INDEX

1. INTRODUCTION AND OBJECTIVES

Next generation sequencing (NGS) holds significant promise for improving public health (Struelens & Brisse, 2013). Nevertheless, transforming powerful NGS techniques into practical public health solutions will be a major challenge in the coming years (Carriço, Sabat, Friedrich, Ramirez, & Markers, 2013). The purpose of this paper is to identify how NGS can improve estimates of the burden of foodborne diseases and contribute to improving food safety in general. Examples of poultry-related studies are used to demonstrate NGS data analysis techniques, with particular emphasis on the potential variability associated with the analysis of NGS data.

The objective of this capstone paper is to demonstrate multi-disciplinary learning relevant to Environmental Health Science, Epidemiology, and Bioinformatics. During May to August 2013, Mr. Caudill collected and prepared poultry-related samples for NGS (Site Description below). Mr. Caudill also gains experience in setting up a Linux-based computational resource to process NGS data with QIIME 1.8.0, which is a software package that integrates various data analysis algorithms (Caporaso, Kuczynski, et al., 2010). This capstone paper incorporates some of the methods used during Mr. Caudill's internships and integrates learning relevant to both Environmental Health Science and Epidemiology.

2. SITE DESCRIPTION AND MISSION

United States Department of Agriculture (USDA) is an executive department of the federal government. The overall mission of the USDA is to “provide leadership on food, agriculture, natural resources, rural development, nutrition, and related issues based on sound public policy, the best available science, and efficient management” (USDA, 2013, Fiscal Year 2013 Agency Financial Report).

The USDA contains a number of agencies that are grouped within seven mission areas. One of the mission areas is the Research, Education, and Economics mission area, which contains four agencies, including the Agricultural Research Service (ARS) (USDA, 2013, Fiscal Year 2013 Agency Financial Report). The mission of ARS includes the objective to “ensure high-quality, safe food, and other agricultural products” (ARS, 2014, www.ars.usda.gov).

The ARS has an annual budget of approximately $1.1 billion, which it uses to conduct research at over 90 locations (ARS, 2014). The Richard B. Russell Research Center (RRC) is one of these ARS research locations and is located on College Station Road in Athens, GA. The principle investigator involved in the four poultry-related studies included in this paper is Dr. Michael J. Rothrock Jr., Ph.D., who is a Research Microbiologist at the RRC.

3. ANALYSIS OF THE PROBLEM

This section discusses the challenges in obtaining accurate and precise estimates of the burden of foodborne diseases, the consequences of inaccurate estimates, and the potential for next generation sequencing to contribute to the solution.

3.1 The problem of estimating the burden of foodborne diseases

3.1.1 Original foodborne disease estimations

Foodborne pathogens cause a substantial burden of disease in the United States. Quantifying the burden of disease caused by foodborne pathogens, however, has proven to be challenging. From the mid-1980s through most of the 1990s, estimates of foodborne illnesses varied significantly (Mead et al., 1999). For example, one study estimated that there were 12.6 million cases of foodborne diseases in the United States each year (Todd, 1989), while another study reported a range from 23 million to 81 million or more cases (Archer & Kvenberg, 1985). In the mid-1990s, the Centers for Disease Control and Prevention (CDC), the United States Department of Agriculture (USDA), the Food and Drug Administration (FDA), and selected state health departments collaborated to create the Foodborne Diseases and Active Surveillance Network or more simply, FoodNet (Centers for Disease Control and Prevention, 1997, MMWR, 46(12), 258). One of the primary goals of FoodNet was to assist in developing more precise estimates of foodborne diseases in the United States (Centers for Disease Control and Prevention, 1997, MMWR, 46(12), 258). In 1999, Mead and authors published the most definitive (over 6,300 citations from 1999 to 2014) (Google Scholar, 2014) estimates of the burden of foodborne diseases to date, using data from FoodNet, four other surveillance systems, three national surveys, the National Vital Statistics System, and selected published studies (Mead et al., 1999). Mead and authors' estimates of the annual burden of foodborne pathogens in the United States are shown in Table 1. Mead et al. estimated that foodborne pathogens were responsible for 76 million annual illnesses, including 14 million illnesses from known pathogens and 62 million illnesses from unknown agents (Mead et al., 1999). Mead et al. further estimated that foodborne pathogens caused 325,000 hospitalizations and 5,000 deaths each year (Mead et al., 1999).

Table 1. **Estimates of the Annual Burden of Foodborne Pathogens in the United States**

(a) from Mead et al., 1999

	Illnesses	Hospitalizations	Deaths
Known pathogens	14 million (18%)	60,000 (18%)	1,800 (36%)
Unknown agents	62 million (82%)	265,000 (82%)	3,200 (64%)
Total	76 million (100%)	325,000 (100%)	5,000 (100%)

(b) from Scallan, Hoekstra, et al., 2011 and from Scallan, Griffin, et al., 2011

	Illnesses	Hospitalizations	Deaths
Known pathogens	9.4 million (20%)	55,961 (44%)	1,351 (44%)
90% CrI[a]	6.6-12.7 million	39,534-75,741	712-2,268
Unknown agents	38.4 million (80%)	71,878 (56%)	1,686 (56%)
90% CrI[b]	19.8-61.2 million	9,924-157,340	369-3,338
Total[c]	47.8 million (100%)	127,839 (100%)	3,037 (100%)
90% CrI[d]	26.4-73.9 million	49,458-233,081	1,081-5,606

[a] 90% credible interval as reported in Scallan, Hoekstra, et al., 2011

[b] 90% credible interval as reported in Scallan, Griffin, et al., 2011

[c] Summation of estimates of known pathogens and unknown agents; not reported summed by Scallan et al.

[d] Approximation of 90% credible interval by summation of the component intervals; not reported by Scallan et al.

3.1.2 Uncertainty in foodborne disease estimations

The results from Mead et al. have faced scrutiny in more recent years. The issue is that estimating illnesses, hospitalizations, and deaths resulting from foodborne diseases requires a sizeable number of parameters, and some of these parameters have a substantial amount of uncertainty (Morris Jr, 2011). For example, Powell and authors calculated that for a median value for *E. coli* O157:H7 of 75,000 cases per year, the 95% credible interval ranged from 50,000 to 120,000 cases per year, indicating substantial uncertainty in the estimate (Powell, Ebel, & Schlosser, 2001). In another example, Phillips and LaPole demonstrated the effect of introducing uncertainty into the percentage of gastroenteritis cases caused by Noroviruses that was attributable to food (Phillips & LaPole, 2003). Mead and authors estimated that 40% of total illnesses caused by Noroviruses were attributable to food (Mead et al., 1999). Phillips and LaPole used Monte Carlo simulations with percentages ranging from 20% to 60% (as the actual percentage was not well-established) and found that only about 50% of the probability means fell within the already wide range of 50-100 million foodborne illnesses per year (Phillips & LaPole, 2003). Thus, a reasonable assumption of variation for a single parameter of one pathogen was shown to result in considerable uncertainty in the reported 76 million total foodborne pathogen cases per year. In another study, Frenzen explained the challenges in estimating deaths from unknown pathogens (Frenzen, 2004). Frenzen highlighted the high degree of uncertainty involved in these estimations and consequently, casted significant doubt concerning the accuracy of the unknown pathogen death estimates presented by Mead et al. (Frenzen, 2004).

3.1.3 Updated foodborne disease estimations

In light of the limitations of the Mead et al. study, requests were initiated to the CDC for the estimates to be re-calculated with better methodology (Morris Jr, 2011). Scallan and authors answered these requests with two papers published in 2011 (Scallan, Griffin, Angulo, Tauxe, & Hoekstra, 2011; Scallan, Hoekstra, et al., 2011). Instead of combining all of the estimates for foodborne pathogens into a single paper, Scallan and authors published one paper with estimates for 31 major known pathogens (Scallan, Hoekstra, et al., 2011) and another paper with estimates for unspecified agents (Scallan, Griffin, et al., 2011). While these estimates were published separately, they are also shown totaled in Table 1. Comparing section (a) and section (b) of Table 1 could lead to the conclusion that food safety dramatically increased during the approximate decade between the papers. In fact, directly comparing the two estimates would suggest that foodborne illnesses decreased by about 37% over about a decade. While this conclusion would be appealing to many food safety stakeholders, this conclusion is not appropriate (Morris Jr, 2011). To elaborate, the methods employed by Scallan et al., while improved over the methods used by Mead et al., were different enough to make any direct comparison of the estimates between papers inappropriate (Morris Jr, 2011). Notably, Scallan and authors improved the uncertainty calculations and included 90% credible intervals in their publications (Morris Jr, 2011). While a direct comparison of the estimates between papers is not appropriate, it can be noted from Table 1 that the percentage of the contribution from known pathogens increased in the Scallan et al. estimates compared to the Mead et al. estimates. The most dramatic difference is seen in the estimates of foodborne-pathogen-related hospitalizations, in which Mead et al. estimated that known pathogens contributed to only 18% of the hospitalizations, while Scallan et al. estimated that known pathogens contributed to 44% of the hospitalizations.

3.2 The problem of inaccurate foodborne disease estimates

Clearly, the goal for the nation is to continue to reduce foodborne illnesses and sequelae. Nevertheless, there are extensive costs associated with a high level of food safety (Antle, 1999). For example, the Government Accountability Office estimated that in 1999 the federal government spent $1 billion on food safety and the state governments spent an additional $300 million (General Accounting Office, 2001, GAO-01-177). The high costs of food safety is one reason that getting good estimates of the burden of foodborne pathogens is critical. Policymakers need to be able to make decisions informed by both the cost of food safety and the cost of foodborne diseases. Policymakers will have more accurate information upon which to base their decisions if there are better estimates of the burden of foodborne diseases and subsequently, better estimates of the costs associated with foodborne diseases. Two papers published by Scharff illustrate how changing the estimates of the burden foodborne diseases affects the estimates of the cost of foodborne diseases (Scharff, 2010, 2012). Using the 1999 Mead et al. estimates, Scharff estimated the annual cost of foodborne illnesses to be $152 billion (95% CI: $39-$265 billion) (Scharff, 2010). Using similar methods based on the 2011 Scallan et al. estimates, Scharff estimated the annual cost of foodborne illnesses to be $77.7 billion (90% CI: $28.6-$144.6 billion) (Scharff, 2012). Thus, the updated cost of illness estimate based on Scallan et al. was substantially lower, about half of the original estimate. Scientists must strive to produce the most accurate estimates of the burden of foodborne diseases so that policymakers can be informed by the most accurate estimates of the magnitude and cost of these diseases.

3.3 NGS as part of the solution

3.3.1 Addressing the problem of inaccurate estimates

While policymakers deserve the best information, producing accurate estimates of the burden of foodborne diseases is not easy. For example, Table 1 shows that both studies estimated that about 80% of the total foodborne illnesses involved unknown pathogens. Furthermore, the credible intervals reported by Scallan et al. were much wider for unknown pathogens than for known pathogens (Scallan, Griffin, et al., 2011; Scallan, Hoekstra, et al., 2011). While this is expected, it is important to note that the vast majority (~80%) of the estimated total burden of foodborne diseases is comprised of data with a high degree of uncertainty. Therefore, even the improved methods employed by Scallan et al. produced results that are imprecise.

One strategy to improve accuracy of the estimates would be to increase the percentage of illnesses attributable to known pathogens. This, of course, is much easier said than done. A basic problem with this strategy is that pathogens can be tricky to identify. In 1985, Stanley and Konopka noted that “only a few percent of the bacterial cells enumerated by direct microscopic count can be cultured and identified” (Staley & Konopka, 1985). A more recent study reported that scientists have probably not yet identified even half of the pathogens that have clinical significance (Oakley et al., 2013). In the past few decades, technology has been the major limiting factor in investigating these unknown bacterial communities. In 1985, Stanley and Konopka had to rely on community metabolism and other indirect measures to estimate species diversity (Staley & Konopka, 1985). But scientists now have new tools available to identify pathogens: next generation sequencing (NGS) and subsequent metagenomic analysis (Oakley et al., 2013). Genome sequence information has the potential to shed light on previously underexplored bacterial communities and identify sets of organisms that have previously eluded cultivation techniques (Rappé & Giovannoni, 2003). Implementing NGS technology in the realm of food safety may eventually allow more “unknown agents” to become “known pathogens,” which should increase the accuracy of the burden of foodborne diseases and subsequently, provide policymakers with more accurate information concerning the magnitude and cost of foodborne diseases.

3.3.2 Addressing the problem of stagnant food safety initiatives

In addition to providing policymakers with more accurate estimates of the burden and cost of foodborne diseases, policymakers need information about another topic: the effectiveness of government-funded food safety initiatives. As previously discussed, it is not appropriate to simply compare estimates of foodborne diseases in one study to another study; comparing estimates from two different studies would only be appropriate if the estimation methods were the same or very similar. One solution for addressing the effectiveness of food safety programs would be to examine FoodNet data over time (Morris Jr, 2011). One of the original goals of FoodNet was actually for the purpose of evaluating the Hazard Analysis Critical Control Point (HACCP) rule implemented by the USDA (Voetsch et al., 2004). FoodNet and HACCP were both implemented in the mid-1990s, and over 75% of meat and poultry production plants had implemented a HACCP plan by January 1998 (Voetsch et al., 2004). By examining the FoodNet data, Morris concludes that the USDA’s regulatory changes, including HACCP, were associated with a decrease in foodborne diseases (Morris Jr, 2011). Unfortunately, Morris also concludes that “after the initial decline since the USDA regulatory changes in 1995, one does not see evidence of sustained improvement” (Morris Jr, 2011). Thus, while foodborne illnesses may not be getting more prevalent in the United States, presumably the current estimate shown in Table 1 of almost 50 million foodborne illnesses annually is not acceptable as the level of illnesses the United States desires to maintain. In regards to the more or less stagnant levels of foodborne illnesses in the United States, NGS may also be part of the solution (See Section 3.4). Once again, if more “unknown agents” become “known pathogens,” it is likely that food safety techniques can become more targeted to novel pathogens, and presumably, food safety programs will be more successful in reducing the burden of foodborne diseases.

3.4 NGS one step at a time: using poultry as an example

Applying NGS techniques to improve food safety will be a long process. Just as it has taken many years to develop successful and useful methods of bacterial cultivation for some of the known pathogens, it will likely take many years to learn how to make NGS part of the food safety solution. Many food safety risks, and therefore many potential interventions, exist along the “farm-to-fork” continuum (Batz et al., 2005). Thus, determining where to most effectively intervene will be challenging. Food safety interventions may be most successfully designed if there is an ability to associate particular food sources with specific illnesses (Morris Jr, 2011; Pires et al., 2009). With that in mind, NGS applications discussed later in this paper will focus on only one food source, poultry. The application of NGS techniques to poultry production will serve as an example of how learning more about the microbiomes associated with a particular food source can be the foundation for future food safety interventions. Furthermore, it should be noted that examining poultry is relevant as it was found to be the third leading vehicle of foodborne illness outbreaks from 1990-2003 (Dewaal, Hicks, Barlow, Alderton, & Vegosen, 2006).

4. ANALYSIS OF NGS SOLUTIONS

This section gives background information on NGS generally and more specially in relation to bacterial communities. This section also examines multiple sources of variation, with particular emphasis on data analysis. Variation from data processing is demonstrated with examples based on four poultry-related studies.

4.1 Introduction to Next Generation Sequencing (NGS)

In light of the current limitations on bacterial cultivation (Rappé & Giovannoni, 2003) and the significant burden of unknown agents (Scallan, Griffin, et al., 2011), the previous section proposed that NGS could assist in producing better estimates of the burden of foodborne diseases and potentially lead to reductions in the burden of these diseases. Nevertheless, while NGS is becoming increasingly more affordable (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013), NGS is not a straightforward solution. Substantial differences exist among preparation steps, sequencing methods, and data analyses (Metzker, 2010). Just as the different methods employed by Mead et al. and Scallan et al. made direct comparison of their results inappropriate, the different combinations of methods used in NGS and subsequent data analyses can make comparisons among these studies challenging (Metzker, 2010). That notwithstanding, NGS technologies offer unprecedented speed, affordability, and potential (Shendure & Ji, 2008). “It is an exciting time to be a molecular ecologist” (Glenn, 2011).

One of the biggest differences in NGS is the sequencing platforms themselves. Glenn compared the following manufacturers and platforms: Life Technologies / Applied Biosystems (3730, capillary; SOLiD); Roche / 454 (FLX; GS); Illumina (GA IIx; HiSeq; MiSeq; NextSeq);

Ion Torrent (Proton; PGM); and Pacific Biosciences (RS II) (Glenn, 2011). These platforms have substantial differences in terms of their sequencing methods, amplification methods, read lengths, cost per run and per read, error types and rates, etc. (Glenn, 2011). The choice of which platform or combination of platforms to use is highly dependent on the type and amount of data to be processed (Glenn, 2011). Nevertheless, for the past few years, Illumina platforms have held the highest percentage of the NGS market (Karrow & Toner, 2011). For most molecular ecologists, the Illumina MiSeq has been the best choice, but Illumina's new NextSeq 500 platform may gain popularity in the coming years (Glenn, 2011).

4.2 Choosing appropriate samples

In addition to choosing which platform will be most appropriate for the data, researchers must choose what and how much they want to sample. Using poultry as the example, researchers must choose a more specific final product: meat or eggs. In regards to either of these products, the ultimate food safety concern is the final product available to the consumer. Nevertheless, if the final product is unsatisfactory in terms of food safety, the contamination had to have come from somewhere. Thus, in addition to collecting samples from the final product itself, it makes sense to sample the flock to attempt to determine if there are potential ways to intervene at the farm or during processing. When choosing to sample the flock, however, there are many important considerations. Factors such as age, environment, and diet can all impact the bacterial communities (Gabriel, Lessire, Mallet, & Guillot, 2006). Other considerations would include whether to sample feces and/or organs. When sampling the gastrointestinal tract, it is important to take into account that different areas of the gastrointestinal tract harbor varying bacterial communities (Sekelja et al., 2012). Thus, for even just one food source, it is evident that developing the best sampling techniques for the application of NGS will require substantial methodological research.

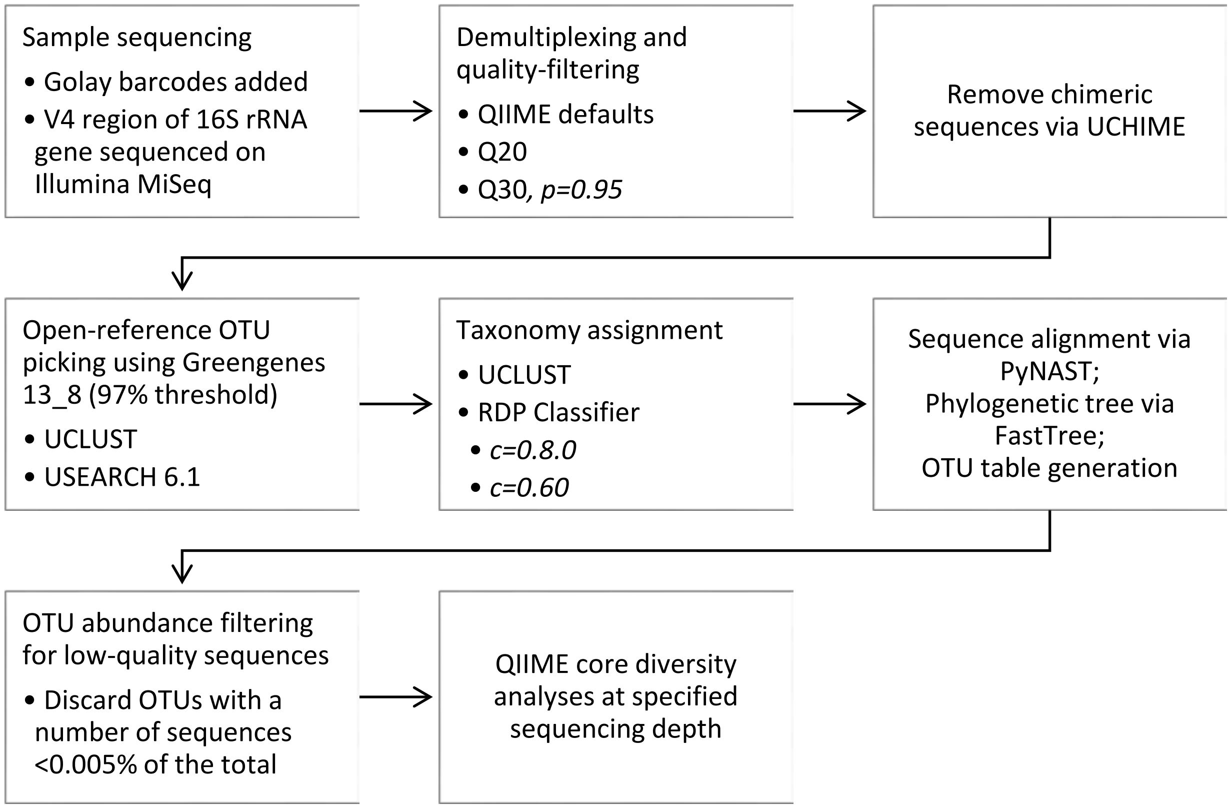
4.3 Effect of extraction methods

Once samples that are appropriate for the particular research question have been collected, researchers must choose how to extract the genetic material that will later be prepared for sequencing. With complex environmental samples, such as feces and soils, that contain substantial quantities of organic and inorganic matter, the extraction method used to isolate bacterial genetic material from the other sample components could potentially have a significant impact on the resulting quantity and quality of the bacterial DNA. There are currently two main methods for extracting DNA from these complex environmental samples: (1) mechanical disruption using bead-beating and (2) enzymatic disruption to release bacterial cells and inhibit PCR inhibitors chemically (Rothrock, Hiett, Caudill, Cicconi-Hogan, & Caporaso, 2014). Rothrock and authors found that the mechanical disruption resulted in higher quantities of 16S rDNA, the enzymatic disruption appeared to be associated with greater diversity, and a novel hybrid method produced results sharing aspects of both approaches (Rothrock et al., 2014). Thus, in addition to having variability from different sequencing platforms and substantial variability from different sample types (even when related to the same flock or bird), it is important for researchers to note that the extraction method can also significantly impact their findings.

4.4 NGS data analysis

In addition to the variability from the sequencing platform, sampling methodology, and extraction method, the programs and parameters used in analyzing the sequencing data can produce considerable variability. This subsection will examine the effects of differing data analysis programs and parameters by using data from four studies conducted by Dr. Michael Rothrock, Jr. at the Russell Research Center of the USDA Agricultural Research Service in Athens, GA. Figure 1 illustrates a general flowchart for the data analysis used in these four studies. Before consider the details of the data analysis methods, this section will provide a brief introduction to microbial sequencing theory and the methods used in these four studies.

Figure 1. **Flowchart of NGS Data Analysis Used in Four Poultry-Related Studies**


Sample sequencing
• Golay barcodes added
• V4 region of 16S rRNA gene sequenced on Illumina MiSeq
Demultiplexing and quality-filtering
• QIIME defaults
• Q20
• Q30, *p=0.95*
Remove chimeric sequences via UCHIME
Open-reference OTU picking using Greengenes 13_8 (97% threshold)
• UCLUST
• USEARCH 6.1
Taxonomy assignment
• UCLUST
• RDP Classifier
• *c=0.8.0*
• *c=0.60*
Sequence alignment via PyNAST;
Phylogenetic tree via FastTree;
OTU table generation
OTU abundance filtering for low-quality sequences
• Discard OTUs with a number of sequences <0.005% of the total
QIIME core diversity analyses at specified sequencing depth

4.4.1 Background on microbial DNA sequencing

To assign taxonomy and to determine phylogenetic relationships among bacteria, the most commonly used stable code segment is the 16S ribosomal RNA (rRNA) gene, also called 16S rDNA (Clarridge, 2004). The 16S rRNA is the small subunit of the bacterial ribosome, and the gene that codes for this small subunit has been highly conserved among different bacterial species (Clarridge, 2004). Highly conserved regions of the 16S rRNA gene make good candidates for designing universal primers that can target desired regions of the gene (Chakravorty, Helb, Burday, Connell, & Alland, 2007). Of course, if all of the 16S rDNA were highly conserved, the gene would not be useful for distinguishing one group of bacteria from another. Fortunately, despite the overall high level of conservation, the 16S rRNA gene contains nine hypervariable regions, labeled V1-V9, that can serve as molecular fingerprints for differing bacteria (Chakravorty et al., 2007).

4.4.2 Sequencing on the Illumina MiSeq platform

With about 15 million reads per run for the Illumina MiSeq v2 instrument (Glenn, 2011), using the capacity of the machine cost-effectively for microbial data requires a method for combining multiple samples or even multiple studies per run. More specifically, individual samples can be labeled with sequences such as the Golay barcodes described by Caporaso and authors (Caporaso et al., 2012). A mapping file can then be used after sequencing to demultiplex the data such that the base pair barcodes are replaced by meaningful sample identifiers (Navas-Molina et al., 2013). The studies conducted by Dr. Rothrock were processed in this manner. After extracting and preparing the DNA, the samples were sent to the Argonne National

Laboratory, where the V4 region of the 16S rRNA gene was PCR amplified with primers containing MiSeq sequencing adapters and Golay barcodes. Libraries were constructed and run on the MiSeq platform. The resulting forward and reverse paired-end reads and an index read for the barcodes were made available for download.

4.4.3 Data analysis using QIIME software

The number of programs available to analyze NGS data is almost overwhelming. Fortunately, bioinformaticians have reduced the complexity and improved the data analysis workflow by “wrapping” many algorithms into unified software packages. One particularly popular wrapper for microbial metagenomics is the wrapper called ‘quantitative insights into microbial ecology,’ or more simply, QIIME (Caporaso, Kuczynski, et al., 2010). According to the QIIME web information (http://qiime.org/), QIIME version 1.8.0 released in December 2013 is the most stable version as of April 28, 2014. QIIME can be run on a variety of computer platforms, including Linux (ex. Ubuntu), the Amazon EC2 cloud, or via VirtualBox (QIIME Team, 2014). For the four studies discussed in this section, data analysis was conducted via QIIME 1.8.0 installed on local computer in Dr. Rothrock’s lab running Ubuntu 13.10.

4.4.3.1 Upstream data analysis using QIIME 1.8.0

The first step in upstream data analysis is demultiplexing and quality-filtering the data (Navas-Molina et al., 2013). For the four studies included in this section, demultiplexing and quality-filtering was performed with the QIIME *split_libraries_fastq.py* script, using the fastq files provided by the Argonne National Laboratory. This script demultiplexes by using a mapping file to replace the Golay barcodes with meaningful sample identifiers. Quality-filtering was performed with two different sets of parameters. The first set of parameters were the QIIME 1.8.0 defaults, which were three maximum consecutive low-quality base calls (*r=3*), 75% consecutive high-quality base calls (*p=0.75*), Q4 as the minimum Phred quality score (*q=3*), and zero ambiguous bases (*n=0*) (Navas-Molina et al., 2013). For later comparison, the data were also quality-filtered with Q20 as the minimum Phred quality score (*q=19*). To discover rare OTUs (Bokulich et al., 2013), the data were also quality-filtered with Q30 as the minimum Phred quality score (q=29) and 95% consecutive high-quality base calls (*p=0.95*).

After demultiplexing and quality-filtering, chimeric sequences were removed from the data. “Chimeras are hybrid products between multiple parent sequences that can be falsely interpreted as novel organisms, thus inflating apparent diversity” (Haas et al., 2011). As recommend by Navas-Molina and authors, chimeric sequences were identified by the UCHIME (Edgar, Haas, Clemente, Quince, & Knight, 2011) algorithm that is integrated into USEARCH 6.1 (Edgar, 2010). Compared to Chimera Slayer (Haas et al., 2011), UCHIME has been shown to have increased sensitivity and speed when detecting chimeric sequences (Edgar et al., 2011). In QIIME, chimeric sequences were identified via UCHIME by using the *identify_chimeric_seqs.py* script with the *-m usearch61* option. The *filter_fasta.py* script was then used to remove the list of identified chimeric sequences from the data.

The remainder of the upstream analysis was conducted in QIIME by applying the *pick_open_reference_otus.py* script to the forward reads. This script picks the operational taxonomic units (OTUs), assigns taxonomy, aligns the sequences, creates a phylogenetic tree, and generates the OTU tables used in downstream analysis (Caporaso, Kuczynski, et al., 2010). Picking OTUs (*pick_otus.py*) can be accomplished using a variety of methods. For these four studies, OTUs were picked using the recommended open-reference approach, in which sequences were matched using a subset of the Greengenes 13_8 database (DeSantis et al., 2006) filtered at 97% identity, and then non-matching sequences were added *de novo*, clustering OTUs at a 97% threshold to correspond approximately with species identity (Navas-Molina et al., 2013). To examine potential differences between clustering algorithms, clustering was performed with the QIIME 1.8.0 default of UCLUST (Edgar, 2010) and also with USEARCH 6.1.

Taxonomy was also assigned (*assign_taxonomy.py*) using two different algorithms: (1) UCLUST and (2) RDP Classifier (Wang, Garrity, Tiedje, & Cole, 2007). With both algorithms, taxonomy was assigned using the Greengenes 13_8 reference database. The primary taxonomy assignment algorithm used was the QIIME 1.8.0 default of UCLUST, using default parameters. For comparison to methods frequently employed in QIIME 1.7.0, the previous default taxonomy assignment of the RDP Classifier was used at two different minimum confidence levels for recording an assignment, the default level of 80% (c=0.8) and a reduced level of 60% (c=0.6).

Sequence alignment (*align_seqs.py*) was performed by PyNAST (Caporaso, Bittinger, et al., 2010) using the Greengenes core set (DeSantis et al., 2006), and a phylogenetic tree was constructed (*make_phylogeny.py*) by FastTree (Price, Dehal, & Arkin, 2010) for use in downstream analysis methods relying on phylogenetic distance. An OTU table was also generated (*make_otu_table.py*) in the BIOM format (McDonald et al., 2012) for downstream analysis.

4.4.3.2 Downstream data analysis using QIIME 1.8.0

Despite the overall high accuracy of Illumina sequencing, some false positive OTUs can appear as a result of errors in the sequencing process (Bokulich et al., 2013). In the absence of a mock bacterial community that can calibrate the upstream results, the recommended procedure is to discard OTUs with a number of sequences <0.005% of the total number of sequences (Bokulich et al., 2013; Navas-Molina et al., 2013). This can be accomplished in QIIME using the *filter_otus_from_otu_table.py* script with the --*min_count_fraction 0.00005* option (Navas-Molina et al., 2013).

The rest of the downstream analysis can be largely carried out with the *core_diversity_analyses.py* script, which is a QIIME workflow script that combines many other scripts such as: *core_diversity_analyses.py*, *beta_diversity_through_plots.py*, *summarize_taxa_through_plots.py*, *make_distance_boxplots.py*, *compare_alpha_diversity.py*, *otu_category_significance.py*, and *biom summarize-table*. The charts, plots, and statistics generated by the *core_diversity_analyses.py* script can be modified with a user-specified parameter file. Additionally, it is critical to note that these downstream analyses depend on the sequencing depth (Navas-Molina et al., 2013). While the appropriate sequencing depth will depend on the data, Navas-Molina and authors recommend at least “rarefying over 1000 sequences per sample for Illumina MiSeq, because samples below this level often suffer from other quality issues as well” (Navas-Molina et al., 2013).

4.4.4 Examples of the effects of differing data analysis methods

This subsection will briefly examine the variability associated with choosing different data analysis methods using QIIME 1.8.0 with four studies conducted by Dr. Rothrock. All of the studies are poultry-related, but their particular objectives are unimportant for this subsection. The purpose of this subsection is simply to demonstrate in broad terms how different data analysis methods can affect the findings.

4.4.4.1 Examples of effects on upstream data analysis

Table 2 and Table 3 are presented in Appendix A. The data in these tables are the output of the steps described in Section 4.4.3.1. Table 2 shows that the effect of removing chimeric sequences varies by study. Removing chimeric sequences had virtually no effect on the EggIsolate study, but about 1.5% of the sequences were removed by filtering for chimeras in the Production study. Table 2 also shows the effect of filtering by a Phred quality score of Q4 and better versus a score of Q20 and better. For two of the studies, only about 10% of additional sequences were removed by increasing the score to Q20, but for the other two studies, about 25% of additional sequences were removed.

Table 3 shows the effect of differing parameters on the number of OTUs, the number of taxa (to the L7, or species level), and the OTU table density (i.e. fraction of non-zero cells). Table 3 shows that the upstream data analysis methods can result in widely varying numbers of OTUs and taxa. Obviously, the number of OTUs and taxa identified in the upstream analysis affect the findings from the later downstream analysis. Therefore, it is critical when interpreting studies incorporating NGS to thoroughly evaluate the data analysis methods to ensure that appropriate and accepted techniques have been employed.

4.4.4.2 Examples of effects on downstream data analysis

Figure 2 and Figure 3 are presented in Appendix B. The data in these figures are the output of the steps described in Section 4.4.3.2. Because the steps used to generate Figure 2 and Figure 3 were identical except for the sequencing depth, these figures illustrate that downstream analysis is dependent on sequencing depth. In Figure 2, the sequencing depth is 220, but in Figure 3, the sequencing depth is 11,365. In each figure, part (a) is bar charts of taxa (to the L6, or genus level), part (b) is boxplots of the Chao1 richness estimate, and part (c) is rarefaction plots of the Chao1 richness estimate. The relatively low sequencing depth in Figure 2 has the advantage of including more samples, as any samples containing less than 11,365 sequences were excluded from analysis in Figure 3. The greater sequencing depth in Figure 3, however, has substantially increased estimates of the Chao1 richness estimate. The rarefaction curve in Figure 3 part (c) appears to level off considerably more than the curve in Figure 2 part (c). When the rarefaction curve levels off, it suggests that increasing the sampling depth beyond that point does not have a major impact on the estimate. Likewise, the non-leveling curve in Figure 2 part (c) suggests that a better estimate would be achieved by increasing the sampling depth. Despite the differences between the figures, it is interesting to note that the general relationships among the sample categories remain similar for this estimate in this study.

5. DEVELOPMENT OF A PUBLIC HEALTH AGENDA

NGS has the potential to offer unique solutions for improving food safety (Diaz-Sanchez, Hanning, Pendleton, & D’Souza, 2013). In addition to detecting molecular signatures of known pathogens, NGS techniques can also cluster non-matching sequences into novel OTUs and then assign taxonomy as appropriate (Edgar, 2010). Eventually, scientists may be able to move more of the “unknown agents” into the “known pathogens” category, which would improve foodborne illness estimates and hopefully provide more opportunities for successful food safety interventions. To help accomplish this goal, a public health agenda should include research aimed at determining how the microbiomes differ in humans with and without gastrointestinal illnesses. This research could attempt to establish illness patterns and likely determine novel pathogens.

NGS can also serve to improve food safety by providing better understanding of pathogen introduction, transportation, and fate. As previously mentioned, if there are unacceptable levels of pathogens on the final product, they had to come from somewhere. In this regard, Singer and authors argue that “to achieve further reductions in foodborne illness levels in humans, effective pre-harvest interventions are needed” (Singer et al., 2007). Therefore, in addition to using NGS on human samples, a public health agenda should include research using NGS at the farm level. Potential research questions could investigate topics such as (1) where known pathogens first become introduced or where they spike, (2) differences between samples collected on the farm, during processing, and on the final product, (3) and the effect of management practices on pathogen load, including examination of conventional and alternative farming practices.

Another critical aspect of a public health agenda is the development of more standardized methods that will allow for better comparisons among studies using NGS techniques related to food safety (Carriço et al., 2013). Standardizing methods will be a major challenge, however, because the pace of development is so rapid with both NGS instruments and with data analysis methods. Consequently, the current public agenda should probably focus on conducting research to determine the most useful methods for sampling, sequencing, and data analysis in regards to a particular food safety issue or product. Future research can then focus on standardizing methods to an extent where NGS can become a well-integrated aspect of the food safety system.

6. SUMMARY AND CONCLUSIONS

Producing accurate and precise estimates of the burden of foodborne diseases has proven to be challenging (Morris Jr, 2011). Part of the problem is that about 80% of foodborne illnesses have been attributed to “unknown agents” (Mead et al., 1999; Scallan, Griffin, et al., 2011; Scallan, Hoekstra, et al., 2011), and estimating the burden of unknown agents has considerable uncertainty associated with it (Frenzen, 2004; Morris Jr, 2011; Phillips & LaPole, 2003; Powell et al., 2001). Policymakers need accurate estimates of the burden of foodborne diseases so that they can have accurate representations of the magnitude and costs of foodborne diseases. Policymakers also need to be able to appropriately evaluate government-funded food safety initiatives and be able to improve these initiatives such that the burden of foodborne diseases will continue to decrease.

Next generation sequencing (NGS) may be able to improve foodborne illness estimates by identifying novel pathogens and consequently, reducing the percentage contribution of “unknown agents.” NGS is also promising in its ability to explore previously unsurmountable food safety research queries (Diaz-Sanchez et al., 2013). Hopefully, NGS can be implemented as a tool that will improve current trends of food safety stagnation by providing new insights into intervention strategies. Despite the powerful potential of NGS, researchers will face obstacles in creating standardized methodologies, especially in light of the rapid pace of NGS development (Carriço et al., 2013). In the meantime, researchers should continue striving to produce the best NGS methods for their particular research questions. In the future, full integration of NGS into the food safety system is likely to transform the practice of public health.

7. REFERENCES

Antle, J. M. (1999). Benefits and costs of food safety regulation. *Food policy, 24*(6), 605-623.

Archer, D. L., & Kvenberg, J. E. (1985). Incidence and cost of foodborne diarrheal disease in the United States. *Journal of food protection (USA)*.

ARS. (2014). Agricultural Research Service (ARS): About Us. Retrieved April 27, 2014, from http://www.ars.usda.gov/AboutUs/AboutUs.htm

Batz, M. B., Doyle, M. P., Morris Jr, J. G., Painter, J., Singh, R., Tauxe, R. V., . . . Lo Fo Wong, D. (2005). Attributing illness to food. *Emerging infectious diseases, 11*(7).

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., . . . Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods, 10*(1), 57-59.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics, 26*(2), 266-267.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods, 7*(5), 335-336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., . . . Bauer, M. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal, 6*(8), 1621-1624.

Carriço, J., Sabat, A., Friedrich, A., Ramirez, M., & Markers, E. S. G. f. E. (2013). Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. *Euro surveillance: bulletin Européen sur les maladies transmissibles= European communicable disease bulletin, 18*(4), 20382.

Centers for Disease Control and Prevention. (1997). Foodborne Diseases Active Surveillance Network, 1996. *MMWR. Morbidity and mortality weekly report, 46*(12), 258.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods, 69*(2), 330-339.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews, 17*(4), 840-862.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology, 72*(7), 5069-5072.

Dewaal, C. S., Hicks, G., Barlow, K., Alderton, L., & Vegosen, L. (2006). Foods associated with foodborne illness outbreaks from 1990 through 2003. *Food protection trends, 26*(7), 466-473.

Diaz-Sanchez, S., Hanning, I., Pendleton, S., & D'Souza, D. (2013). Next-generation sequencing: The future of molecular genetics in poultry production and food safety. *Poultry science, 92*(2), 562-572.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics, 27*(16), 2194-2200.

Frenzen, P. D. (2004). Deaths due to unknown foodborne agents. *Emerging infectious diseases, 10*(9), 1536.

Gabriel, I., Lessire, M., Mallet, S., & Guillot, J. (2006). Microflora of the digestive tract: critical factors and consequences for poultry. *World's poultry science journal, 62*(03), 499-511.

General Accounting Office. (2001). *Food Safety: Overview of Federal and State Expenditures*. (GAO-01-177). United States General Accounting Office.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources, 11*(5), 759-769.

Google Scholar. (2014). Google Scholar. Retrieved April 27, 2014, from http://scholar.google.com/

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., . . . Sodergren, E. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research, 21*(3), 494-504.

Karrow, J., & Toner, B. (2011). In Sequence Annual Survey: Illumina Leads Market but Most Users Believe PacBio will Provide the Next Big Leap. *Genome Web, January, 25*, 2011.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell, 155*(1), 27-38.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., . . . Meyer, F. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience, 1*(1), 7.

Mead, P. S., Slutsker, L., Dietz, V., McCaig, L. F., Bresee, J. S., Shapiro, C., . . . Tauxe, R. V. (1999). Food-related illness and death in the United States. *Emerging infectious diseases, 5*(5), 607.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Morris Jr, J. G. (2011). How safe is our food? *Emerging infectious diseases, 17*(1), 126-128.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., . . . Knight, R. (2013). Chapter Nineteen - Advancing Our Understanding of the Human Microbiome Using QIIME. In F. D. Edward (Ed.), *Methods in Enzymology* (Vol. Volume 531, pp. 371-444): Academic Press.

Oakley, B. B., Morales, C. A., Line, J., Berrang, M. E., Meinersmann, R. J., Tillman, G. E., . . . Seal, B. S. (2013). The poultry-associated microbiome: network analysis and farm-to-fork characterizations. *PLoS One, 8*(2), e57190.

Phillips, C. V., & LaPole, L. M. (2003). Quantifying errors without random sampling. *BMC Medical Research Methodology, 3*(1), 9.

Pires, S. M., Evers, E. G., van Pelt, W., Ayers, T., Scallan, E., Angulo, F. J., . . . Hald, T. (2009). Attributing the human disease burden of foodborne infections to specific sources. *Foodborne Pathogens and Disease, 6*(4), 417-424.

Powell, M., Ebel, E., & Schlosser, W. (2001). Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. *International journal of food microbiology, 69*(3), 209-215.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One, 5*(3), e9490.

QIIME Team. (2014). QIIME: Quantitative Insights Into Microbial Ecology. Retrieved April 27, 2014, from http://qiime.org/

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology, 57*(1), 369-394.

Rothrock, J., Michael J., Hiett, K. L., Caudill, A. C., Cicconi-Hogan, K. L., & Caporaso, J. G. (2014). A semi-automated, hybrid DNA extraction method for the qualitative and quantitative assessment of bacterial communities from environmental poultry production samples. Manuscript submitted for publication.

Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R. V., & Hoekstra, R. M. (2011). Foodborne illness acquired in the United States—unspecified agents. *Emerging infectious diseases, 17*(1), 16.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M.-A., Roy, S. L., . . . Griffin, P. M. (2011). Foodborne illness acquired in the United States--major pathogens. *Emerging infectious diseases, 17*(1).

Scharff, R. L. (2010). Health-related costs from foodborne illness in the United States.

Scharff, R. L. (2012). Economic burden from health losses due to foodborne illness in the United States. *Journal of Food Protection®, 75*(1), 123-131.

Sekelja, M., Rud, I., Knutsen, S., Denstadli, V., Westereng, B., Næs, T., & Rudi, K. (2012). Abrupt temporal fluctuations in the chicken fecal microbiota are explained by its gastrointestinal origin. *Applied and environmental microbiology, 78*(8), 2941-2948.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology, 26*(10), 1135-1145.

Singer, R. S., Cox Jr, L. A., Dickson, J. S., Hurd, H. S., Phillips, I., & Miller, G. Y. (2007). Modeling the relationship between food animal health and human foodborne illness. *Preventive veterinary medicine, 79*(2), 186-203.

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology, 39*(1), 321-346.

Struelens, M., & Brisse, S. (2013). From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Euro Surveill, 18*, 20386.

Todd, E. C. (1989). Preliminary estimates of costs of foodborne disease in the United States. *Journal of food protection, 52*.

USDA. (2013). *Fiscal Year 2013 Agency Financial Report*. United States Department of Agriculture.

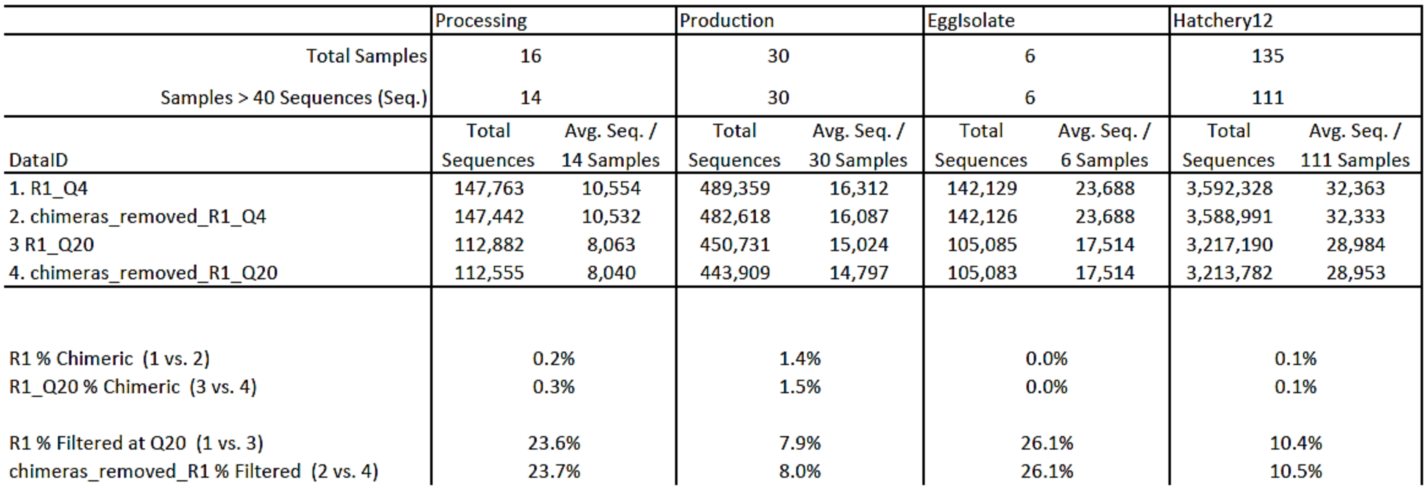
Voetsch, A. C., Van Gilder, T. J., Angulo, F. J., Farley, M. M., Shallow, S., Marcus, R., . . . Tauxe, R. V. (2004). FoodNet estimate of the burden of illness caused by nontyphoidal Salmonella infections in the United States. *Clinical infectious diseases, 38*(Supplement 3), S127-S134.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology, 73*(16), 5261-5267.

APPENDICES

Appendix A

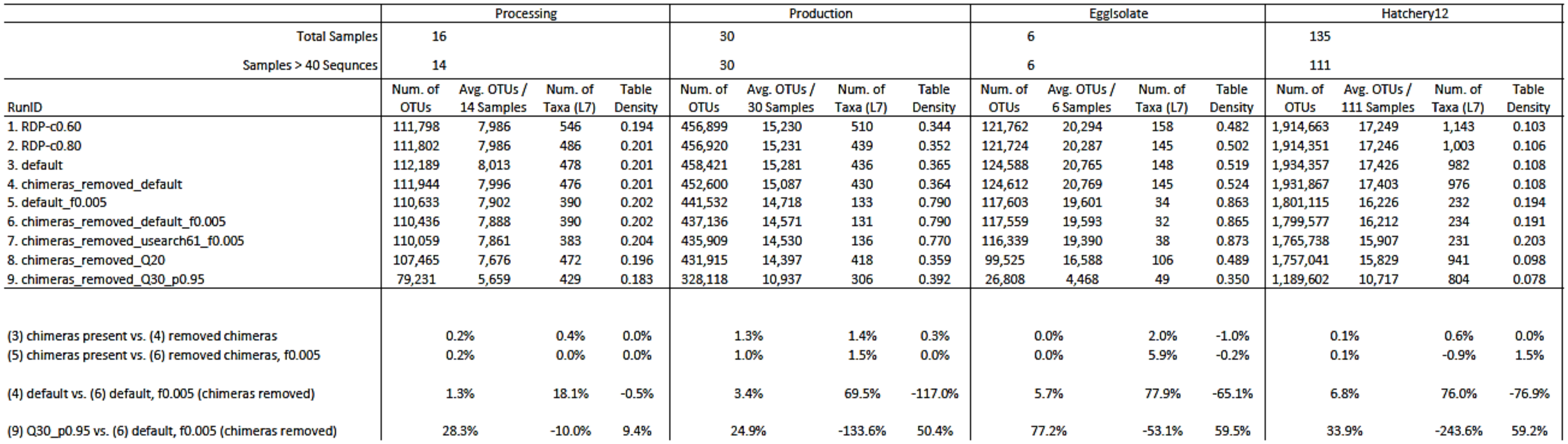
Table 2. **Effects of Quality and Chimera Filtering on the Number of Sequences**



	Processing		Production		EggIsolate		Hatchery12	
Total Samples	16		30		6		135	
Samples > 40 Sequences (Seq.)	14		30		6		111	
DataID	Total Sequences	Avg. Seq. / 14 Samples	Total Sequences	Avg. Seq. / 30 Samples	Total Sequences	Avg. Seq. / 6 Samples	Total Sequences	Avg. Seq. / 111 Samples
1. R1_Q4	147,763	10,554	489,359	16,312	142,129	23,688	3,592,328	32,363
2. chimeras_removed_R1_Q4	147,442	10,532	482,618	16,087	142,126	23,688	3,588,991	32,333
3 R1_Q20	112,882	8,063	450,731	15,024	105,085	17,514	3,217,190	28,984
4. chimeras_removed_R1_Q20	112,555	8,040	443,909	14,797	105,083	17,514	3,213,782	28,953
R1 % Chimeric (1 vs. 2)	0.2%		1.4%		0.0%		0.1%	
R1_Q20 % Chimeric (3 vs. 4)	0.3%		1.5%		0.0%		0.1%	
R1 % Filtered at Q20 (1 vs. 3)	23.6%		7.9%		26.1%		10.4%	
chimeras_removed_R1 % Filtered (2 vs. 4)	23.7%		8.0%		26.1%		10.5%	

Table 3. **Effects of Data Analysis Parameters**

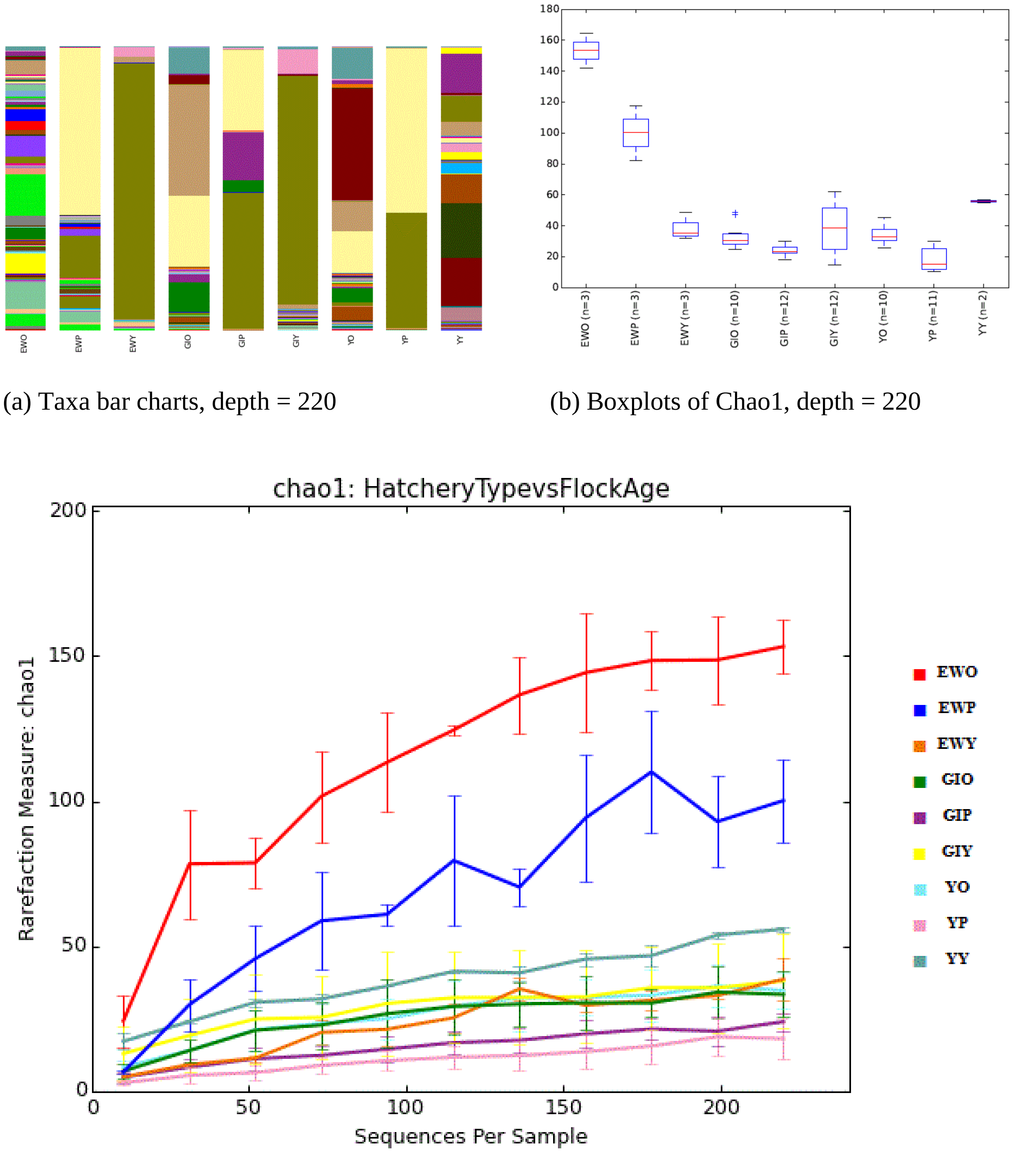
RunID	Quality Filtering	OTU Picking	Chimeras	Taxonomy Assignment	OTU Abundance Filtering
1. RDP-c0.60	Default	UCLUST (UCLUST defaults)	Present	RDP (c = 0.6)	None
2. RDP-c0.80	Default	UCLUST (UCLUST defaults)	Present	RDP (c = 0.8)	None
3. default	Default	UCLUST (QIIME defaults)	Present	UCLUST	None
4. chimeras_removed_default	Default	UCLUST (QIIME defaults)	Removed	UCLUST	None
5. default_f0.005	Default	UCLUST (QIIME defaults)	Present	UCLUST	<0.005%
6. chimeras_removed_default_f0.005	Default	UCLUST (QIIME defaults)	Removed	UCLUST	<0.005%
7. chimeras_removed_usearch61_f0.005	Default	USEARCH 6.1 (QIIME defaults)	Removed	UCLUST	<0.005%
8. chimeras_removed_Q20	≥ Q20	UCLUST (QIIME defaults)	Removed	UCLUST	None
9. chimeras_removed_Q30_p0.95	≥ Q30, p=0.95	UCLUST (QIIME defaults)	Removed	UCLUST	None



	Processing				Production				EggIsolate				Hatchery12			
Total Samples	16				30				6				135			
Samples > 40 Sequnces	14				30				6				111			
RunID	Num. of OTUs	Avg. OTUs / 14 Samples	Num. of Taxa (L7)	Table Density	Num. of OTUs	Avg. OTUs / 30 Samples	Num. of Taxa (L7)	Table Density	Num. of OTUs	Avg. OTUs / 6 Samples	Num. of Taxa (L7)	Table Density	Num. of OTUs	Avg. OTUs / 111 Samples	Num. of Taxa (L7)	Table Density
1. RDP-c0.60	111,798	7,986	546	0.194	456,899	15,230	510	0.344	121,762	20,294	158	0.482	1,914,663	17,249	1,143	0.103
2. RDP-c0.80	111,802	7,986	486	0.201	456,920	15,231	439	0.352	121,724	20,287	145	0.502	1,914,351	17,246	1,003	0.106
3. default	112,189	8,013	478	0.201	458,421	15,281	436	0.365	124,588	20,765	148	0.519	1,934,357	17,426	982	0.108
4. chimeras_removed_default	111,944	7,996	476	0.201	452,600	15,087	430	0.364	124,612	20,769	145	0.524	1,931,867	17,403	976	0.108
5. default_f0.005	110,633	7,902	390	0.202	441,532	14,718	133	0.790	117,603	19,601	34	0.863	1,801,115	16,226	232	0.194
6. chimeras_removed_default_f0.005	110,436	7,888	390	0.202	437,136	14,571	131	0.790	117,559	19,593	32	0.865	1,799,577	16,212	234	0.191
7. chimeras_removed_usearch61_f0.005	110,059	7,861	383	0.204	435,909	14,530	136	0.770	116,339	19,390	38	0.873	1,765,738	15,907	231	0.203
8. chimeras_removed_Q20	107,465	7,676	472	0.196	431,915	14,397	418	0.359	99,525	16,588	106	0.489	1,757,041	15,829	941	0.098
9. chimeras_removed_Q30_p0.95	79,231	5,659	429	0.183	328,118	10,937	306	0.392	26,808	4,468	49	0.350	1,189,602	10,717	804	0.078
(3) chimeras present vs. (4) removed chimeras	0.2%		0.4%	0.0%	1.3%		1.4%	0.3%	0.0%		2.0%	-1.0%	0.1%		0.6%	0.0%
(5) chimeras present vs. (6) removed chimeras, f0.005	0.2%		0.0%	0.0%	1.0%		1.5%	0.0%	0.0%		5.9%	-0.2%	0.1%		-0.9%	1.5%
(4) default vs. (6) default, f0.005 (chimeras removed)	1.3%		18.1%	-0.5%	3.4%		69.5%	-117.0%	5.7%		77.9%	-65.1%	6.8%		76.0%	-76.9%
(9) Q30_p0.95 vs. (6) default, f0.005 (chimeras removed)	28.3%		-10.0%	9.4%	24.9%		-133.6%	50.4%	77.2%		-53.1%	59.5%	33.9%		-243.6%	59.2%

Appendix B

Figure 2. **Downstream Data Analysis Results at a Sequencing Depth of 220**

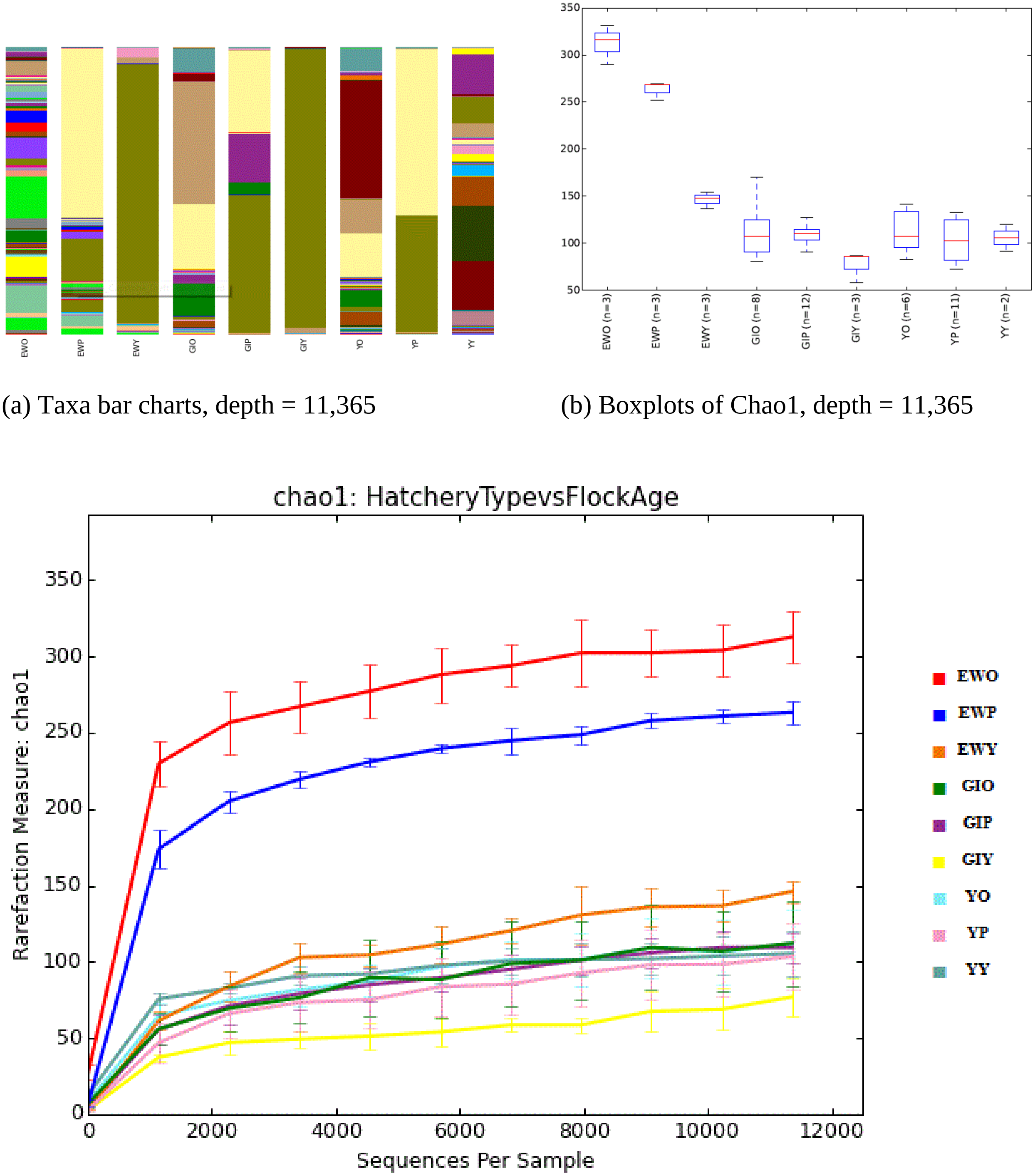


(a) Taxa bar charts, depth = 220

(b) Boxplots of Chao1, depth = 220

(c) Rarefaction plot of Chao1, depth = 220

Figure 3. **Downstream Data Analysis Results at a Sequencing Depth of 11,365**



(a) Taxa bar charts, depth = 11,365

(b) Boxplots of Chao1, depth = 11,365

(c) Rarefaction plot of Chao1, depth = 11,365