similar to those from the Aransas River, which is consistent with the close geographical proximity and similar land uses in the two watersheds.

Other Watersheds

AgriLife SCSC continued BST analysis in support of SARA's watershed characterization efforts. A total of 130 *E. coli* isolates from 14 water samples collected in 2018-2019 were isolated, verified as *E. coli*, fingerprinted by ERIC-RP and compared against Texas *E. coli* BST Library ver. 12-17 for source identification.

As part of the "BST on Tributaries of Trinity and Galveston Bays" project funded by TCEQ (Contract Number: 582-18-80240), monthly water samples were collected by TWRI from April 2018 thru April 2019 at five sampling locations: Dickinson Bayou, Cedar Bayou, Buffalo Bayou, Double Bayou and Clear Creek. AgriLife SCSC cultured and ERIC-RP fingerprinted a total of 241 *E. coli* isolates that were compared against Texas *E. coli* BST Library ver. 03-20 for source identification. The Galveston-Trinity project is in the final stages of data analysis and will be completed in summer 2020.

AgriLife SCSC is also doing library-dependent BST in support of the "Geronimo and Alligator Creeks Watershed Protection Plan Implementation – Environmental Education Site Coordinator and Bacteria Source Tracking" project funded by TCEQ and USEPA (Federal ID #9961423). Monthly water samples from two sites in the Geronimo Creek watershed were collected by Guadalupe-Blanco River Authority from April 2018 through March 2019 and processed using EPA Method 1603. Plates containing *E. coli* colonies were then shipped to AgriLife SCSC for *E. coli* isolation, ERIC-RP fingerprinting and source ID. Approximately 100 *E. coli* isolates are being fingerprinted and compared against Texas *E. coli* BST Library ver. 03-20 for source identification. The BST for this project is scheduled to be completed in May 2020.

Evaluation of the Texas *E. coli* BST Library

BST in Texas began in the early 2000s to apply molecular techniques to the challenges of determining the source(s) of fecal pollution in lakes, rivers and streams. While elevated concentrations of *E. coli* in a water body indicated the potential presence of fecal pollution and increased health risk for recreational use, more information concerning where that *E. coli* originated was warranted. If the source of the fecal contamination could be determined, BMPs could efficiently mitigate that pollution. BST is based upon the premise that different bacterial strains have adapted to different guts and therefore have genotypic and phenotypic differences specific to different hosts that can be detected and allow for the identification of the original host species and source of the fecal contamination.

Numerous BST methods have been developed that use DNA fingerprints and bacterial markers to identify sources of fecal pollution. Several factors should be considered when choosing a method. *E. coli* has been chosen as the target bacterium because it provides a direct link with water quality standards. This indicator organism has standardized methods of detection and is present when fecal contamination is present and absent when it is not (see FY15 report TR-496 for evidence against role of naturalized *E. coli*).

While library-independent based methods using *E. coli* have recently been developed (Seinkbeil et al. 2019), using *E. coli* for source tracking has largely required library-dependent methods. Based on a multi-year study initiated in 2002, the state of Texas selected the two-method approach using ERIC-PCR and RiboPrinting (ERIC-RP), as this approach was found to be the most accurate and cost-effective. The BST method should be able to detect the host sources present and of concern. *Bacteriodales* and other library-independent BST methods are limited by which markers have been developed (ruminant, human, hog, horse, dog, chicken and cow) and cannot directly detect the impact of all wildlife. While wildlife is more difficult to manage due to the impracticality of limiting wildlife access to a water body, our previous studies have found that wildlife often accounts for at least 50% of the source identifications. There must also be verification that the marker developed in one area can be applied to another watershed, being able to detect the animal of interest, without false positive interference. Therefore, known source isolates from the local watershed should still be collected.

The Texas *E. coli* BST Library was first developed as a way to counter one of the biggest drawbacks of library-dependent source tracking methods — the need to collect large numbers of known source samples for every watershed study. An identification library should reflect the large host inter- and intra-species variation in *E. coli*, so that the DNA fingerprints of *E. coli* isolates found in the water can be matched and their sources identified. We theorized that there would be enough geographical and temporal stability in the host specificity of *E. coli* populations to allow the DNA fingerprints from known source isolates from different watershed studies to be pooled together. Developing a statewide BST library using *E. coli* isolates from local watershed libraries allows for time and cost savings.

Through the Texas BST Program, over 20 watershed studies have been conducted in over a dozen watersheds across the state of Texas. The Texas *E. coli* BST Library is dynamic, with new isolates being added with each successive BST project. This has resulted in 3,764 ERIC-RP composite DNA fingerprints from 3,062 samples from over 50 source subclasses of wildlife, domestic animals and humans, representing a collection effort of over 4,300 samples from over 140 subclasses, with more than 11,000 known source *E. coli* isolates archived (Table 6).

Table 6. Sample collection, fingerprinting and screening for Texas *E. coli* BST Library (ver. 03-20).

Watershed	# of total samples collected	# of (+) samples	# of isolates archived	# of isolates ERIC-PCR	# of isolates Ribo-Printed	# of isolates local library	# of samples local library	# of isolates self-validated	# of samples self-validated	# of isolates in TXSV 03-20
San Antonio	1013	786	3330	2107	947	933*	778 **	457	403	389
Waco-Belton	1143	834	3224	2275	1079	958	813	537	481	488
Lake Granbury	74	59	198	173	80	80	59	60	48	43
Oyster Creek	355	298	292	286	286	286	286	166	166	129
Trinity River	193	130	129	128	128	128	128	67	67	46
Buck Creek	60	28	53	53	31	31	28	20	20	13
Little Brazos River	75	66	166	63	85	85	66	66	57	51
Leon (SCSC)	30	30	146	146	72	72	30	58	27	41
Leon (UTSPH)	95	71	323	204	133	132	71	85	60	75
Lampasas	118	85	384	244	145	143	83	97	67	79
Big Cypress	30	19	73	73	34	34	19	28	16	22
Attoyac	156	113	494	113	113	113	113	72	72	56
Leona	260	201	900	201	201	201	201	94	94	77
Arroyo Colorado	254	99	409	274	144	142	98	74	60	57
Infra 2013 Leon	25	24	120	72	31	31	24	26	20	23
Infra 2013 SA	75	72	358	216	125	125	72	120	70	108
Riesel	56	46	189	116	116	58	46	56	45	53
Birds and Bridges	20	6	24	24	24	11	6	11	6	8
Big Elm Creek	47	44	218	132	79	79	44	62	40	49
Plum Creek	59	53	106	76	76	76	52	53	40	42
Trinity River SCSC	91	77	151	151	76	76	71	46	44	33
Mission Aransas	72	63	151	126	70	70	60	44	42	30
TOTAL	4301	3204	11438	7253	4075	*3764+ 100 zoo	**3062 +86 zoo	2299	1945	1912

Escherichia coli, *E. coli*; Enterobacterial repetitive intergenic consensus sequence polymerase chain reaction, ERIC-PCR

The standards and methods used to create the Texas *E. coli* BST Library have developed over time. Currently, three steps are used to refine the Texas *E. coli* BST Library: de-cloning, self-validation and cross-validation of isolates. Underlying these steps are the use of host classes and how to define a match. To better evaluate these steps, all 3,764 known source ERIC-RP patterns in the database were consecutively ran against all known human isolates (730), all known domestic animal isolates (1,438) and all known wildlife isolates (1,596). The overall best match for each isolate was also calculated by determining the matching host class with the highest percent of similarity.

Defining Host Class

BST is based on the premise that different strains of *E. coli* have adapted to different gut environments to become host specific. There are many caveats to consider when dealing with *E. coli* populations. There are different strains of *E. coli* in a single individual while different strains of the bacteria will exist in different individuals. Similar strains may be present in similar environments. Strains of *E. coli* present may be dependent upon animal species, gut type, diet, environment and interactions with other individuals/species. The questions of temporal and geographical variations are especially relevant to a pooled library, such as the Texas *E. coli* BST Library, since it is built over time from different watersheds. While library-independent markers attempt to identify a limited specific animal species, our library-dependent approach can work with these realities of *E. coli* populations while still giving practical results by using host classes. The most supportable division of known sources is into three host classes: human, domestic animals and wildlife. These embrace the adaptations to a shared environment, allow the use of a wide variety of wildlife and do not penalize for the cross-identification seen between livestock. The division of water isolates into human, domestic animals and wildlife is also practical for making decisions about BMPs. There is also more statistical strength when small numbers of isolates are divided into fewer categories. To ensure a variety of host classes are included in collections for the library, domestic animals and wildlife are further divided into subsets. Below are the 3- and 7-way split categories that were used for categorizing *E. coli* isolates and which we have most frequently used for characterizing watersheds:

3-way split

1. Domesticated animals and livestock (livestock and pets)
2. Wildlife (including feral hogs)
3. Humans

7-way split

1. Cattle
2. Other livestock, non-avian (non-avian livestock other than cattle; sheep, etc.)
3. Other livestock, avian (chickens, etc.)
4. Pets (dogs, cats)
5. Avian wildlife (ducks, geese, sparrows, etc.)
6. Non-avian wildlife (deer, feral hogs, coyotes, etc.)
7. Humans

The 80% Similarity Cut-Off Standard

The BioNumerics software (Applied Maths) used to analyze DNA fingerprints creates a matrix comparing each ERIC-RP composite DNA fingerprint to every other fingerprint in that set. The percent similarity between each pair of fingerprints is based on band position and intensity. We

know from the quality control strains that have been run on each ERIC gel and day of RP that the reproducibility of those fingerprint patterns varies only slightly from batch to batch, day to day, over time (decades), laboratories and technicians. Quality Control (QC) strains from ERIC gels generally cluster together at 85% similarity or greater, while RP patterns for QC strains cluster at a slightly higher similarity. Therefore, isolates that match each other at 85% or more similarity can be considered the same strain. Using a composite of these two fingerprint types seems to build in more variation. Consequently, isolates that are less than 80% similar to each other, based on ERIC-RP, should be considered different strains. Currently, the same similarity cut-off standard has been used for both water identification applications and library screening/building. This makes sense for attempting to determine if a water isolate is the same strain as a known-source library isolate. An isolate whose best match is less than 80% similar is considered to not have a match. For a known source isolate in a Jackknife analysis, this means the isolate has a unique pattern and has been called a "singleton (SVU)." For a water isolate, it means it remains unidentified. There must be a balance between having definitive but few matches versus many but less definitive matches. Early studies considered the similarity cut-off both in terms of the number of water isolates left unidentified and the rates of correct classification in library Jackknife analyses. While an 85% similarity cut-off may improve by a few percentage points the rate of classification (100*number of good matches divided by the number of match attempts, i.e. correct matches plus incorrect matches; does not include isolates left unidentified), it may also increase the number of isolates left unidentified (best match is below the similarity cut-off) significantly.

Use of Best Match

As discussed above, the BioNumerics software (Applied Maths) used to analyze DNA fingerprints creates a matrix comparing each ERIC-RP composite DNA fingerprint to every other fingerprint in that set. Those *E. coli* isolates that are designated as "unknowns" are compared to the other isolates in the set that are considered "knowns." A customized algorithm, or script, created by Applied Maths, then lists the best match to a single "known" isolate for each "unknown" isolate, giving the percent similarity to each other. Although the match is to a single isolate that came from a particular known animal fecal sample, it is considered a match to the host class. A match is wholly dependent upon the composition of the library.

Currently, library screening/building is based on Jackknife analysis and whether a known source isolate has found a "correct" best match, an "incorrect" best match, or has been left "unidentified." If a known source isolate cannot find a best match to another known source isolate in the library at greater than 80% similarity, it is considered a SVU or unique pattern. The term "unidentified" really should not apply since its true source is known, but it did not correctly match to its specific host class during Jackknife analysis. When *E. coli* isolates from known source samples are being compared to each other, determining that an ERIC-RP composite DNA fingerprint from one source is a "correct match" to an ERIC-RP pattern from another host class downplays the fact that both are legitimate patterns from their respective known sources.

As additional *E. coli* fingerprints are added to the Texas *E. coli* BST Library, there is a need to examine other strategies for testing unknown water isolates against the library. Using the best match-approach, even though several known source isolate DNA fingerprints from the Texas *E. coli* BST Library may be greater than 80% similar to an "unknown" water isolate, its source identification will be determined only by the isolate with the highest percentage match. For example, if there are 10 human isolates in the library that match an unknown at 87% similarity

and one cattle isolate that matches it at 88% similarity, the unknown will be identified as a cattle isolate using the best match approach. To better explore this phenomenon, all 3,764 ERIC-RP patterns in the database were consecutively ran against all known human isolates (730), all known domestic animal isolates (1,438) and all known wildlife isolates (1,596). For any host class, at least 70% of the isolates could find a similar ERIC-RP fingerprint (of at least 80% similarity) in either or both of the other host classes. A substantial portion of isolates (27-44%) were at least 90% similar to an isolate from another host class (see cosmopolitan discussion below). While the overall best match was to the correct host class for the largest fraction of known source isolates (even before removal of cosmopolitan isolates and other steps used in library development/validation), this does indicate potential complications for the best match approach as the library expands. Cosmopolitan isolates and strategies for reducing their impacts on source identification are discussed further in the subsequent section on cross-validation.

De-cloning vs 1 isolate per sample

The de-cloning screening process compares the ERIC–PCR patterns from up to three isolates per individual known source fecal sample. Isolates that are greater than 80% similar are considered clones (the same strain) and subsequently, only one isolate is selected for further consideration. All de-cloned isolates from individual source samples are included in their respective local watershed library, independent of their similarity to other library isolates. Keeping clones of multiple isolates from a single sample does not help identify water isolates and only inflates rates of correct classification in a library Jackknife analysis, besides adding the expense of RP superfluous isolates. An alternative to the de-cloning screening process would be to only fingerprint one isolate per sample. This however, would increase the number of known source samples needed and would require balancing the time and effort needed to collect the additional known source samples against the benefits of reduced de-cloning efforts. Generally, screening three isolates per known source sample results in an average of 1.5 de-cloned isolates per sample. Specifically, for the 26 known source samples from the Mission Aransas project that were processed by UTSPH-EP, 75 isolates were screened by ERIC-PCR, resulting in 37 de-cloned isolates. For over 60% of the samples (16), all of the screened isolates were clones resulting in one selected isolate for each sample. Approximately 1/3 of the samples (9) each had two unique patterns, while in one sample, all three isolates had ERIC-PCR fingerprints that were very different from each other. Overall, 37 isolates from 26 known source samples were available for library inclusion, which maximized the collection effort by 142%.

Self-validation

The local watershed library consists of the limited numbers of de-cloned isolates from known source samples collected from the same time and place as the water isolates. These act to supplement the statewide library with isolates that are temporally and geographically contemporary to the water isolates. As such, they can also be treated as unknowns to mimic how any temporal and geographic shifts may affect the identification of isolates from the new watershed.

Currently, self-validation of the local watershed library composite ERIC–RP fingerprints is performed using Jackknife analysis to identify isolates that are correctly classified using a 7-way split of source classes (i.e., human, pets, cattle, other non-avian livestock, avian livestock, avian wildlife and non-avian wildlife (including feral hogs)). Every local known isolate is compared to every other local known isolate from the watershed to determine its best match (highest similarity between ERIC-RP composite fingerprints). If there is no match (<80% similarity), the

isolate is considered a SVU with a unique pattern. The host classes of the isolate and its best match are compared. If they have the same 7-way host class, they are considered SV. The SV and SVU isolates make up the SV library that goes on to the next step of the library screening process, typically about 2/3 of the original isolates.

The remainder of the local library isolates are considered to be "incorrect" if their best matches were above 80% similarity and did not agree at the 7-way split subset host class level. The original rationale for such a strict self-validation step on the local library level was to apply the most stringent test for identifying host-specific known source isolates when there were no confounding temporal or geographical factors. The 7-way watershed self-validation step may be too conservative, especially when the local libraries are very small and not diverse. When *E. coli* isolates from known source isolates are being compared to each other, determining that an ERIC-RP composite DNA fingerprint from one source is an "incorrect match" to an ERIC-RP pattern from another host class downplays the fact that both are legitimate patterns from their respective known sources. Another confounding factor is that while the "incorrectly matched" isolate is removed, the isolate that it matched to may itself have a higher similarity best match that is correct, and therefore is kept in the library.

Of the 3,764 DNA fingerprint patterns in the database, 39% (1,465) were considered "incorrect" using a 7-way split of source classes in their local watershed library. When compared to all patterns in the database, 38% of these (554) then found "correct" matches, with 205 isolates still matching back to their original watershed local libraries at a 3-way split of host classes. The 7-way self-validation step seems to most adversely affect Other Non-Avian Livestock (55% loss) and Pets (53% loss), even though 47% and 34% of those lost, respectively, found "correct" matches when compared to all patterns in the database.

Of the total isolates in the database, 1,659 isolates were designated as SV in their local libraries. Of those, 84% had a best match to their 3-way host class, when compared to all patterns in the database, with 80% of those still coming from their original watershed study. As discussed in previous reports, a watershed exclusive comparison may be the best way to tease out the more geographically and temporally stable patterns.

Of the total isolates in the database, 640 (17%) were designated as SVU in their local libraries. When compared to all patterns in the database, 48% of these still remain unique patterns, while 29% found a confirming best match. All isolates considered unique patterns are carried through to the final version of the identification library. There is some concern that since these isolates have not been validated to their source, they may not be as dependable in identifying water isolates, although they were collected from samples from known sources. However, this could also mean that the DNA fingerprint pattern is unique and a record of that pattern had not yet been added to the library, highlighting the need for additional known-source isolates to be added to the library. Of the Mission River water samples, 10 of the 121 water isolates were identified by known source isolates originally designated as SVU. This was also true for the 10 of the 120 water isolates from the Aransas River.

Perhaps the best use of the local library isolates is to act as 1) a first identification screen of water isolates and 2) as a test for any identification libraries. When the Mission River water isolates were tested against the 70 isolate local source library (referred to as MAF 70), 31% of the water isolates (37) found best matches to the patterns from local human sources, including 22 isolates with greater than 85% similarity match. Using the Texas *E. coli* BST Library v. 03-20, only 3% isolates (4) identified as being from human sources. When the local library was treated

as unknowns and tested against the state library, only one of the 11 fecal isolates from local human source was correctly identified as human, with over half (6/11) identified as wildlife. The local library results should be interpreted cautiously since they are based on a relatively small known-source library, but they do indicate a potential consideration for future studies. Based on this, it is recommended that future projects investigate the need to give more weight to high similarity matches (85% or 90%) of water isolates to local source isolates. The self-validation step is currently used as a simple way to screen the known source isolates of the current project and add to the previously screened known source isolates.

Cross-validation

The final step in creating the Texas *E. coli* BST known source identification library has been cross-validation. The SV local watershed libraries described above were pooled together. In an attempt to remove non-specific *E. coli* source isolates, serial Jackknife analyses of the combined SV libraries were then performed to remove isolates that cross-identified between human, domestic animals and wildlife with the goal of 100% ARCC using a 3-way split of source classes and an 80% similarity cut-off.

While high rates of correct classification of library isolates are desirable, a goal of perfect library numbers may defeat the real purpose. The goal of a known source identification library in library-dependent BST is to correctly identify all the sources of the *E. coli* isolated from lakes, rivers and streams. This requires a large and diverse set of DNA fingerprints that reflect both the variation in *E. coli* populations and the variety of potential host sources.

Correct identification is complicated by the fact that not all *E. coli* are source-specific. Some strains of *E. coli* may be actively shared between different host classes, while other strains from different source classes may just be very similar to each other. The strains that are found in many different animals and humans are referred to as “cosmopolitan.” While the general definition of a cosmopolitan isolate is one that is found in more than one source class, a specific, measurable definition is needed, and that definition also needs to account for geographical and temporal variability. Several attempts have been made to develop a screening method that can identify such isolates. The Jackknife analyses used in the self-validation and cross-validation steps of library screening are based on the best match (highest percentage similarity) to a single isolate. An isolate that finds an incorrect match (80% or greater similarity to an isolate from another source class) is removed. It should be noted, however, that the isolate it incorrectly matched to may find its own best match to be correct (at a higher percent similarity). Subsequently, this isolate stays in the library. Therefore, a cosmopolitan strain may have been masked and pass through best match self-validation and cross-validation screenings. Previous studies have found that temporal and geographical cohorts are especially good at masking otherwise cosmopolitan isolates. On a per-watershed basis, 40-70% of isolates found their best matches with another isolate from their local watershed cohort (but from a different source sample due to de-cloning).

It is important to begin to think of the composite ERIC-RP fingerprints more as genotypes with different degrees of relatedness to each other. To determine how the patterns cross-identify, all 3,764 ERIC-RP patterns in the known source database were consecutively ran against all known human isolates (730), all known domestic animal isolates (1,438) and all known wildlife isolates (1,596). To take a conservative approach in our definition of cosmopolitan, a 90% similarity cut-off was set. If an isolate was at least 90% similar to fingerprint from an isolate in another source class, that isolate was considered cosmopolitan. Since each isolate was ran against each source

class, there could be no masking. There were a total of 1,661 known source isolates that could be considered cosmopolitan, 44% of the total ERIC-RP patterns in the database (Table 7.) Over half of these (861) crossed all three source classes (from a known source is one class and 90% or more similar to an isolate in each of the other two source classes.) Cosmopolitan isolates were rather evenly distributed over source classes, with 40% of the known source human isolates, 47% from domestic animals and 43% of the known source isolates from wildlife found to be cosmopolitan. Of the 1,465 known source isolates that were labeled "incorrect" by the self-validation step, 60% of them were cosmopolitan, but so were 45% of isolates designated "SV". Since 52% of the cosmopolitan isolates still found their best match overall within their own source class, it is not surprising that 627 of the 1,912 known source isolates in the Texas *E. coli* BST Library ver. 03-20 (33%) were cosmopolitan isolates based on the 90% similarity measure.

Table 7. Distribution of cosmopolitan isolates (from library of 3,764 isolates).

	Origin of Known-Source Isolate			
Cross-Match	Human	Domestic Animals	Wildlife	Totals
H—DOM	52	48	----	100
H—WILD	83	----	62	145
DOM—WILD	----	293	262	555
H—DOM—WILD	155	339	367	861
Total	290 (40% of 730)	680 (47% of 1438)	691 (43% of 1596)	1661 (44% of 3764)

human-domestic, H-DOM; human-wildlife, H-WILD; domestic-wildlife; DOM-WILD; human-domestic-wildlife; H-DOM-WILD

This analysis to find cosmopolitan isolates also found DNA fingerprint patterns that did not cross-identify with any different source class at 90% or greater similarity. Approximately 15% (526) of the total known source isolates still cross-identified, finding the wrong best match with the 80% similarity cut-off. However, 1,577 isolates had a best match to their own source class (sometimes at greater than 90% similarity) or were unique patterns SVUs. We combined these non-cosmopolitan isolates together as another library (Cosmo-free). We then treated the 70 Mission Aransas known source isolates (MAF 70) as unknowns and attempted to identify them using different library subsets;

- MAF 70 vs local Mission Aransas
- MAF 70 vs Texas *E. coli* BST Library ver. 03-20
- MAF 70 vs the Cosmo-free library
- MAF 70 vs all known source isolates

The results of the known source comparisons to different library subsets is shown in Table 8. The limited local library leaves the most isolates unidentified. The other libraries give very similar rates of correct classification. However, it should be noted that they do differ in their incorrect identifications.

Table 8. Library comparison using Mission Aransas known source samples (70 isolates).*

	MAF (70 isolates)	**03-20 (1912 isolates)**	**COSMO FREE (1577 isolates)**	**ALL (3764 isolates)**
	---------- Breakdown of known source isolates in each library ----------			
H%, DOM%, WILD%	16%, 39%, 46%	22%, 31%, 47%	20%, 35%, 45%	20%, 38%, 42%
	---------- Source identification of water isolates by each library ----------			
H (11 isolates)	50% (45%) (3; 5)	10% (9%) (1; 1)	10% (9%) (1; 1)	10% (9%) (1;1)
DOM (27 isolates)	38% (41%) (6; 11)	58% (11%) (14; 3)	58% (11%) (14; 3)	58% (11%) (14;3)
WILD (32 isolates)	63% (41%) (12; 13)	80% (22%) (20; 7)	80% (22%) (20; 7)	70% (16%) (19;5)

Results given as RCC (100#corrrect / (#correct + #incorrect)) with % left unidentified (#unidentified/(#correct + #incorrect + #unidentified = total) in 1st parentheses and raw numbers for #correct and #unidentified given in 2nd parentheses. Library names also list # of total isolates and % in each host class = library composition. All libraries are MAF watershed inclusive.
human, H; domestic animals, DOM; wildlife, WILD

We then used the same libraries to compare how they identified the Mission River water isolates (Table 9) and Aransas River water isolates (Table 10). As discussed previously, the local watershed library found more human signature from the Mission River. Otherwise, the libraries are consistent in ranking the sources of fecal contamination in both the Mission River and the Aransas River as wildlife, then domestic animals, followed by human. The greatest difference in percentage between groups is the extent of wildlife vs domestic animals.

Table 9. Comparison of library identification of Mission River water isolates.

Mission River Water (MRW)	**MAF 70**	**03-20**	**COSMO FREE**	**ALL**
HUM	31% (37)*	3% (4)	2% (2)	7% (8)
DOM	9% (11)	8% (10)	18% (22)	25% (30.5)
WILD	29% (35)	66% (80)	57% (69)	47% (56.5)
UNID	31% (38)	22% (27)	23% (28)	21% (26)

*Numbers in parentheses represent the # of isolates in that library subset for each host class (121 total)
Free of cosmopolitan isolate, COSMO FREE; Mission and Aransas 70 isolate local source library, MAF70

Table 7. Comparison of library identification of Aransas River water isolates.

Aransas River Water (ARW)	MAF 70	03-20	COSMO FREE	ALL
HUM	2% (2)*	6% (7)	2% (2)	4% (5)
DOM	23% (28)	12% (14)	30% (36)	29% (35)
WILD	49% (59)	71% (85)	54% (65)	57% (68)
UNID	26% (31)	12% (14)	14% (17)	10% (12)

*Numbers in parentheses represent the # of isolates in that library subset for each host class (120 total)
Free of cosmopolitan isolate, COSMO FREE; Mission and Aransas 70 isolate local source library, MAF70

Removing all known source cosmopolitan isolates from the library may be unrealistic, since they will be present in water samples as well. It is important to remember that a cosmopolitan isolate still came from a specific known source. One possible library approach would be using the local known source isolates supplemented by host-specific temporally and geographically stable known source isolates, with any water isolates left unidentified to then be ran against all available DNA fingerprint patterns. While the current state library incorporates local known source isolates, they usually only make up a small portion of the total library. The 30 known source isolates from the current Mission Aransas project makes up less than 2% of the 1,912 known source isolates in the Texas *E. coli* BST Library ver. 03-20. Perhaps a more tiered approach should be considered. Instead of serial Jackknife analysis to remove known source isolates, a serial library approach could be used to identify water isolates. Different similarity cut-offs could be potentially be used as part of this approach.

In a tiered approach, the first step may be to use the local watershed known source isolates. Note that 27 of the 70 local Mission Aransas known source isolates (39%) were considered cosmopolitan isolates when ran against all known source patterns. This may reflect some temporal or geographical shift in hosts. While the traditional self-validation step may be too stringent, perhaps a local cosmopolitan screening (all vs H, all vs DOM, all vs WILD with a 90% similarity requirement) should be done before using the local isolates. Note that self-validation would remove many of these potential cosmopolitan genotypes but not all. Any water isolates that do not find a match with 85% or greater similarity would then go to the second tier. The second-tier library would include all local known source isolates pooled with all previous known source patterns, and then screened to remove cosmopolitan isolates as described before. When attempting to identify water isolates, an 80% or 85% similarity cut-off could be used. Lastly, for a third-tier library any water isolates not yet identified could be ran against all known source patterns (without the removal of any cosmopolitan isolates) with the 80% similarity cut-off. Experimental tiered approaches to identify the water isolates from the current projects still showed the same ranking of results with wildlife being the largest contributor, followed by domestic animals and then humans.

This approach to finding cosmopolitan isolates (running all known source patterns against those known to be from human, domestic animals and wildlife, in parallel) could also be used in future work to begin to understand the probability that a particular pattern is associated with a particular host. It is clear that determining the accuracy of using DNA fingerprints of *E. coli* to identify the sources of fecal contamination in rivers, lakes and streams is a challenge, but a necessary challenge that can help policy makers and watershed authorities formulate BMPs that protect the safety of our recreational waters. Future work to expand the Texas *E. coli* BST

Library in order to better represent the temporal and geographical behavioral variation of *E. coli* will continue to be needed to achieve these goals.

Future Development of the Texas *E. coli* BST Library

As indicated in the preceding sections, continued evaluation, expansion and development of the Texas *E. coli* BST Library is needed as projects move into new watersheds and additional potential sources (e.g., human and domestic animals) are added. One key area for potential advancement is through more detailed statistical analysis of the library. There is concern about potential library bias since isolates from wildlife make up nearly 50% of the Texas *E. coli* BST Library, which should be examined by a random sampling, or similar, technique. Questions of certainty in water isolate identification should also be examined with the goal of calculating confidence intervals when determining sources. While all further analysis may give more insight into the biology and ecology of *E. coli* in the environment, the focus remains on how this information can be applied to the identification of the sources of fecal pollution in watersheds, and how this can be presented to stakeholders in a clear and useful manner.

Evaluation of Library-Independent PCR Markers

In an effort to expand the BST toolbox for future projects, additional library-independent markers and platforms were evaluated by AgriLife SCSC. As detailed below, the human-specific marker HumM2 was used for quantitative PCR (qPCR)-based analysis of water samples and preliminary tests were conducted on a droplet digital PCR-based assay for the human-specific marker HF183.

In response to Hurricane Harvey in August 2017, AgriLife SCSC, in collaboration with Dr. Michael LaMontagne at the University of Houston-Clear Lake (UHCL), initiated a project entitled "Characterization of Microbial Community Structure and Fecal Contamination of Floodwaters Generated by Hurricane Harvey" with funding from NSF (Project #CBET-1759540). Surface water samples were collected by UHCL at six locations in the southeastern Houston area immediately after the hurricane and then every one to two weeks thereafter over a two-month period. Samples were immediately transported to AgriLife SCSC for enumeration of *E. coli* using the IDEXX Quanti-Tray/2000 system and Colilert tests. Water samples were also filtered, had DNA extracted and were analyzed via qPCR for a total *Bacteroides* marker (Bernhard and Field 2000) and the HumM2 human-specific *Bacteroides* marker (Shanks et al. 2009, 2016). Protocols for two human-specific qPCR assays (HF183 and HumM2) were simultaneously evaluated by AgriLife SCSC for potential use. The HumM2 assay was optimized first, so it was used for subsequent sample analysis given the short timeline of the project.

The surface water samples collected immediately after the hurricane had elevated levels of *E. coli* ranging from 488 to 1,733 most probably number (MPN)/100 mL with a geometric mean of 1,019 MPN/100 mL across all sites. However, after one week, the *E. coli* levels had decreased to <100 MPN/100 mL at all sites (Figure 7, top panel). Levels of total *Bacteroides* were higher than *E. coli* but followed similar trends with elevated levels immediately following the hurricane and a rapid decrease in levels as the floodwaters dissipated (Figure 7, middle panel). In contrast, levels of human-specific *Bacteroides* (HumM2) were highest around one week after the hurricane passed and then decreased (Figure 7, bottom panel). The relatively low levels of human *Bacteroides* detected at the first sampling date, during flooding and when maximum *E. coli* and total *Bacteroides* levels were observed, suggests that non-human fecal sources were primarily responsible for contamination during the initial flooding. However, the delayed (one

week) spike in human *Bacteroides* marker abundance, and increased fraction over time (relative to total *Bacteroides*), indicates the prevalence of human sources under normal conditions. Based on these results, it appears that HumM2 was useful for detecting shifts in fecal contamination sources and should be considered for deployment in future BST projects where delineation of human-specific contamination is needed.

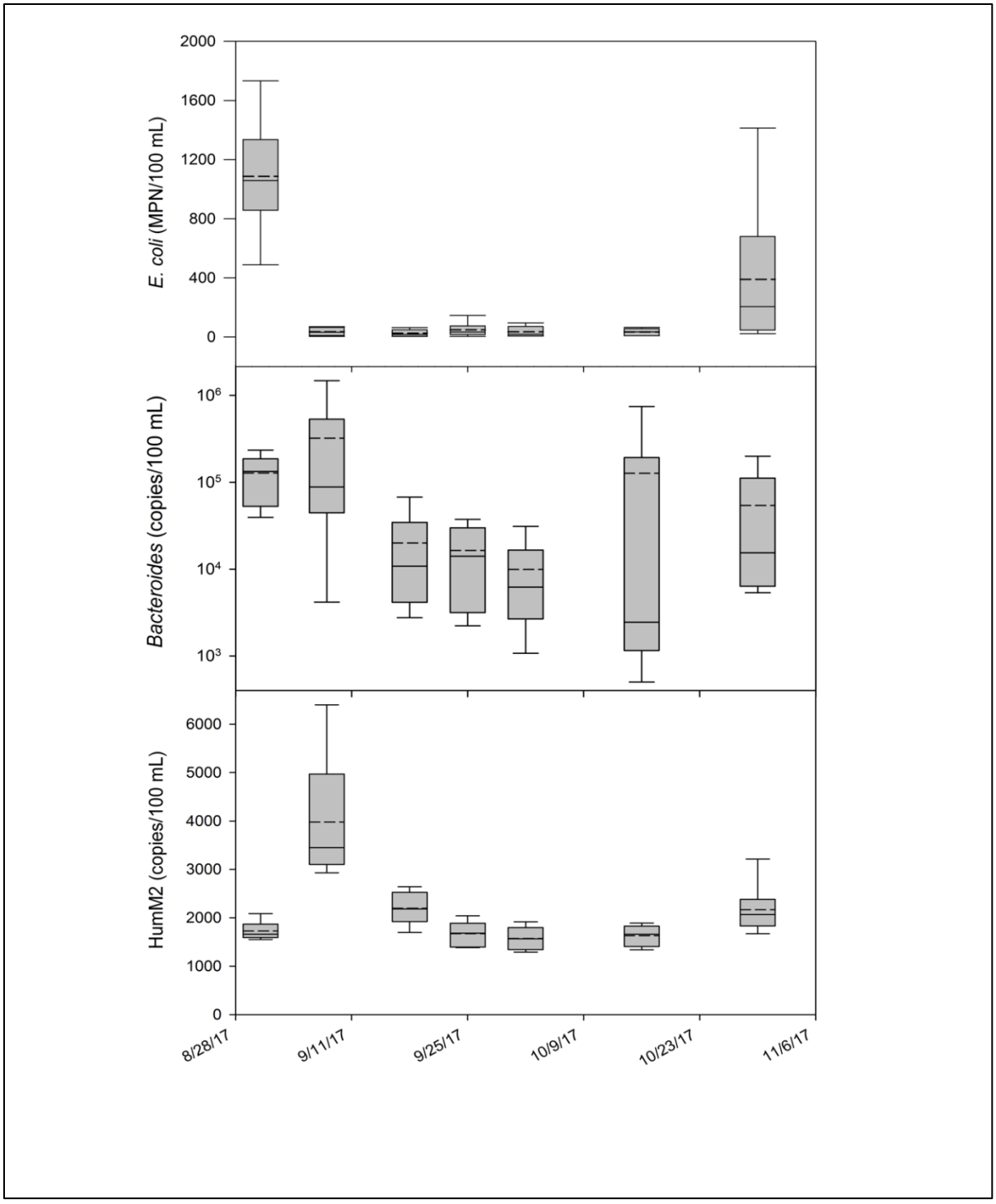

Figure 7. Measured levels of *E. coli* (top panel), total *Bacteroides* (middle panel) and human-specific *Bacteroides* (bottom panel) in surface water samples from six sites near Houston, TX for two months following Hurricane Harvey.

AgriLife SCSC also evaluated a droplet digital PCR (ddPCR) protocol for the human marker HF183. A major advantage of the ddPCR platform is not having to include a standard for the quantification of unknown samples. The HF183 ddPCR protocol described by Cao et al. (2015)

was evaluated using a synthetic standard of known copy number, synthesized by Integrated DNA Technologies (IDTDNA, Coralville, Iowa, USA) constructed from a portion of the 16S ribosomal ribonucleic acid (rRNA) gene of *Bacteroides dorei*. The ddPCR equipment used by AgriLife SCSC, generously made available by Texas A&M Institute for Genome Sciences and Society, in order of workflow included: 1) QX200 AutoDG (Droplet Generator) Instrument, 2) PX1 PCR Plate Sealer, 3) C1000 Touch™ Thermal Cycler with 96-Deep Well Reaction Module, 4) QX200 Droplet Reader and 5) QuantaSoft Software.

The results, shown in Table 11, indicate good sensitivity of the ddPCR method, successfully detecting ~5 target copies per µL. Furthermore, detection was somewhat linear with increasing target copies per reaction. However, ddPCR appeared to underestimate copy numbers, especially at the higher concentrations. For example, in the 50,500 copies/µL sample, the droplet reader estimated 5,570 copies/µL. Further refinement and validation of the protocol will be required before we can successfully use the approach in future BST projects.

Table 11. *Bacteroides dorei* HF183 gene marker spiked controls compared to the ddPCR output of actual quantification.

***B. dorei* HF183 gene marker expected outcome (gene copies/µL)**	**ddPCR actual quantification (gene copies/µl)**
50,500	5,570
5,050	375
505	19
5	2

droplet digital polymerase chain reaction, ddPCR; *Bacteroides dorei*, *B. dorei*; microliters, µl

AgriLife SCSC and UTSPH-EP continued to review the literature for publication of new markers that may be useful for expanding the Texas BST toolbox. One marker that appears especially promising is the human-specific H8 marker for *E. coli*, which codes for a sodium/hydrogen exchanger precursor (Hughes et al. 2017; Senkbeil et al. 2019). Senkbeil et al. (2019) determined the H8 marker was 92% specific and 100% sensitive to human isolates by validation with the conventional PCR method and the qPCR method using reference human and animal fecal samples, and later implementing qPCR on environmental samples. They also reported a strong correlation (R^2 = 0.89) between the H8 and HF183 qPCR assays. Initial plans for the next BST Infrastructure project are for AgriLife SCSC to selectively screen previously archived *E. coli* isolates for the H8 marker via endpoint-PCR to determine if sensitivity and specificity are similar to values reported in the literature and to investigate the use of H8 qPCR for analysis of water samples.

BST Program Outreach

Outreach regarding BST was a focus area of the project, which included presentations at conferences and meetings, a website redesign and maintenance of the Texas BST Library website.

Different aspects of the BST program were presented at three conferences. AgriLife SCSC presented a poster entitled "Hurricane Harvey Impacts on Fecal Indicator Bacteria Levels in Houston, TX Water Bodies" at the Water Microbiology Conference in Chapel Hill, NC on May 22-24, 2018. A presentation on bacteria source tracking was given at the Southern Region Water Conference in College Station from July 23-25, 2019. Another presentation providing an overview of BST was given at the Soil Science Society of America meetings in San Antonio on November 13, 2019. A seminar titled "Microbial Assessment of Water & Soil Quality: From Hurricanes to Cropping Systems" was held at Oklahoma State University on November 19, 2018. The seminar included an overview of the Texas BST Program. A presentation discussing BST activities in the State of Texas was also provided at the annual Texas State Soil and Water Conservation Board (TSSWCB) meeting held on October 29-30, 2018. AgriLife SCSC also gave a presentation on BST at the EPA Region VI Stormwater Conference in Denton, TX in July 2019. Individual and small group meetings were also held during the project's duration throughout Texas. Meetings and discussions about BST were held with the Tarrant Regional Water District, the Plum Creek Watershed Partnership, stakeholders of Sycamore Creek (facilitated by Atkins Global) and the North Central Texas Council of Governments.

TWRI hosted and maintained the Texas BST Library website. From Februrary 1, 2018 through March 31, 2020, there were 385 visits from 309 visitors (Figure 8). Of the 385 visits, 331 were from the United States and 217 were from Texas (predominantly College Station, Austin, Houston, Dallas and San Antonio). The Czech Republic was second to the United States in number of visits with 19. There were 713 page views, for a result of 1.85 pages per session. On average, users stayed on the site for 1 minute and 42 seconds. Peak visits occurred in the January 2019.

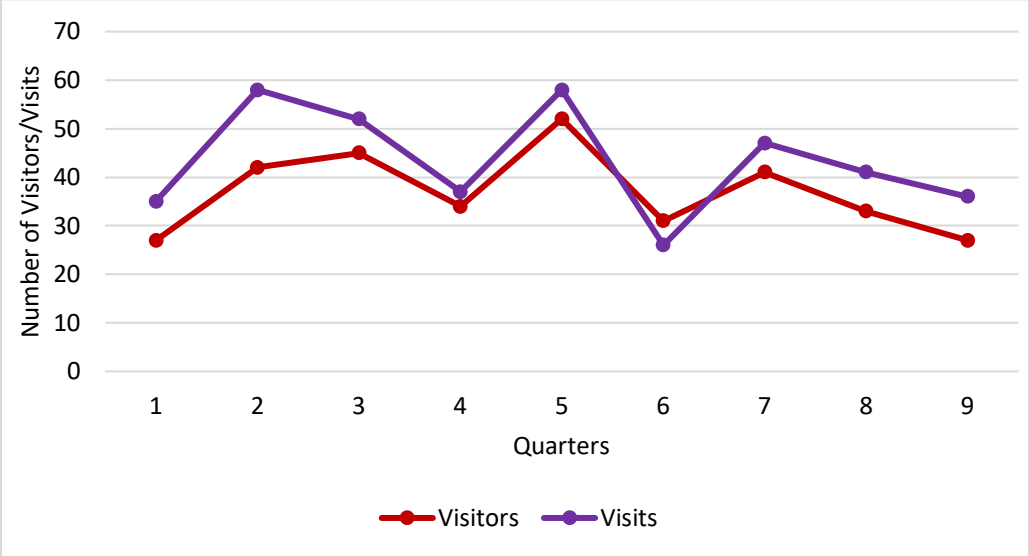

Figure 8. Number of visits and visitors to the Texas BST Program website from January 1, 2018 to March 31, 2020.

Literature Cited

Bernhard, A.E., Field, K.G. 2000. A PCR assay to discriminate human and ruminant feces based on host differences in Bacteroides-Prevotella 16S ribosomal DNA. Appl Environ Microbiol 66: 4571-4574.

Cao, Y., Raith, M.R., Griffith, J.F. 2015. Droplet digital PCR for simultaneous quantification of general and human-associated fecal indicators for water quality assessment. Water Research 70: 337-349

Hughes, B., Beale, D.J., Dennis, P.G., Cook, S., Ahmed, W. 2017. Cross-contamination of human wastewater-associated molecular markers in relation to fecal indicator bacteria and enteric viruses in recreational beach waters. Appl Environ Microbiol 83: e00028-17

Senkbeil, J.K., Ahmed, W., Conrad, J., Harwood, V.J. 2019. Use of *Escherichia* coli genes associated with human sewage to track fecal contamination source in subtropical waters. Science of the Total Environment 686: 1069-1075

Shanks, O.C., Kelty, C.A., Oshiro, R., Haugland, R.A., Madi, T., Brooks, L., Field, K.G., Sivaganesan, M. 2016. Data acceptance criteria for standardized human-associated fecal source identification quantitative real-time PCR methods. Appl Environ Microbiol 82: 2773–2782.

Shanks, O.C., Kelty, C.A., Sivaganesan, M., Varma, M., Haugland, R.A. 2009. Quantitative PCR for genetic markers of human fecal pollution. Appl Environ Microbiol 75: 5507–5513.

[TCEQ] Texas Commission on Environmental Quality. Chapter 307 – Texas Surface Water Quality Standards. 2018. https://www.tceq.texas.gov/assets/public/legal/rules/rules/pdflib/307.pdf. March 1, 2018