A PRELIMINARY REPORT ON *ESCHERICHIA COLI* STRAIN DIVERSITY IN COWS

A Senior Project

presented to

the Faculty of the Biological Sciences Department

California Polytechnic State University, San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science

by

Alison T. Stivers

March 2015

ABSTRACT

A Preliminary Report on Escherichia coli Strain Diversity in Cows

Alison T. Stivers

Pyroprinting is a strain typing method that relies on the simultaneous pyrosequencing of the multi-copy rRNA intergenic transcribed spacer regions of *E. coli* (Black, et al., 2014). These pyroprints can be used to identify the source of *E. coli* in the environment. Currently, Cal Poly's Center for Applications in Biotechnology (CAB) is augmenting the existing *E. coli* pyroprint library. By pyroprinting the intestinal *E. coli* of cows, we can quantify the strain diversity present, evaluate persistence, and determine the minimum sample size required for a complete overview of the cow intestinal *E. coli* population. These pyroprints can then be added to the current pyroprint library to be used for comparison to environmental samples. 504 unique pyroprints from two cows were generated over the course of this study, and from these, we determined that the sample size necessary to capture all of the diversity present in a cow at any given time would be over 300 isolates. The intestinal *E. coli* population of a cow is almost perfectly even, with most strains only being represented by one or two isolates. There is low persistence between sampling events, and very little transmission between cows. These data suggest that intestinal *E. coli* diversity in cows will be difficult to completely capture, and is not a prudent use of the CAB's time or money.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

Escherichia coli is a Gram-negative bacillus and a normal inhabitant of the mammalian gut (Duriez, et al., 2001). *E. coli* is naturally shed in the host feces (10^4-10^6 cfu/gram), giving us a snapshot of the *E. coli* density and diversity at that given time (Dowd, et al., 2008). Currently, there is a lack of data regarding the microbial diversity in the gut of livestock, mostly due to the expense of investigating the subject (Dowd, et al., 2008). Pyrosequencing has allowed us to investigate this topic, and furthermore, to add these data into an *E. coli* strain library for use in other projects. Characterizing diversity of *E. coli* in cows is important for establishing the sample sizes necessary to obtain from a single cow to evaluate the presence of a given strain.

Pyroprinting is a strain typing method that takes advantage of the intergenic transcribed spacer regions (ITS). These ITS have low conservation, and therefore, vary in sequence between strains of *E. coli*. The ITS regions used in this experiment are located in the ribosomal RNA (rRNA) operon between the 16S and 23S sequences and the 23S and 5S sequences.

Since the rRNA operon is present in the cell with seven separate copies, when the ITS regions are amplified by PCR, all seven copies serve as templates, producing amplified products that vary in their sequence. When the products are sequenced by pyrosequencing, all seven sequences are read by the pyrosequencer simultaneously, yielding a pyrogram that has a jumbled sequence, but a strain-specific pattern (called a pyroprint) (Black, et al., 2014). Pyroprints from separate isolates can be compared using a Pearson correlation, and isolates with a correlation value higher than 0.99 are said to be of the same strain (Black, et al., 2014).

To identify the amount of *E. coli* diversity in cattle, we pyroprinted approximately 350 *E. coli* isolates from two different cows at two different time points and compared all pyroprints to

each other using a Pearson correlation. This resulted in a list of clusters of related *E. coli* isolates, in which we can see how many isolates belong to each cluster (where each cluster is effectively one *E. coli* strain). A large volume of strains indicates a large amount of *E. coli* diversity, and vice versa for a small number of strains. We can also observe strain persistence over time by comparing strains in the same cow at the two different time points.

CHAPTER 2

Materials and Methods

Study Design and Sampling

Samples were collected from two cow subjects twice each in November 2013. Cow A was Cal Poly dairy cow #468 and Cow B was Cal Poly dairy cow #382. A sample of fresh feces was obtained immediately after defecation.

The feces samples were immediately diluted 10^{-3} with water and then plated on MacConkey plates using glass beads. The plates were then incubated at 37°C for approximately 24 hours. The plates were inspected for at least 100 red cfu and deficient plates were made with a new and more appropriate dilution of the original fecal sample. Red cfu indicate that the colony can ferment lactose and is an initial indicator of *E. coli*.

E. coli Confirmation

MacConkey agar was used as the initial confirmation step. To isolate *E. coli*, red colonies from the original MacConkey plate were streaked onto a second MacConkey plate. From this second MacConkey plate, red colonies were streaked onto Luria-Bertani (LB) agar. Isolates on LB were confirmed as *E. coli* with three subsequent steps: The first was plating on eosin-methylene blue agar, where *E. coli* results in a metallic green sheen; The second was an indole spot test using indole spot reagent, where *E. coli* results in a blue color; The last step was plating on Simmon's citrate agar, where *E. coli* shows no growth (Black, et al., 2014).

Pyroprinting

Two separate colony PCR reactions amplified each of two genomic intergenic transcribed spacer regions of the ribosomal RNA operon (23-5 ITS and 16-23 ITS) using forward and reverse primers for the respective ITS regions (Black, et al., 2014). Gel electrophoresis (2.5% agarose in

Tris-acetic acid-EDTA buffer) confirmed the PCR amplification was successful. A successful 16-23 PCR amplification would result in two bands at about 400 and 500 base pairs. A successful 23-5 PCR amplification would result in one band at about 200 base pairs. PCR products were pyrosequenced using a Qiagen Pyromark® Q24 and ITS-specific dispensation sequences with a length of 95 dispensations (Black, et al., 2014). Pyroprint data was exported from the Pyromark® software and collected for statistical analysis. Any pyroprints with wide peaks or double peaks at a dispensation were excluded and the isolate was re-sequenced.

Statistics

Pyroprints for the same ITS region were compared between different *E. coli* isolates using Pearson's correlation to assess matches. An algorithm described by Montana et al. (Montana, et al., 2013) clustered isolates into strains using correlation values between isolates from both ITS regions (Black, et al., 2014) (Montana, et al., 2011). Pairwise correlations of pyroprints were placed in a correlation matrix for strain identification (Black, et al., 2014). The algorithm clustered isolates if both ITS loci are over 0.99, and additional isolates were added to the cluster if both regions have a correlation value of over 0.99 against all isolates already in that cluster (Black, et al., 2014).

Simpson's index, the Shannon-Weaver index, species richness, and species evenness were calculated for Cow A and Cow B across both sampling events. Simpson's index was calculated using the sum of the square of the probability of randomly picking an individual of a particular strain (Σp_i^2). The Shannon-Weaver index was calculated using the inverse of the sum of the probabilities of randomly picking an individual of a particular strain times the natural log of that same probability ($1/\Sigma p_i \ln(p_i)$). Richness is the number of strains found, and evenness was the Shannon-Weaver index divided by the natural log of the richness value ($H/\ln(S)$). Rarefaction curves and statistics were performed in R.

CHAPTER 3

Results

Temporal Strain Diversity and Stability

Cow B had more representative isolates from each strain analyzed than did Cow A, even though there were more total strains found in Cow B (Figure 1).

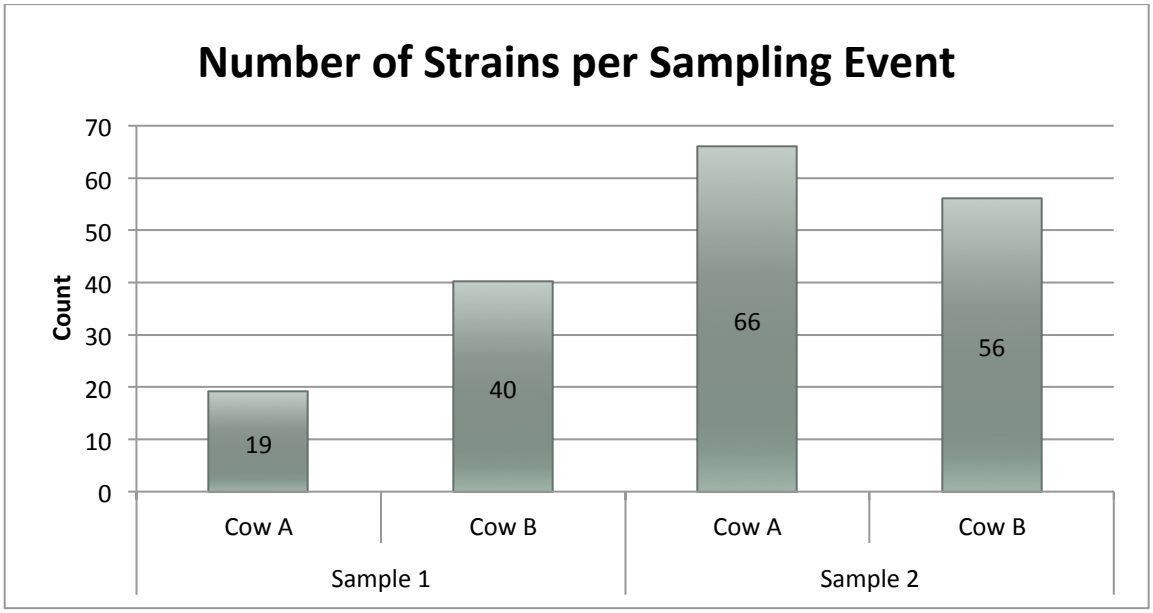

Figure 1. Number of strains per sampling event. The grey bar represents how many identifiable strains are present in each sampling event (also represented by the number within the grey bar). Sample 1 is the first sampling event on 11/4/2013, and Sample 2 is the second sampling event on 11/14/2015. Cow A Sample 1 n=21, Cow B Sample 1 n=75, Cow A Sample 2 n=83, Cow B Sample 2 n=75.

Cow A had the least amount of clustering between the two cows. For Sample 1, out of 21 isolates, only two clustered for a total of 19 strains. 17 of those strains consisted of one isolate, while the remaining two strains (11% of the total number of strains) each had two isolates. Sample 2 was slightly better in terms of clustering, with 83 isolates clustering into 66 strains. 54 of those strains only consisted of one isolate, while the other 12 strains (18%)

contained between two and four isolates. Cow B overall had more clustering and larger strains. Sample 1 had 75 isolates clustering into 40 strains. 26 of those strains only consisted of one isolate, while the remaining 14 strains (35%) had between two and seven isolates. Sample 2 had 75 isolates clustered into 56 strains, 45 of which consisted of one isolate, while the remaining 11 strains (20%) contained between two and four isolates. Looking at just the percentage of strains from each sampling event that contained more than one isolate, it is clear that Cow B (35% and 20%) had more strains that contained multiple representative isolates than did Cow A (11% and 18%).

The isolates were not only clustered within sampling events, but also across both sampling events, and then across both cows. The resulting strains can be viewed in Figure 2.

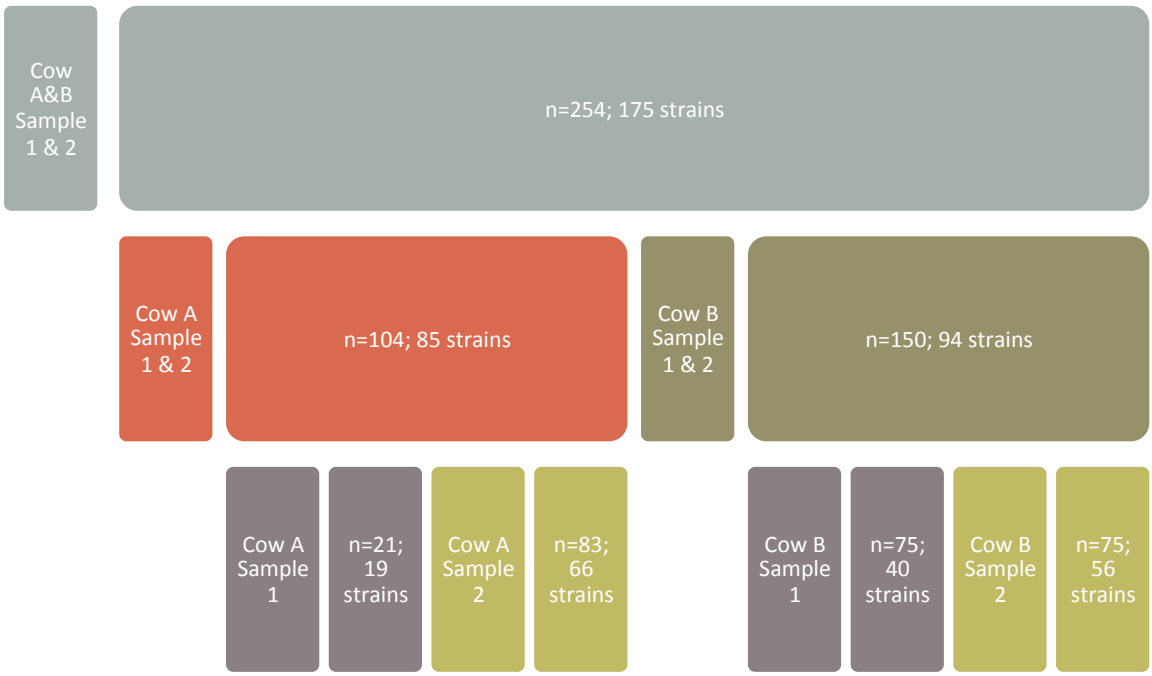

Figure 2. Hierarchy of isolates analyzed, strains containing only one isolate (singles), and strains containing more than one isolate (clusters).

The bottom row of Figure 2 is a summary of what was shown in in Figure 1. Cow A Sample 1 consisted of 21 isolates divided into 17 strains containing only one isolate (referred to as a single) and two strains containing more than one isolate (referred to as a cluster). This continues across the row, with Cow A sampling events on the left, and Cow B sampling events on the right. These sampling events are then combined in row two of this figure. Cow A Sample 1 & 2 show the results of combining both sampling events. 104 isolates were clustered into 71 strains containing only one isolate, and 14 strains (16% of the total number of strains) containing between two and four isolates. Cow B Sample 1 & 2 had 150 isolates clustered into 69 strains containing only one isolate, and 25 strains (27%) containing between two and seven isolates. The top row combines all sampling events for both cows, and the 254 isolates were clustered into 135 strains containing only a single isolate, and 40 strains (23%) containing between two and eight isolates. This suggests that not only is there low representation from each strain within a single cow and a single sampling event, but also that there is little strain overlap between the cows.

E. coli Abundance and Strain Distribution among Subjects

Over the course of this study, 254 *E. coli* isolates were collected, pyroprinted, and clustered into 175 different strains. Since only 104 *E. coli* were collected from Cow A between the two sampling events compared to the 150 from Cow B, direct comparisons concerning abundance are difficult to make because of differences in sample size. However, calculations were made for indirect comparisons.

Table 1. Table of Simpson's Index (Σp_i^2), Shannon-Weaver Index ($1/\Sigma p_i \ln(p_i)$), richness (number of strains), and evenness ($E=H/\ln(S)$) values for all isolates collected from Cow A (n=104) and all isolates collected from Cow B (n=150).

	Simpson's D	Shannon-Weaver H	Richness S	Evenness E
Cow A Sample 1 & 2	0.0142	4.27	85	0.96
Cow B Sample 1 & 2	0.0174	8.64	94	0.95

The Simpson's index (D) (Table 1) is the probability that any two randomly selected individuals belong to the same strain, which suggests how diverse a population is. Cow B has a slightly larger value (0.0174) than Cow A (0.0132), indicating that there is a higher abundance of strains in Cow A (corrected for sample size), which is corroborated by Figure 1. The Shannon-Weaver Index (H) measures the entropy within the system. Again, Cow B has a higher value (8.64) than Cow A (4.27). A large Shannon-Weaver value means that for each new isolate, there is a high uncertainty as to what strain it belongs to. A Shannon-Weaver value close to zero means that there is low evenness, and all individuals belong to the same strain. Since neither Cow A nor Cow B's Shannon-Weaver value approaches zero, both cows have high entropy within the system. Cow B's value is larger, indicating more uncertainty in picking an individual from a particular strain than Cow A. Richness (S) is a measure of the number of different strains measured in the sample. Evenness (E) is a measure of similarity of the abundance of different strains. High similarity would be if each strain had the same number of individuals. Low similarity would be if one strain contained many individuals, and the rest of the strains contained a variable number of individuals. Cow A has a slightly higher value (0.96) than Cow B (0.95), indicating that there is higher evenness in that sample (where E=1 would be complete evenness).

There is high evenness and high diversity within both cows, but that diversity can also be explored between the two cows. High strain overlap between the cows would indicate that there could be transmission of *E. coli* between animals.

Table 2. *E. coli* strains shared between Cow A and Cow B, and number of contributing isolates from each subject.

	Number of Isolates by Subject		
Shared Strains	Cow A	Cow B	Total Isolates
Strain 26	1	1	2
Strain 32	1	1	2
Strain 98	1	1	2
Strain 100	1	1	2
Strain 103	2	1	3
Strain 105	4	4	8
Strain 112	1	3	4
Strain 114	1	1	2
Strain 127	1	1	2
Strain 145	1	1	2
Strain 159	1	2	3
		Total	32

Out of the total 175 strains detected between Cow A and Cow B (Table 2), only 11 strains (with a total of 33 isolates) out of a total of 175 strains (or approximately 6%) were shared between the two cows. This result implies that most *E. coli* strains are not shared between cows, or that there is such high strain diversity that many more isolates would need to be collected to accurately analyze the number of strains shared between cows. There were zero strains shared between the two sampling events for Cow A, and only two strains shared between sampling events for Cow B. This indicates low strain stability over time.

To investigate if more sampling would be needed to fully explore the diversity within a cow, a rarefaction curve was created. A line reaching an asymptote would indicate that sufficient sampling has been completed and there are few strains still to be discovered.

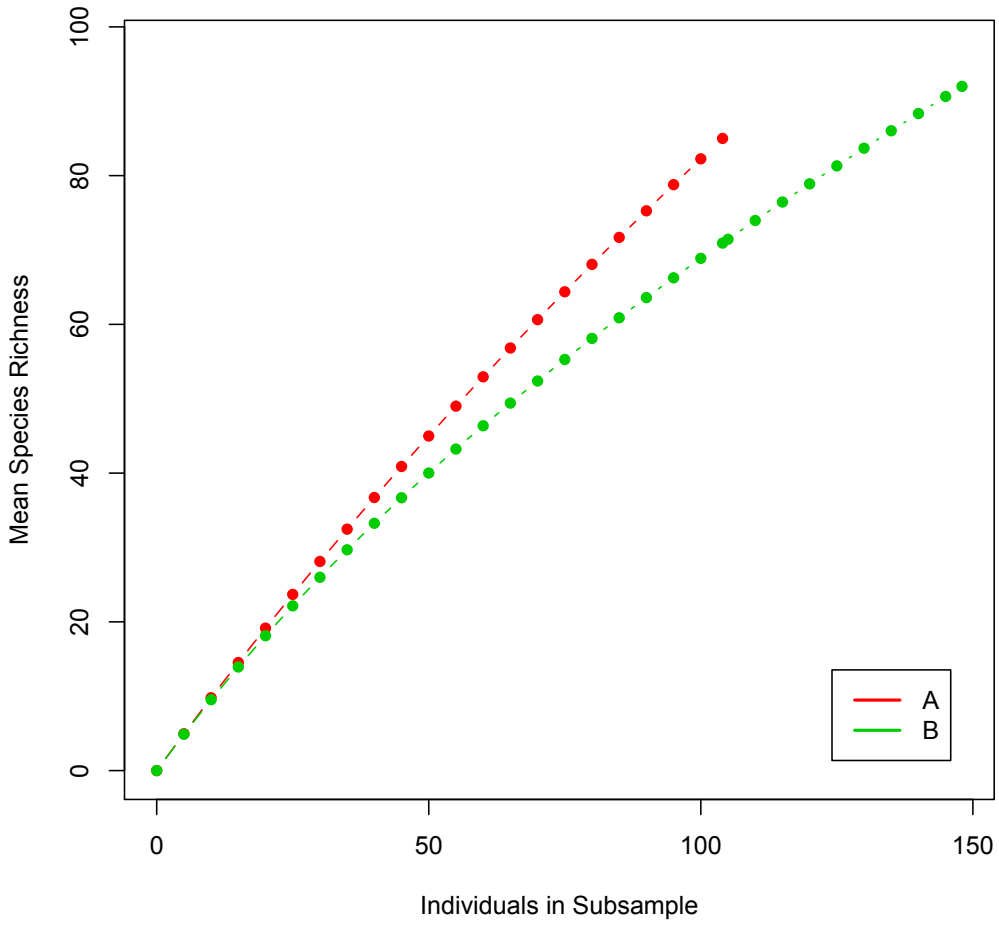

Figure 3. Rarefaction curve for Cow A and Cow B using subsample=5.

The rarefaction curves for Cow A and Cow B (Figure 3) indicate that there is a large amount of strain richness, as both curves are relatively steep and neither reaches an asymptote. The lack of an asymptote means that more sampling would need to be completed for an accurate representation of the diversity in the cows. However, Cow A appears to have higher

strain richness than does Cow B because the line has a steeper slope. The slope is indicative of how often an individual of a new species would be randomly picked from the sample. A steep slope indicates an individual from a new species is picked every sampling event, while a gentler slope would indicate that there are fewer new species being discovered with each sampling event.

Since the rarefaction curve did not reach an asymptote, strain richness can be estimated using statistical tests.

Table 3. Strain richness estimator ACE (Abundance-based Coverage Estimator); calculated using R.

	Cow A	Cow B
Strains observed	85	92
ACE Estimator	323.90	274.58
Standard Error	4.87	9.08

The ACE estimator approximates the minimum number of strains that would be found in the sample if the entire sample were censused. ACE estimated strain richness in Cow A to be approximately 324±4.87, and in Cow B to be approximately 275±9.08. Out of the 104 isolates that were collected from Cow A, we found 85 strains. To sample from Cow A and have at least one isolate from each strain, we would need to collect approximately 396 isolates. Out of the 150 isolates collected from Cow B, we found 94 strains. To sample from Cow B and find at least one isolate from each strain, we would need to collect 439 isolates.

CHAPTER 4

Discussion

Strain Diversity and Stability

Between 104 and 150 *E. coli* were sampled from both cows over two sampling periods with the hope that this sampling effort would accurately represent the amount of *E. coli* diversity present in the cows at the time of sampling. Unfortunately, the results collectively suggest that a larger number of isolates would have to be collected for the majority of strains present in the cows to be represented. The low Simpson's indices in Table 1 indicate that there is a very low probability that two isolates chosen at random from the sample will be the same (both Cow A and Cow B have Simpson's indices of between one and two percent). These numbers hint at the high diversity present in the two cows. The Shannon-Weaver indices are both high, showing the high amount of uncertainty in the prediction of the strain type of any particular isolate, corroborating the results of the Simpson's indices. Table 1 also discussed strain richness and evenness, which can be further visualized in Figure 1. The high amount of strain richness (high when compared to sample size) and the corresponding high evenness confirm that there are many strains present and most are represented by a single isolate.

This result is further explored with the rarefaction curve in Figure 3. The lack of any asymptotes verify that many more than 100-150 isolates per cow will be necessary to completely capture the *E. coli* diversity present in the animal. The richness estimator ACE suggests that 396 isolates would need to be collected from Cow A and 439 from Cow B to completely represent the strain diversity at any one time. Ultimately, these data collectively suggest that pyrosequencing the intestinal *E. coli* of cows to observe strain sharing between animals is not feasible under normal time constraints.

Pyrosequencing Challenges

Over the course of this study, it was determined that the pyrosequencing protocols needed to be adjusted to increase the yield of quality pyroprints. Many pyroprints contained wide peaks and double peaks, and the isolates had to be re-sequenced several times. After a pilot study of the reliability of the results between the two different protocols, it was determined that the Pearson correlation between isolates pyrosequenced with different protocols was too low. Therefore, all of the isolates pyrosequenced with the "old" protocol needed to be redone. This amounted to 238 regions (16-23 or 23-5) of the 508 total regions (two regions for each of the 254 samples analyzed). Unfortunately, the data had already been analyzed and clustered, and so this paper had to be written with a combination of the "old" and "new" data.

During the pilot study, the old protocol and new protocol matched each other 50% of the time, but only 17% of the time did the two protocols (which were performed simultaneously) match the sample already present in the database. The sample present in the database had previously been pyrosequenced with the old protocol, and so the lack of a match between an isolate pyrosequenced with the same protocol at different times was concerning. This indicates that there are issues with correlation between protocols, and these issues may need to be corrected before completely accurate data may be obtained.

Since the original problem was the presence of wide peaks and double peaks in the pyroprints, there may be a relationship between peak width and pyroprint accuracy. This relationship should be further explored to see if we could exclude pyroprints based on a peak width threshold. Despite these obstacles, I am still confident that the general result of high

diversity and evenness in cow *E. coli* is correct. The isolates pyrosequenced with the new protocol do seem to cluster with other isolates more frequently (55% of isolates pyrosequenced with the new protocol cluster, compared to 41% pyrosequenced with the old protocol). Using a paired t-test, this result is statistically significant (*paired t-test:* p=0.0405, df=1, t=15.71), meaning that it is most likely that more isolates should cluster with one another if all of the isolates were pyrosequenced with the new protocol. However, the difference in clustering is only approximately 14%, and even a 14% reduction in diversity leaves us with approximately 279 strains in Cow A and 236 strains in Cow B. Assuming all other data is correct, over 300 isolates would still need to be collected from each cow to observe total diversity in the animals. This is still not a feasible number.

WORKS CITED

Black, M., Goodman, A., Dekhtyar, A., & Kitts, C. (2014). Pyroprinting: a novel strain differentiation method. *Journal of Microbiological Methods, 105*, 121-129.

Black, M., Vanderkelen, J., Montana, A., Dekhtyar, A., Neal, E., Goodman, A., et al. (2014). Pyroprinting: A rapid and flexible genotypic fingerprinting method for typing bacterial strains. *Journal of Microbiological Methods*, 121-129.

Dowd, S., Callaway, T., Wolcott, R., Sun, Y., McKeehan, T., Hagevoort, R., et al. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rRNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology, 8* (125).

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventre, A., Elion, J., et al. (2001). Commensal Escherichia coli isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology, 147*, 1671-6.

Gomi, R., Matsuda, T., Matsui, Y., & Yoneda, M. (2014). Fecal Source Tracking in Water by Next-Generation Sequencing Technologies Using Host-Specific Escherichia coli Genetic Markers. *Environmental Science & Technology, 48*, 9616-23.

Montana, A., Dekhtyar, A., Black, M., Kitts, C., & Goodman, A. (2013). *Ontological hierarchical clustering for library-based microbial source tracking.*

Montana, A., Dekhtyar, A., Neal, E., Black, M., & Kitts, C. (2011). Chronology-sensitive hierarchical clustering of pyrosequenced DNA samples of E. coli- a case study. *2011 IEEE International Conference on Bioinformatics and Biomedicine*, (pp. 155-159).

Neal, E. (2013). *Escherichia coli strain diversity in humans: effects of sampling effort and methodology.* Master Thesis, Cal Poly, San Luis Obispo.