ESCHERICHIA COLI STRAIN DIVERSITY IN HUMANS: EFFECTS OF SAMPLING EFFORT AND METHODOLOGY

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Biological Sciences

by

Emily R. Neal

May, 2013

COMMITTEE MEMBERSHIP

TITLE:	*Escherichia coli* Strain Diversity in Humans: Effects of Sampling Effort and Methodology
AUTHOR:	Emily R. Neal
DATE SUBMITTED:	May, 2013
COMMITTEE CHAIR:	Dr. Michael Black, Professor Biological Sciences Department California Polytechnic State University- San Luis Obispo, CA
COMMITTEE MEMBER:	Dr. Christopher Kitts, Department Chair, Professor Biological Sciences Department California Polytechnic State University- San Luis Obispo, CA
COMMITTEE MEMBER:	Dr. Anya Goodman, Associate Professor Chemistry & Biochemistry Department California Polytechnic State University- San Luis Obispo, CA
COMMITTEE MEMBER:	Dr. Alexander Dekhtyar, Professor Computer Science Department California Polytechnic State University- San Luis Obispo, CA

ABSTRACT

Escherichia coli Strain Diversity in Humans:
Effects of Sampling Effort and Methodology

Emily R. Neal

Studies investigating *Escherichia coli* strain diversity and demographics in human hosts are frequently inconsistent regarding sampling effort and methodology while current strain typing methods are often expensive or laborious. To rectify these inconsistencies, sampling effort was investigated by comparing the diversity of 15-isolate collections to 100-isolate collections from 3 human subjects. Temporal variation in *E. coli* strain diversity was also studied by collecting 15 isolates once every 6 months. Additionally, strain identification and diversity collected by different sampling methods (fecal swabs vs. anal swabs collected at different times around defecation) were compared to identify any inherent biases in sampling method. This study employed pyroprinting, a new inexpensive and simple strain typing method using pyrosequencing, to generate DNA fingerprints (or pyroprints) based on the Intergenic Transcribed Spacer sequences in the ribosomal RNA operon to differentiate *E. coli* strains. Differences in strain diversity were apparent when comparing sampling efforts. The sampling effort investigation suggested that certain subjects hosted very large and highly diverse *E. coli* strain populations such that even 100 isolates may not fully represent *E. coli* strain populations in human hosts. Instead, the sampling effort required to accurately represent strain demographics may depend on strain richness and evenness within each host. The temporal investigation yielded similar or greater strain abundance and diversity compared to other typing methods in the literature suggesting pyroprinting is a similarly discriminating tool. When agglomerated over time or by subject, no significant differences in diversity were observed between subjects or between sampling methods despite visible differences in strain richness and evenness.

Keywords: *Escherichia coli*, strain diversity, sampling effort

ACKNOWLEDGMENTS

For Arley:
I will always love you.

The success and completion of this Master's thesis belongs truly to the faithfulness and continued sustenance provided by my Lord and God, Jesus Christ, to whom is owed all my praise.

Many thanks to my always encouraging family and supportive friends, but most of all to my husband. Joel, your unconditional love and continued selfless support through this uphill climb gave me strength to endure.

Dr. Black and Dr. Kitts, you pushed and challenged me to grow and think in new ways. I am ever grateful for your belief in me. Thank you for all your support and backing.

This thesis was made possible in part by the WM Keck Foundation, the Center for Applications in Biotechnology, and the College of Science and Mathematics Biological Sciences Department at California Polytechnic State University, San Luis Obispo.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Escherichia coli (*E. coli*) are generally commensal inhabitants of the mammalian and avian gastrointestinal tract (1, 2). Because they are shed in the host feces, *E. coli* are commonly used as fecal indicator bacteria (FIB) when found in environments such as soil, resource water, or recreational water (3–5). Government agencies like the USFDA, USEPA and state water quality control boards use *E. coli* counts to regulate food and water resources (5–7) because their presence indicates fecal contamination that can carry enteric pathogens (3, 4, 8). While pathogenic and antibiotic-resistant strains of *E. coli* are heavily investigated (9–11), relatively little is known about the ecology and strain dynamics of commensal *E. coli* within humans and other hosts of interest.

Because of its status as a FIB, researchers often use the *E. coli* species as a tool for tracking the source of fecal contamination (5, 12, 13). Microbial source tracking (MST) is a relatively recent and rapidly growing effort to find the sources of microbes associated with fecal contamination. The general approach for MST identifies a particular fecal-associated microbial species (or group of species) from the environment and matches them to a host species (e.g. cow, human, swine, etc.) or location, based often on host specific probes or a library-dependent set of diagnostic patterns from potential sources (5, 13). Using *E. coli* for MST is only effectual because of its typically commensal relationship with many host species. Characterizing the diversity and ecology

of commensal *E. coli* is important for the use of *E. coli* as a suitable MST species and for establishing the parameters (e.g. sample sizes necessary) for library-dependent MST methods.

Variations in microbial diversity and ecology in humans are often associated with changes in health. Several studies establish a relationship between the *E. coli* phylotypes and strains present and incidence of both Crohn's disease and inflammatory bowel disease (IBD) (14–17). Scanlan et al. (18) found significantly lower temporal stability for major fecal bacteria in Crohn's disease patients compared to healthy patients, suggesting that the temporal stability of other commensals (e.g. *E. coli*) may also be related to disease onset or progression. Therefore, commensal *E. coli* strain presence, abundance, and diversity may establish a healthy norm for medical comparisons, acting as a potential indicator for general stability of the intestinal microflora.

The strain diversity of *E. coli* in humans is commonly evaluated using a few isolates from many human subjects to provide a snapshot of diversity within a population (12, 19). Very few studies describe strain diversity and stability in individual hosts (12). Additionally, many studies only identify *E. coli* phylotypes (A, B1, B2, & D) (9, 19–24), which are broad phylogenetic groups encompassing many different strains. Therefore, more sensitive strain typing methods are required to track specific *E. coli* strains within a host population.

E. coli strains are commonly differentiated using both phenotypic and genotypic strain typing methods (5, 13). Unfortunately, the strain typing method used appears to influence the observed strain demographics and diversity in humans (25, 26) and the sampling effort required to accurately represent a population. In addition, studies

investigating temporal strain dynamics in individual hosts were inconsistent regarding collection methods, number of isolates collected, and strain typing method used (10, 25, 27, 28, 29, 30).

Early studies by Sears et al. (27, 28) indicate one or two strains persisting over at least a year while several less abundant strains were also detected. In these studies, isolates from several human subjects were obtained from feces over several years and strains were determined by antigen typing. The number of isolates collected varied with each sampling event.

Caugant et al. (29) also found stably maintained *E. coli* strains in human hosts. In this study, 550 *E. coli* isolates were collected from a single human subject and were grouped into 53 strains using electrophoretic typing of 15 enzymes. Over the course of 11 months, one *E. coli* strain persisted for 7 months and one strain persisted for 5 months. The number of isolates collected during each sampling event ranged from 11 to 74. While differences between sampling techniques were investigated, the different sample types were not taken on the same day, leaving time as a variable for change.

Anderson et al. (25) evaluated several human subjects over a seven month period and observed a higher degree of diversity than Sears et al. This group used two different typing methods to evaluate isolates taken from a single fecal sample from each subject from each subject per month. Antibiotic resistance analysis (ARA) was used to type 25 isolates while ribotyping was used to type 5 isolates. Ribotyping yielded less diversity than antibiotic resistance typing, though this could be an artifact of sampling effort. A maximum of 31 unique antibiotic resistance profiles and 11 unique ribotypes were observed in a single human host over the sampling period. Most strains detected by either

method did not persist for more than one month. However, one strain did persist for more than five months. Though diversity comparisons between subjects were not directly measured, the persistence results also seemed to indicate variable degrees of strain richness and diversity across human subjects.

Lastly, Johnson et al. (10) investigated *E. coli* persistence and sharing between six household members (including one dog) using rapid amplified polymorphic DNA (RAPD) analysis and pulse field gel electrophoresis (PFGE) to identify strains. Five isolates were collected from six fecal samples from each subject over a three year period. Only a total of 14 *E. coli* strains were identified in the entire three-year study from which extensive strain sharing was observed between all household members. Certain strains persisted in human subjects up to almost three years. In a similar study, Damborg et al. (30) investigated *E. coli* strain diversity in 18 human subjects over six months and reported strains shared within a household commonly persist in individuals for over six months. *E. coli* isolates were collected using either rectal or fecal swabs from subjects grouped in 8 households. Only a *maximum* of ten isolates collected over the entire six months from each subject were analyzed using amplified fragment length polymorphism (AFLP). In general, the investigation found high diversity over the six months detecting an average of six strains per ten *E. coli* isolates. The data presented did not report any impact of the different sampling methods used. Because subjects in both the Johnson and Damborg studies were members of the same households, strains could have been re-acquired between housemates rather than persisting in an individual.

Overall, the existing temporal studies in the literature review were inconsistent in the number of isolates collected per sampling event. Inconsistencies not only prevent

direct diversity comparisons between sampling events but may also lead to some samplings in which only dominant strains are detected and other samplings in which minor strains can also be detected. Additionally, many of these studies lacked evidence supporting the level of sampling effort used to investigate strain dynamics. The number of isolates necessary for representative sampling has not been investigated for any strain typing method to date. Studies in the literature collected isolates from either feces or anal-rectal swabs, but analysis of differences between sampling strategies was limited to the detection of a particular strain or strain characteristic, usually related to pathogenicity (31–34).

To rectify this lack of information regarding the diversity of commensal *E. coli* and clarify inconsistencies in the literature, we investigated *E. coli* strains using a new genotypic strain typing method called pyroprinting described by Black et al. (unpublished 35, 36). Fifteen isolates were collected from fecal samples once a month for six months. Sampling effort was also investigated by comparing collections of 100 *E. coli* isolates to collections of 15 isolates taken at the same time. In addition, this study is the first to analyze the effects of sampling method on observed commensal *E. coli* strain diversity and detection in humans.

CHAPTER 2

Methods and Materials

Study Design and Sampling

Samples were collected from three human subjects once a month for a total of six months. Subject B (male) was age 21-22 over the course of the study; subjects A and C (females) were ages 25 and 20-21, respectively. During one sampling event, different methods were used to collect *E. coli*: 1) an anal swab directly **before** defecation, 2) an anal swab **immediately** after defecation prior to wiping, 3) an anal swab several hours **later** post-defecation, and 4) a swab from the homogenized **fecal** sample collected for that day (Fig. 1).

Swabs were immediately streaked on MacConkey agar and taken to the lab for processing within 24 hours. Streaked plates were stored at 4°C if necessary prior to an overnight incubation at 37°C. Fifteen isolates from each original MacConkey plate were selected for further investigation. To address questions regarding sampling effort, 100 *E. coli* isolates were collected from homogenized fecal samples on month 1 and month 6 for each subject. Approximately three to five grams were diluted in sterile water to a concentration of 100 mg feces•ml^{-1}. Dilutions of 10^{-2}, 10^{-3}, and 10^{-4} mg•ml^{-1} were plated on MacConkey agar, and isolates were selected for confirmation.

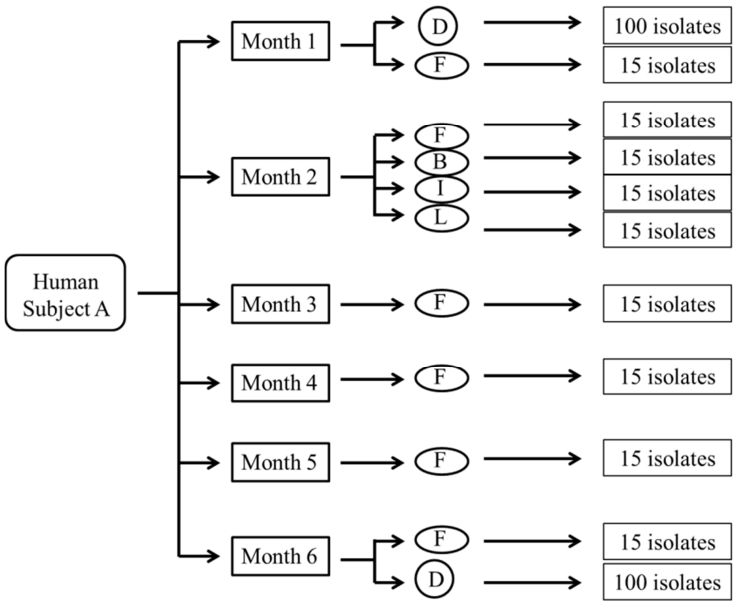

Figure 1. Sampling Plan for Subject A. Letters represent sampling method where D = diluted stool samples, F = fecal swab, B = before swab, I = immediate swab, and L = later swab. The sampling strategy described in month 2 for subject A was exactly the same for subjects B and C except the month selected for different sampling methods. Month 1 was selected for subject B, and month 4 was selected for subject C.

E. coli Confirmation.

MacConkey agar was used as the initial isolation step. To ensure pure cultures, pink colonies selected from the first MacConkey plate were streaked onto MacConkey a second time. From this second plate, pink colonies were streaked onto Luria-Bertani (LB) agar. Isolates on LB were confirmed as *E. coli* if they: 1) produced a green metallic sheen on eosin-methylene blue agar, 2) produced indole in tryptone broth, and 3) remained negative for growth on Simmons' citrate agar.

Pyroprinting

The method described by Black et al. (unpublished 35) was used for this study. Specifically, two separate colony PCR reactions amplified each of two genomic intergenic transcribed spacer regions (ITS1 and ITS2) using flanking primers described in Black et al (unpublished). Gel electrophoresis (2.5% agarose in Tris-acetic acid-EDTA buffer) confirmed PCR success. PCR products were processed for pyrosequencing with a Qiagen Pyromark® Q24 using dispensation sequences established in Black et al (unpublished 35). A pyroprint is defined as the ordered list of light-emission peak heights for each nucleotide dispensation in a sequencing reaction. Pyroprint data were exported from the Pyromark® software and collected for pairwise similarity analysis.

Statistics

Pyroprints for the same ITS regions were compared between different *E. coli* isolates using Pearson's correlation to assess similarity. For clarification, at the end of this analysis two correlation matrices were produced: one comparing pyroprints from ITS1 for all isolates and another comparing pyroprints from ITS2 for the same isolates. A novel algorithm was developed for clustering isolates into strains using correlation values between isolates from *both* ITS regions (35–37). For this study, the algorithm was customized such that subject and sampling time were considered when clustering (37). Isolates collected within a subject were clustered first at month 1, then month 2, and so

on. This process was completed for all three subjects, and only then were clusters compared across subjects.

The Shannon-Weaver index and Pielou's evenness were calculated for every sample type collected during each sampling event from each subject. Diversity indices were used to compare temporal diversity and strain diversity identified by the different sampling methods. One-way repeated measures analysis of variance was performed using Minitab® 16 to compare Shannon-Weaver values and evenness between subjects across all six months and between sampling methods (B, F, I, L) across all subjects.

The Freeman-Halton version of Fisher's exact test was used to compare the distribution of strains identified from the dilution samples and fecal swab samples collected together at the same times. This test was conducted six times using the FREQ procedure in SAS 9.2: one for each sampling event (month 1 and month 6) from all subjects (A, B, and C).

To compare the dilution sampling effort to the fecal swab sampling, rarefaction curves were generated for the same six data sets (described above), EstimateS (38) was used to calculate the number of strains observed (S_{obs}) with the *Mao Tau* analytical method. Data were exported and graphed using Microsoft® Excel 2010.

To assess similarity of sampling methods the frequency of detection across all sampling methods in a given strain was compared using binary logistic regression. Data from each subject were analyzed separately in Minitab® 16, and binary logistic regression was applied for each strain.

CHAPTER 3

Results

Effects of Sampling Effort

Strain distributions from dilutions (100 isolates) and fecal swabs (15 isolates) were compared for month 1 and month 6 in all 3 subjects (Fig. 2). Differences in strain distribution across sampling efforts (100 vs. 15 isolates collected) were assessed using Fisher's exact tests. No significant differences in strain distribution were found between the samples collected from subject A (both month 1 and month 6) and between the samples from subject B at month 6 (P values > 0.05). All other comparisons showed a significant difference in the distribution of strains between the two levels of sampling effort.

Rarefaction curve trajectories measuring number of strains sampled using fecal swabs more closely followed curve trajectories generated from the dilution samples (Fig. 3C, E, F) when the distributions of strains between the two were found to be significantly different. Alternatively, when the distribution of strains was *not* found to be significantly different, the fecal swab rarefaction curves displayed greater curvature than the dilution sample rarefaction curves (Fig. 3A, B, D). In these cases, the fecal swabs detected fewer strains than dilutions for the same number of collected isolates. For example, the fecal swab from subject B month 6 was predicted to detect a total of 6 strains using 13 isolates

whereas the dilution would detect 10 strains using 13 isolates (Fig. 3D). Although dilution rarefaction curves generated from subject A month 6 (Fig. 3B), subject B month 1 (Fig. 3C), and subject C month 1 (Fig. 3E) displayed less curvature than the other dilution rarefaction curves, none of the curves for either sampling type reached an asymptote for number of identified strains indicating that neither 15 nor 100 isolates were sufficient to detect the majority of strains present in the fecal samples.

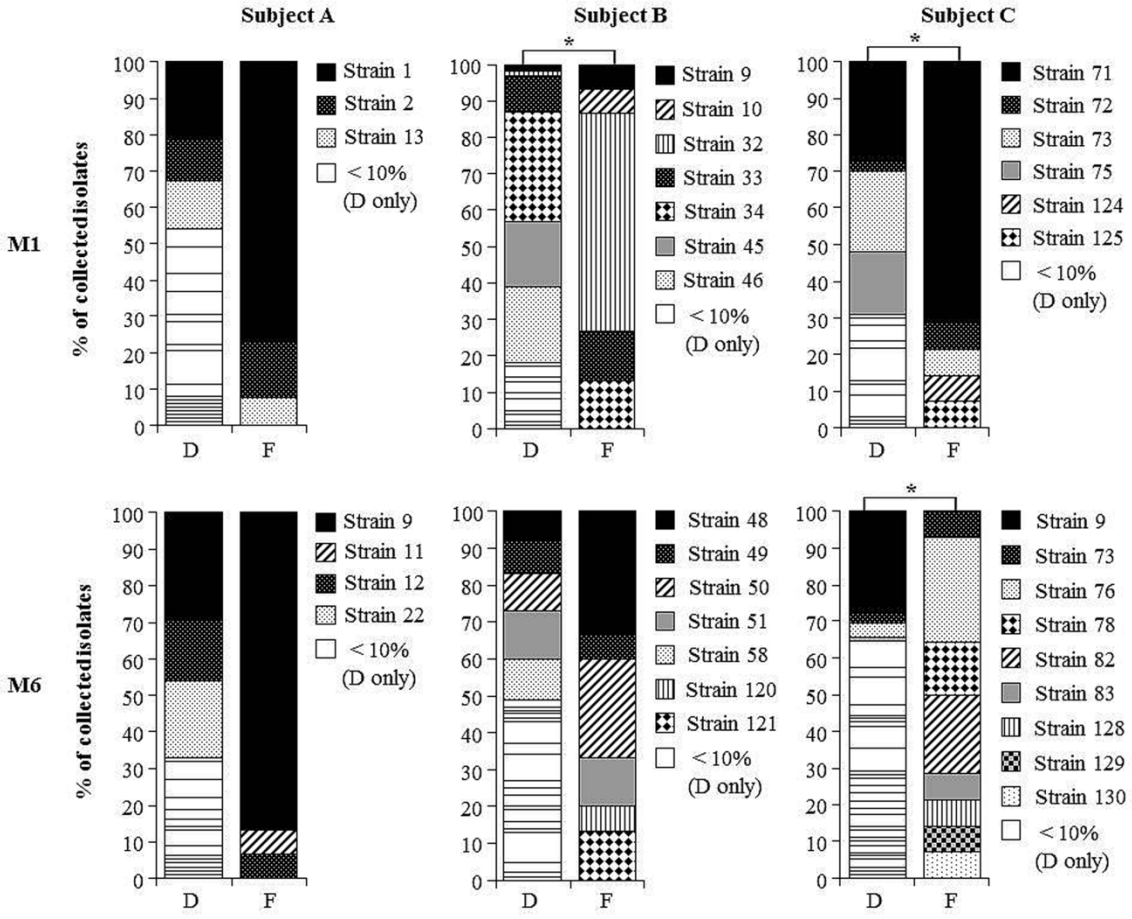

Figure 2. Sampling Effort Distribution Comparisons. Percentage of total isolates collected from a particular sampling method (stool dilution or fecal swab) represented by each strain detected for all three human subjects A, B, & C from month 1 (M1) or month 6 (M6). Strains detected at less than 10% of the total isolates collected from the dilution were categorized together in each figure unless the strain was detected again from the fecal swabs (e.g. strain 32 from subject B M1). * indicates significant differences ($P < 0.05$).

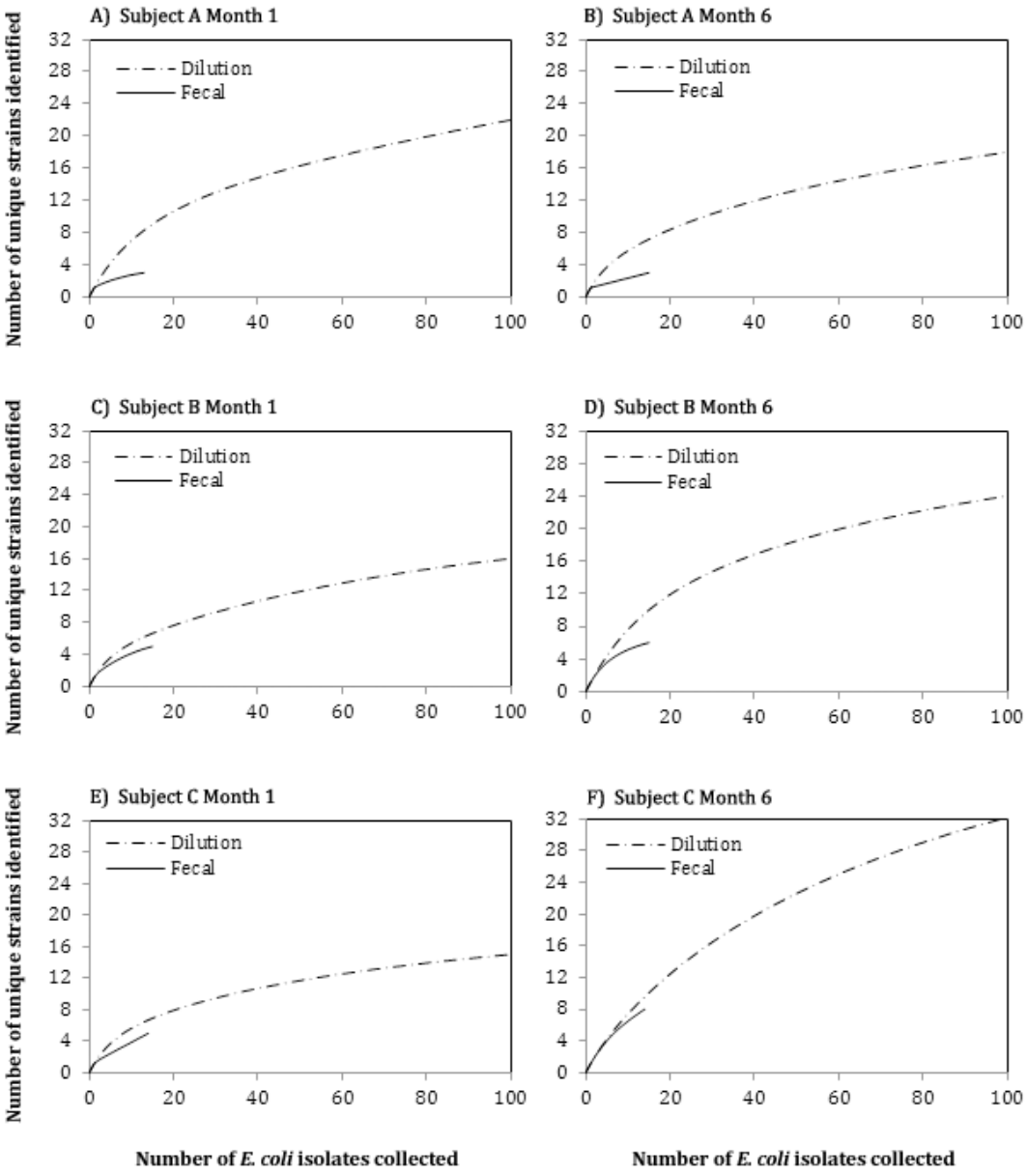

Figure 3. Sampling Effort Rarefaction Curves. Rarefaction curves were generated using the S_{obs} calculated with the Mao Tao method for each sampling effort event conducted. All subjects were sampled at month 1 and month 6. Different numbers of *E. coli* were isolated using the two methods: Dilution (100 isolates) and Fecal (15 isolates).

In general, a majority of isolates collected from fecal swabs (15 isolates total) represented strains that were also isolated from fecal dilutions (100 isolates). In the best case, 100% of the isolates from subject A at month 1 were represented in the

corresponding dilution sample, while in the worst case only 38% of the isolates from subject C at month 6 were represented in the dilution sample (Table 1). Perhaps unsurprisingly, these two extremes in representation across sampling effort correlate to extremes in strain diversity (Fig. 2).

Table 1. *Sampling Effort Strain Representation*
Number of strains detected from both fecal swabs and dilution samples from Months 1 and 6 and percent of strains from fecal swabs represented in the dilution sample.

	Month 1			Month 6		
Subject	No. of F strains	No. of D strains	No. of F strains in D (%)	No. of F strains	No. of D strains	No. of F strains in D (%)
A	3	20	3 (100.0)	3	18	2 (66.7)
B	5	16	4 (80.0)	6	24	4 (66.7)
C	5	15	3 (60.0)	8	32	3 (37.5)

Letters represent subject (A-C) or sample type (F and D) as described in Methods and Materials.
Table displays both the number of strains and percentages of strains detected using fecal swabs that were also detected in dilution samples.

Effects of Sampling Method

Average Shannon-Weaver values for each sampling method were calculated using data from all three subjects and compared using repeated measures ANOVA. No significant differences were observed between sampling methods (Fig. 4A). The test was repeated using Pielou's evenness values, and again no significant differences were observed between mean evenness values of each sampling method (Fig. 4B). However, large variability around the means was observed except in the evenness of strains from swabs taken before defecation.

Data from different swab samples collected during a single sample month (Fig. 1) were analyzed using binary logistic regression to test the hypothesis that sampling method does not affect strain detection. If this were true then each strain would have an

equal chance of detection by each sampling method, that is, a particular strain should not be represented by isolates from predominantly one sampling method. Binary logistic regression requires at least three isolates in each strain to determine significant differences ($P < 0.05$); therefore not all strains detected by the sampling methods could be analyzed.

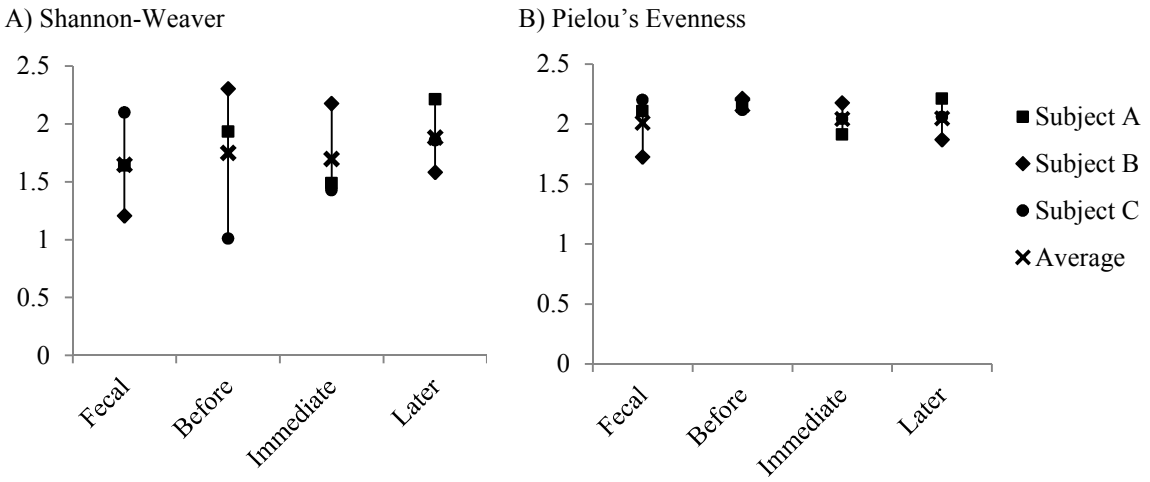

Figure 4. Sampling Method Shannon-Weaver and Pielou's Evenness. A) Shannon-Weaver values from each subject and the average values calculated using all three subjects for every sampling method. No significant differences were found between the averages ($P > 0.05$). B) Pielou's evenness values calculated for every sampling method from each subject and the average values calculated using all three subjects. No significant differences were found between the sample types ($P > 0.05$)

Between two and six strains from each subject were detected in significantly higher proportions by one or two sampling methods, but the proportions of strains with three or more isolates varied by subject (Fig. 5). Seven of the 16 strains collected from subject A were represented by 3 or more isolates while only 2 of these (29%) were detected in significantly different proportions by sampling method. In subject B, 9 out of 22 strains were represented by more than 3 isolates, and 4 of these (44%) were detected in significantly different proportions. Subject C held the extreme with 6 out of 24 strains

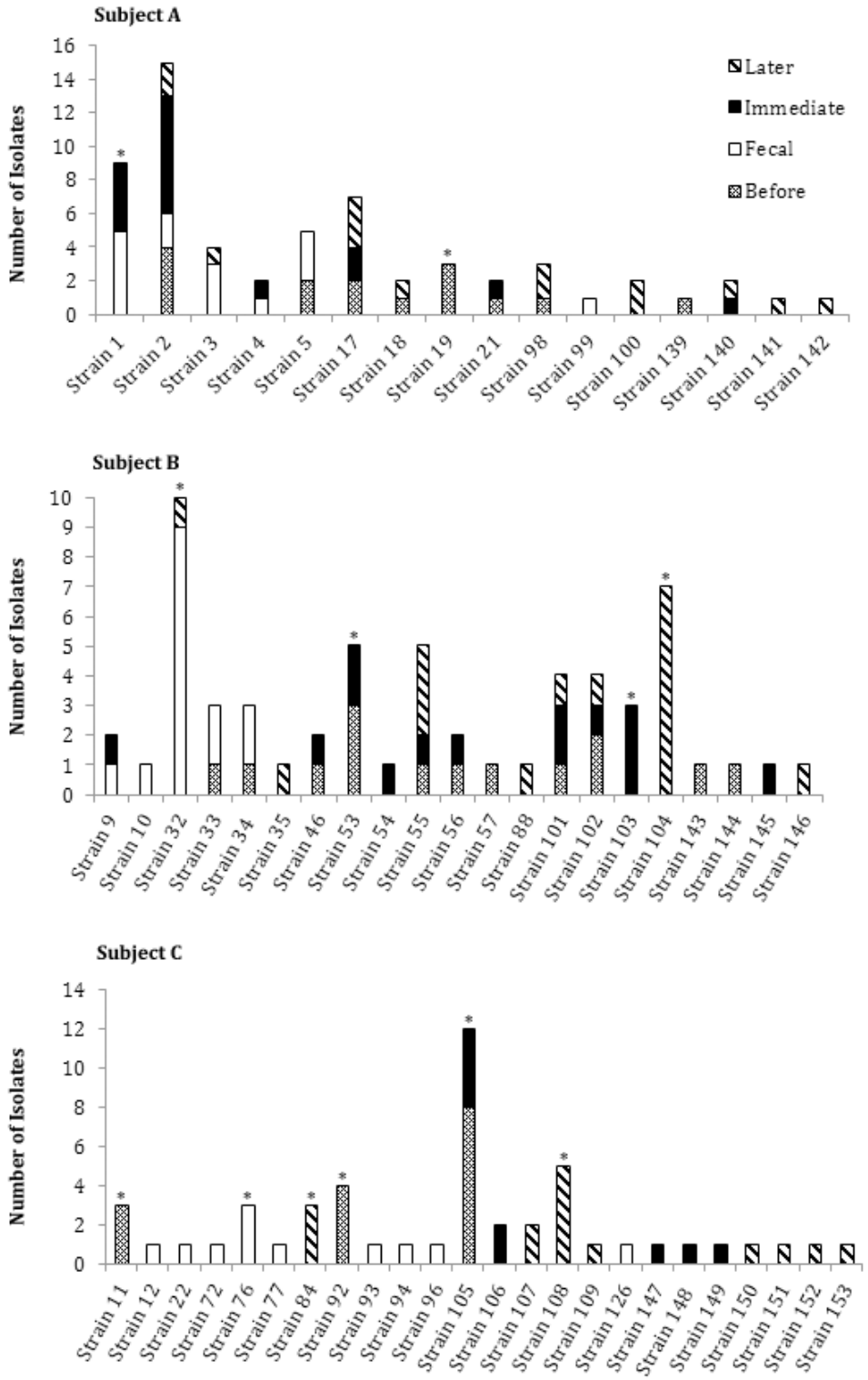

Figure 5. Strains Identified by Sampling Method. Strains identified using the four different sampling methods and analyzed using binary logistic regression. * identifies strains with a p-value less than $\alpha = 0.05$, indicating sampling methods used to detect a particular strain were disproportionately represented.

represented by more than 3 isolates, and all 6 (100%) were detected in significantly different proportions by sampling method.

Temporal strain diversity and stability

Isolates collected from fecal swabs over six months were used for temporal comparisons (Fig. 6 & 7). Subject A had the fewest total strains isolated over the study. Subjects B and C had more similar numbers of strains detected, although the largest number of strains were isolated from C (Table 2).

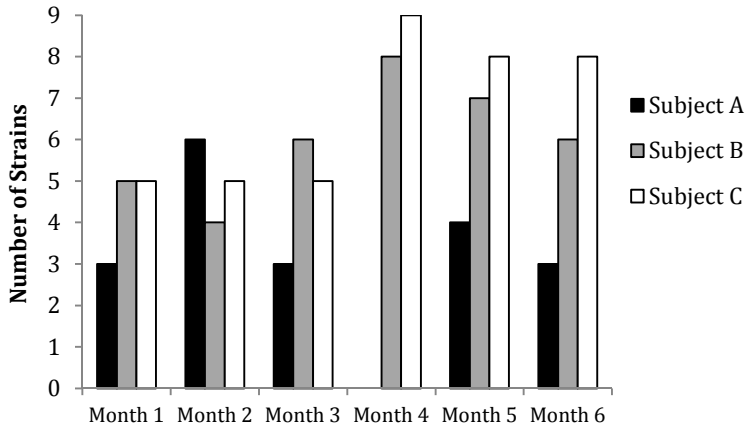

Figure 6. Numbers of Strain Detected Monthly. Number of strains detected from each subject in each of six months using F swab samples.

To compare temporal diversity between subjects, Shannon-Weaver index and Pielou's evenness values were calculated for each sampling event and then averaged for each subject (excluding month 4 for subject A). Although no significant differences were observed between subjects ($P > 0.05$), both diversity indices were lowest from subject A suggesting less strain diversity, while subjects B and C both had higher and more similar values of diversity (Table 2).

Table 2. *Temporal Strain Diversity for Each Subject*
Average strain detection and diversity measurements over the six month study period

Subject (total strains detected)	Average number of strains detected per month (Min-Max)	Shannon index	Pielou's evenness
A (15)	3.17* (≤1-6)	0.93 (± 0.44) a	0.69 (± 0.17) a
B (25)	6 (4-8)	1.56 (± 0.32) a	0.88 (± 0.08) a
C (29)	6.67 (5-9)	1.59 (± 0.42) a	0.84 (± 0.12) a

*Average for subject A calculated using all 6 months, including month 4 in which zero *E. coli* were detected. As this is a descriptive study and the subject did not host *E. coli* at detectable levels, the study aims to present realistic observations which could be applicable to other normal human individuals/populations. The average number of strain collected during months only in which *E. coli* were collected was 3.8.

Within all three subjects, the majority of strains were detected only once during the six-month study period (Fig. 7 & 8). No strains were detected in more than two sampling events from subject A, indicating low strain stability. Low strain stability was also observed in subject C, who had the highest percentage of strains detected in only one sample (Fig. 8). Interestingly, subject C also had the highest percentage of strains detected in four out of six samples, which indicates high strain stability. Strains detected multiple times, however, were not necessarily detected in consecutive months. For example, strain 33 was detected in months 1, 2, 4, and 5 from subject B but did not appear in month 3 (Fig. 7B). In month 4, no *E. coli* were detected from subject A (Fig. 6). Colonies appearing on MacConkey agar from that month were identified as *Salmonella sp.* after analysis using API-20E® strips (data not shown).

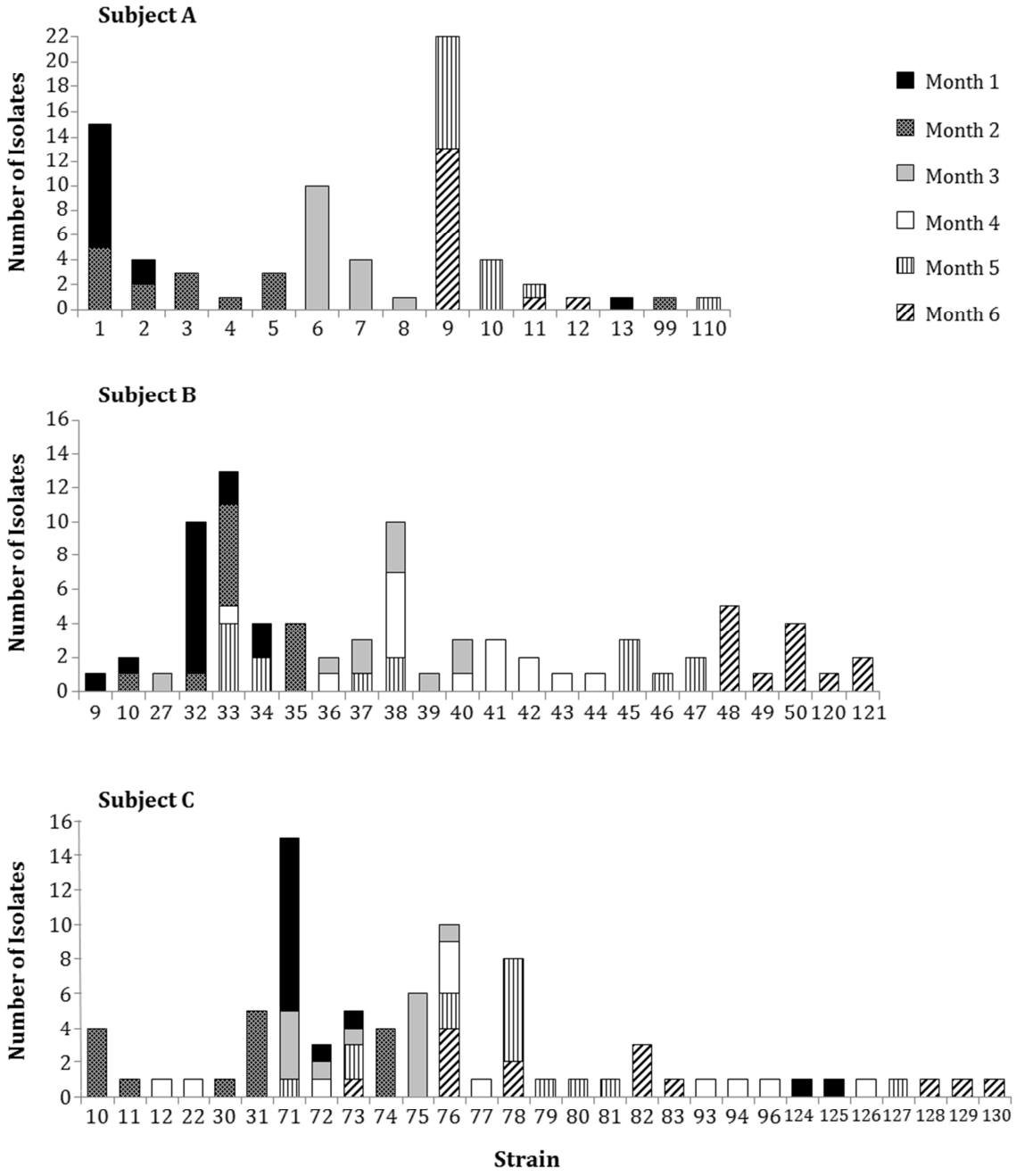

Figure 7. Temporal Strain Persistence. Persistence of each strain detected using F sampling from each subject during the six month study.

No formal log was kept regarding subjects' diets and daily health, but major events were recorded. Most notably, subject A traveled between the month 2 and month 3 collections and experienced a day of diarrhea approximately two weeks before the month

4 collection. These events coincide with changes in the *E. coli* strains detected and a lack of strain persistence in subject A. Two strains were detected from both month 1 and month 2 collections, but after traveling, three different strains appeared at month 3. After the intestinal disruption in month 4, two new strains appeared and persisted through months 5 and 6 (Fig. 7A). Subject B reported consuming antibiotics several weeks prior to the first collection date and experienced gastroenteritis that delayed collection from month 4. However no obvious impacts to stability were apparent as half of the strains detected in month 3 were also detected in month 4. Subject C recorded no major health changes but did note that collections during months 4-6 occurred immediately following menstruation. This coincided with an increase in the number of strains detected during those three months (Fig. 7C).

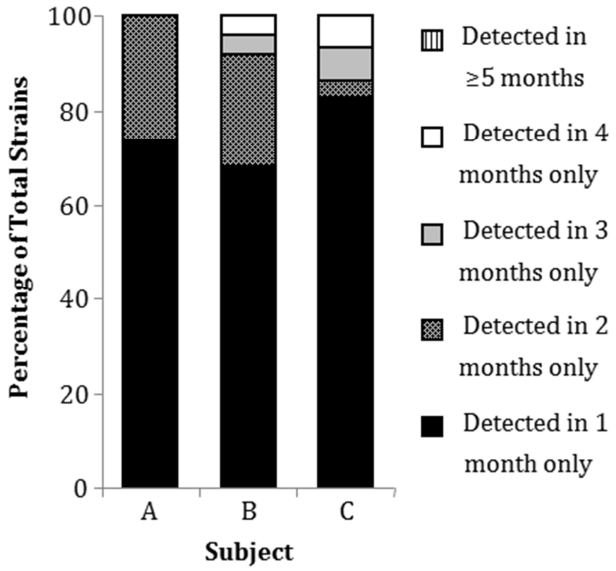

Figure 8. Frequency of Strain Detection. Frequency with which strains were detected for each subject. Data are represented as percentages of the total number of strains detected over six months from F swab samples and the frequency with which a certain percentage of strains were detected. Detecting a strain more than once does not necessarily imply consecutive detection. For example, strain 34 isolated from subject C was detected in month 1 and only again in month 5.

E. coli Abundance and Strain Distribution among Subjects

The number of strains isolated and characterized was a small fraction of the total *E. coli* present in each subject. Fecal samples taken at months 1 and 6 produced counts of *E. coli* around 10^7- 10^8 CFU•g^{-1} feces from all 3 subjects. These concentrations are consistent with other studies (2, 39). Therefore, the largest samplings (100 isolates at months 1 and 6) represent only 0.001% of the total *E. coli* in 1 gram of feces. Over the course of the six-month study, a total of 962 *E. coli* isolates were collected, confirmed, pyroprinted and found to comprise 153 different strains. Subject A hosted the fewest strains, followed in increasing order by subjects B and C (Table 3). The majority of strains were unique to 1 subject while 14 (9.2%) were shared by at least 2 subjects and only 4 strains (2.6%) were shared by all 3 subjects. Subjects A and C shared the most strains (Tables 3 and 4). The most dominant strain (that is, the strain comprised of the most *E. coli* isolates among the isolates sampled) also was shared by all 3 subjects (strain 9, Table 4). No other correlation between strain dominance and sharing was apparent in the data.

Table 3. *Quantity of Detected Strains from All Collected E. coli per Subject*

Subject	Total no. of strains	No. of unique strains*	Percentage shared strains**
A	50	38	24.0
B	56	49	12.5
C	66	53	19.7

*Unique strains are defined in this case as strains detected from only one subject or, not shared by subjects
**Percentage includes strains shared by one or two other subjects divided by the total number of strains detected from that subject.

Table 4. *Shared Strains between Subjects*
Number of *E. coli* strains shared between subjects and number of contributing isolates* from each subject.

	Shared Strains	Number of Isolates by Subject			Total Isolates
		A	B	C	
Two Subjects	Strain 11	2	-	7	9
	Strain 12	18	-	8	26
	Strain 29	4	-	3	7
	Strain 30	3	-	2	5
	Strain 31	1	-	5	6
	Strain 97	1	-	1	2
	Strain 110	1	-	2	3
	Strain 27	2	1	0	3
	Strain 54	-	2	1	3
	Strain 88	-	1	3	4
Three Subjects	Strain 9	51	4	27	82
	Strain 10	5	2	11	18
	Strain 22	21	1	2	24
	Strain 39	1	1	1	3
Total	14 Strains	110	12	73	195

*The number of contributing isolates represents the total number of isolates from the indicated subject contributing to all strains shared by subjects indicated in column

CHAPTER 4

Discussion

Pyroprinting as a Strain Typing Method

A bacterial strain is defined based on the typing method used to differentiate isolates beyond the species level (12). For example, *E. coli* strains were originally defined as isolates with the same surface antigens identified through serological typing (27, 40). However, worldwide collections of enterotoxigenic *E. coli* are differentiated by only a few serotypes. As newer, more discriminating typing methods were introduced, the *E. coli* strains they defined were better differentiated. As a result, literature estimates of strain diversity within a host have generally increased over time (12). The current gold standard for strain-level analysis of *E. coli* is multi-locus sequence typing (MLST) where gene sequences from 7 different housekeeping genes are compared together (41). MLST can differentiate between strains derived from the same parent clone over very short evolutionary periods (41). However, this method is neither rapid nor cheap, making the analysis of large numbers of isolates difficult. This is the first large-scale use of a new method for typing *E. coli* strains in human hosts and a comparison to existing literature values for strain diversity and richness serves to calibrate pyroprinting in relation to other methods.

E. coli strain diversity and richness among this study's subjects parallel results from previous studies using single enzyme ribotyping, ARA, and AFLP strain typing methods. Anderson et al. (25) generated average strain accumulation curves from five human subjects using both ribotyping and ARA to analyze 45 and 15 collected isolates, respectively. Anderson et al. detected approximately 3 different ARA-defined strains using 45 isolates and approximately 2 different ribotype-defined strains using 15 isolates. Rarefaction curves generated in our study indicate pyroprinting can discriminate more strains in both small 15 isolate sampling efforts (3-8 strains) and large 100 isolates sampling efforts (15-32 strains). Over 7 months, Anderson et al. identified between 3 and 11 strains per subject with ribotyping (analyzing 35 total isolates per subject) and between 21 and 33 strains per subject with ARA (analyzing 175 total isolates per subject). Over 6 months, Damborg et al. (30) identified between 1 and 10 strains per subject using AFLP, although only 10 isolates were collected from each subject over the entire length of the study. Over 6 months, our study identified between 15 and 29 strains per subject using pyroprinting to analyze 90 isolates per subject. The results between similar sampling efforts and temporal analyses from this study and the others mentioned provide evidence that pyroprinting may be more discriminating than single enzyme ribotyping and ARA. The sampling effort in Damborg et al. is not sufficient to compare pyroprinting to AFLP.

Effect of Sampling Effort

One hundred isolates were sampled twice from all three subjects with the assumption that this sampling effort would accurately represent the *E. coli* strains present at the time of sampling. Other studies using genotypic methods suggest sampling 15 isolates would represent a majority of strain diversity actually present in the host (25, 26). Lautenbach et al. (26) suggested strains comprising only 5% of the population are likely to be sampled 50% of the time if 15 isolates are collected and analyzed using PFGE. However, we observed significant differences in the distribution of strains, as well as differences in strain representation, between the 100 isolate dilution samples and the 15 isolate fecal swabs. In some cases, strains detected using 15 isolates were not detected when 100 isolates were collected (Fig. 2). These results indicate a large and diverse collection of *E. coli* strains is present in human hosts. We sampled at most 0.001% of the total *E. coli* population present in one gram of feces. A human may host many *E. coli* strains (high strain richness) in relatively equal proportions (high evenness). Under these circumstances many strains will have equal probability for detection, requiring a relatively large sample size to accurately represent the diversity within a population of *E. coli* strains. Alternatively, low strain richness and/or evenness would allow a smaller sample size to accurately represent strain diversity within the host. If richness is high but evenness is very low in the population, a small sample size will likely represent the most abundant strains but will miss detection of the more rare strains.

This interplay between richness, evenness, and diversity can be observed in the strain distribution graphs (Fig. 2) and rarefaction curves (Fig. 3). No significant

differences were found between the distributions of strains in three cases (Fig. 2 AM1, AM6, BM6); yet in these cases the rarefaction curves for the two levels of sampling effort deviate in curvature (Fig. 3A, B, D). This is consistent with an *E. coli* strain population of high richness and low evenness where the most abundant strains are easily detected in both sampling efforts (Fig. 2). In addition, this high richness, low evenness scenario implies a large number of rare strains are more likely detected with a large sampling effort while being missed with a small sampling effort. Thus, the small sample size displays greater curvature than the large sample size curve (Fig. 3A, B, D). Conversely, a high richness, high evenness scenario will result in different distributions (Fig. 2 BM1, CM1, CM6) yet similar rarefaction curves when comparing small and large sampling efforts because all major strains are equally likely to be detected. In concordance, both rarefaction curves display similar curvature and indicate a much larger sample size would be necessary to accurately determine strain richness (Fig. 3C, E, F).

Very few studies collect 100 or more *E. coli* isolates from a single human subject, yet many still draw conclusions regarding the diversity of *E. coli* strains within a given human host. The Anderson et al. (25) accumulation curves displayed high curvature suggesting that the sample sizes selected for ARA and ribotyping in that study adequately represented actual strain richness. However, rarefaction curves from subject A in our study displayed similar curvature for smaller 15 isolate samples sizes as the ARA (45 isolates) and ribotype (15 isolates) curves yet lower curvature using larger 100 isolate samples. Thus, subjects in the Anderson et al. investigation may have hosted strains with low evenness such that rare strains were not detected using only 15 isolates collected for ribotyping and increased strain diversity might have been observed with a larger

sampling effort. Additionally, the high curvature appearinging with a greater number of isolates collected for ARA also suggests the subjects hosted strains with low evenness. It is also possible that ARA ultimately is not a very discriminating strain differentiation method. Unfortunately, since *E. coli* strain diversity can vary between human subjects, the sample size necessary to accurately detect and represent *E. coli* strains may depend on the individual subject. Thus, larger sample sizes of collected *E. coli* isolates from multiple hosts are necessary to accurately estimate richness and diversity. This poses a dilemma for the design and scope of diversity investigations and will likely impact estimates of sharing and persistence among hosts as well as comparisons of sampling methods.

Effects of Sampling Method

The four sampling methods tested in this study yielded significantly different proportions of some *E. coli* strains from the same subject when 15 strains were analyzed from each sample (Fig. 5). This could be a result of inherent biases in the sampling methods. For example, anal swabs might over-sample *E. coli* strains colonizing locations near the sampling site while strains collected from the feces could be colonizing sites along the entire length of the colon. The timing of anal swab sampling with respect to defecation might also affect strain representation in a sample since fecal material may deposit new strains near the sampling site, which could later die off. However, since fecal samples were homogenized before swabbing or dilution, position within a fecal discharge should not have affected strain representation in these samples.

Alternatively, a highly rich and evenly distributed assemblage of strains could produce the observed results by random chance due to insufficient sampling effort. For example, more richness and high evenness was measured from subject C, and almost no strains were detected by more than one sampling method (Fig. 5C) possibly due to a similar detection probability of the many strains present in this host. In contrast, the least diverse and least even *E. coli* population was measured from subject A. This may explain why, even though only 15 isolates were collected using each sampling method, subject A was the only host from which a strain was detected using all four methods (Fig. 5A). Because sampling effort may have confounded the methods investigation, no clear evidence exists for differences between sampling methods and no conclusions can be made.

Sampling method did not significantly impact diversity or evenness measured across the subjects. However, average diversity and evenness values contained large variability except in the evenness of strains from swabs taken before defecation. Additionally, only three subjects participated in this study. Because of the small sample size and large variability that may have masked true differences, we cannot conclude that sampling method had no effect on detected strain diversity or evenness.

Sharing of E. coli Strains between Hosts

Little strain sharing was observed between the subjects in this study even though strain diversity was high within individuals. However, the host sample size in this study was small and may not represent the actual amount of strain sharing in a human population. Future studies should be conducted using pyroprinting to collect *E. coli* from larger human populations. If humans do not share many *E. coli* strains with each other this would suggest that either *E. coli* are not ideal candidates for microbial source tracking or very large databases are needed for library-dependent MST to track human contributions to *E. coli* strains present in the environment.

Strain Diversity and Stability

Although differences in diversity were apparent in the subjects over the course of the study, they were not statistically significant based on the diversity indices used to characterize the *E. coli* strain assemblages. Again, average diversity and evenness values contained large variability, possibly due to sampling effort or variation in strain stability between sampling events, that may have masked true differences in diversity.

Though it is difficult to relate the specific results of this study to concrete conclusions regarding subject health and temporal strain stability, the results seem to compliment other studies regarding the stability of the intestinal microflora from unhealthy patients since the subject reporting more intestinal unrest also exhibited the least temporal stability of *E. coli* strains. Additional use of pyroprinting may continue to

lend new and supportive evidence for using *E. coli* as an indicator for general intestinal health. *E. coli* of the B2 phylogroup are more frequently associated with intestinal disorders such as CD and IBD (14–17). Since the observations of this study correlate host environment, travel, and menstruation with changes in both strain diversity and temporal stability, observed microflora diversity should be considered along with other host variables before making generalized conclusions regarding human health.

LIST OF REFERENCES

1. **Shulman ST, Friedmann HC, Sims RH**. 2007. Theodor Escherich: the first pediatric infectious diseases physician? Clin. Infect. Dis. **45**:1025 –1029.

2. **Tenaillon O, Skurnik D, Picard B, Denamur E**. 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. **8**:207–217.

3. **Ishii S, Sadowsky MJ**. 2008. *Escherichia coli* in the environment: implications for water quality and human health. Microbes Environ. **23**:101–108.

4. **Meays CL, Broersma K, Nordin R, Mazumder A**. 2004. Source tracking fecal bacteria in water: a critical review of current methods. J. Environ. Manage. **73**:71–79.

5. **Simpson JM, Santo Domingo JW, Reasoner DJ**. 2002. Microbial Source Tracking: state of the science. Environ. Sci. Technol. **36**:5279–5288.

6. **Feng P, Weagant SD, Grant MA**. September 2002, posting date. Chapter 4, Enumeration of *Escherichia coli* and the coliform bacteria. In Bacteriological Analytical Manual. FDA. http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/ucm064948.htm.

7. **USEPA**. 2002. Method 1603: *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* agar (modified mTEC). USEPA, Washington, DC. http://www.epa.gov/nerlcwww/documents/1603sp02.pdf

8. **Casarez EA, Pillai SD, Di Giovanni GD**. 2007. Genotype diversity of *Escherichia coli* isolates in natural waters determined by PFGE and ERIC-PCR. Water Res. **41**:3643–3648.

9. **Bailey JK, Pinyon JL, Anantham S, Hall RM**. 2010. Commensal *Escherichia coli* of healthy humans: a reservoir for antibiotic-resistance determinants. J. Med. Microbiol. **59**:1331–1339.

10. **Johnson JR, Clabots C, Kuskowski MA**. 2008. Multiple-host sharing, long-term persistence, and virulence of *Escherichia coli* clones from human and animal household members. J. Clin. Microbiol. **46**:4078–4082.

11. **Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O**. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. **24**:2373–2384.
12. **Gordon D**. 2010. Strain typing and the ecological structure of *Escherichia coli*. J. AOAC Int. **93**:974–984.

13. **Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J**. 2002. Microbial source tracking: current methodology and future directions. Appl. Environ. Microbiol. **68**:5796–5803.

14. **Baumgart M, Dogan B, Rishniw M, Weitzman G, Bosworth B, Yantiss R, Orsi RH, Wiedmann M, McDonough P, Kim SG, Berg D, Schukken Y, Scherl E, Simpson KW**. 2007. Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of *Clostridiales* in Crohn's disease involving the ileum. ISME J. **1**:403–418.

15. **Kotlowski R, Bernstein CN, Sepehri S, Krause DO**. 2007. High prevalence of *Escherichia coli* belonging to the B2+D phylogenetic group in inflammatory bowel disease. Gut. **56**:669–675.

16. **Martinez-Medina M, Aldeguer X, Lopez-Siles M, González-Huix F, López-Oliu C, Dahbi G, Blanco JE, Blanco J, Garcia-Gil LJ, Darfeuille-Michaud A**. 2009. Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. Inflamm. Bowel Dis. **15**:872–882.

17. **Petersen AM, Nielsen EM, Litrup E, Brynskov J, Mirsepasi H, Krogfelt KA**. 2009. A phylogenetic group of *Escherichia coli* associated with active left-sided inflammatory bowel disease. BMC Microbiol. **9**:171.

18. **Scanlan PD, Shanahan F, O'Mahony C, Marchesi JR**. 2006. Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. J. Clin. Microbiol. **44**:3980–3988.

19. **Escobar-Páramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E**. 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ. Microbiol. **8**:1975–1984.

20. **Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A, Elion J, Picard B, Denamur E**. 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. Microbiol. **147**:1671–1676.

21. **Escobar-Páramo P, Grenet K, Le Menac'h A, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy MC, Andremont A, Denamur E, Ruimy R**. 2004. Large-scale population structure of human commensal *Escherichia coli* isolates. Appl. Environ. Microbiol. **70**:5698–5700.

22. **Gordon DM, Bauer S, Johnson JR**. 2002. The genetic structure of *Escherichia coli* populations in primary and secondary habitats. Microbiol. **148**:1513 –1522.

23. **Gordon DM, Stern SE, Collignon PJ**. 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. Microbiol. **151**:15–23.

24. **Gordon DM, Clermont O, Tolley H, Denamur E**. 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ. Microbiol. **10**:2484–2496.

25. **Anderson MA, Whitlock JE, Harwood VJ**. 2006. Diversity and distribution of *Escherichia coli* genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. Appl. Environ. Microbiol. **72**:6914–6922.

26. **Lautenbach E, Bilker WB, Tolomeo P, Maslow JN**. 2008. Impact of diversity of colonizing strains on strategies for sampling *Escherichia coli* from fecal specimens. J. Clin. Microbiol. **46**:3094–3096.

27. **Sears HJ, Brownlee I, Uchiyama JK**. 1950. Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. J. Bacteriol. **59**:293–301.

28. **Sears HJ, Brownlee I**. 1952. Further observations on the persistence of individual strains of *Escherchia coli* in the intestinal tract of man. J. Bacteriol. **63**:47–57.

29. **Caugant DA, Levin BR, Selander RK**. 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics. **98**:467.

30. **Damborg P, Nielsen SS, Guardabassi L**. 2009. *Escherichia coli* shedding patterns in humans and dogs: insights into within-household transmission of phylotypes associated with urinary tract infections. Epidemiol. Infect. **137**:1457–1464.

31. **Greenquist MA, Drouillard JS, Sargeant JM, Depenbusch BE, Shi X, Lechtenberg KF, Nagaraja TG**. 2005. Comparison of rectoanal mucosal swab cultures and fecal cultures for determining prevalence of *Escherichia coli* O157:H7 in feedlot cattle. Appl. Environ. Microbiol. **71**:6431–6433.

32. **Johnson JR, Scheutz F, Ulleryd P, Kuskowski MA, O'Bryan TT, Sandberg T**. 2005. Phylogenetic and pathotypic comparison of concurrent urine and rectal *Escherichia coli* isolates from men with febrile urinary tract infection. J. Clin. Microbiol. **43**:3895–3900.

33. **Khaitsa ML, Bauer ML, Gibbs PS, Lardy GP, Doetkott D, Kegode RB**. 2005. Comparison of two sampling methods for *Escherichia coli* O157:H7 detection in feedlot cattle. J. Food Prot. **68**:1724–1728.

34. **Niu YD, Xu Y, McAllister TA, Rozema EA, Stephens TP, Bach SJ, Johnson RP, Stanford K**. 2008. Comparison of fecal versus rectoanal mucosal swab sampling for detecting *Escherichia coli* O157:H7 in experimentally inoculated cattle used in assessing bacteriophage as a mitigation strategy. J. Food Prot. **71**:691–698.

35. **Black M, Goodman A, Dekhtyar A, Kitts C**. 2013. Pyroprinting: a novel strain differentiation method. unpublished.

36. **Montana A, Neal E, Dekhtyar A, Black M, Kitts C**. 2011. Chronology-sensitive hierarchical clustering of pyrosequenced DNA samples of *E. coli*: a case study. IEEE International Conference on BIBM. Atlanta, GA.

37. **Montana A**. 2013. Algorithms for library-based microbial source tracking. California Polytechnic State University, San Luis Obispo, CA.

38. **Colwell R**. 2009. EstimateS: Biodiversity Estimation. http://viceroy.eeb.uconn.edu/estimates/

39. **Berg RD**. 1996. The indigenous gastrointestinal microflora. Trends Microbiol. **4**:430–435.

40. **Wallick H, Stuart CA**. 1943. Antigenic relationships of *Escherichia coli* isolated from one individual. J. Bacteriol. **45**:121–126.

41. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M**. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. **60**:1136–1151.

42. **Johnson JR, Owens K, Gajewski A, Clabots C**. 2008. *Escherichia coli* colonization patterns among human household members and pets, with attention to acute urinary tract infection. J. Infect. Dis. **197**:218–224.