Potential contribution of next generation sequencing (NGS) to improving food safety and the estimates of the burden of foodborne diseases

College of Public Health, University of Georgia

INTRODUCTION

Next generation sequencing (NGS) holds significant promise for improving public health[1]. Nevertheless, transforming powerful NGS techniques into practical public health solutions will be a major challenge in the coming years[2]. The purpose of this capstone project 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.

During May to August 2013, collected and prepared poultry-related samples for NGS. gained 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[3]. This capstone project incorporates some of the methods used during Mr. Caudill's internships and integrates learning relevant to Environmental Health Science, Epidemiology, and Bioinformatics.

Table 1. Estimates of the Annual Burden of Foodborne Pathogens in the U.S.

(a) from Mead et al., 1999[4]

	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%)

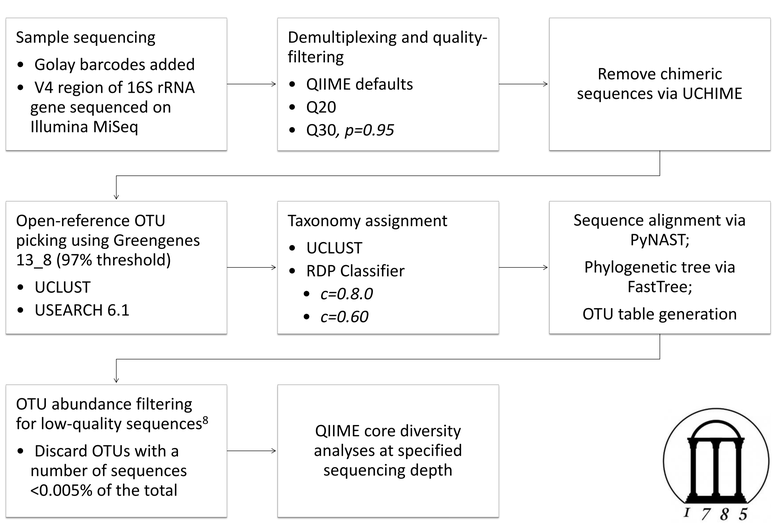
(b) from Scallan, Hoekstra, et al., 2011[5] and from Scallan, Griffin, et al., 2011[6]

	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.

METHODS

Figure 1. Flowchart of NGS Data Analysis[7] Used in Four Poultry-Related Studies



RESULTS

Table 2 shows the effect of differing algorithms and 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 2 shows that the upstream data analysis methods can result in widely varying numbers of OTUs and taxa. 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.

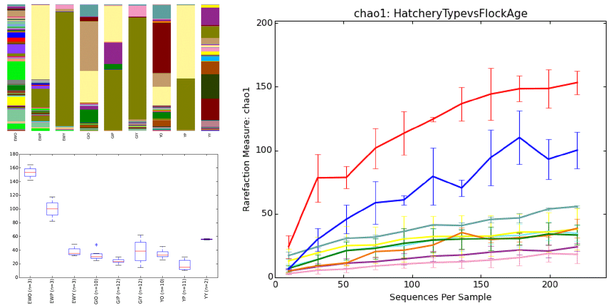
Table 2a. Explanation of Differing Data Analysis Algorithms and Parameters

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

Table 2b. Effects of Differing Data Analysis Algorithms and Parameters

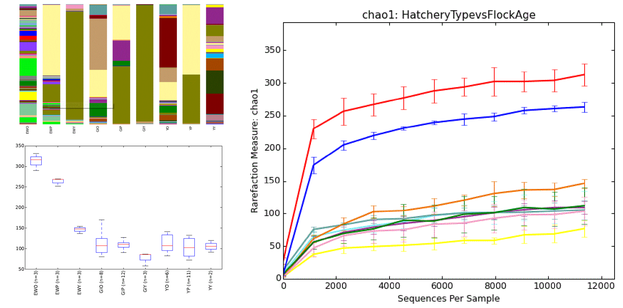
	Processing				Production				EggIsolate				Hatchery12			
Total Samples	16				30				6				135			
Samples > 40 Sequnces	14				30				6				111			
RunID	Num. of OTUs	Avg. OTUs / 14 Samples	Num. of Taxa (L7)	Table Density	Num. of OTUs	Avg. OTUs / 30 Samples	Num. of Taxa (L7)	Table Density	Num. of OTUs	Avg. OTUs / 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%

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



Bar charts of taxa (genus, L6); box plots and alpha rarefaction plot of Chao1 richness estimate

Figure 2b. Downstream Data Analysis Results at a Sequencing Depth of 11,365



Bar charts of taxa (genus, L6); box plots and alpha rarefaction plot of Chao1 richness estimate

SUMMARY AND CONCLUSIONS

Producing reliable estimates of the burden of foodborne diseases has proven to be challenging[9]. Part of the problem is that about 80% of foodborne illnesses have been attributed to "unknown agents"[4-6], and estimating the burden of unknown agents has considerable uncertainty associated with it[9-12]. 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[13]. Hopefully, NGS can be implemented as a tool that will improve current trends of food safety stagnation[9] 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[2]. In the future, full integration of NGS into the food safety system is likely to transform the practice of public health.

REFERENCES

1. Struelens M, Brisse S. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. Euro Surveill. 2013;18:20386.
2. Carriço J, Sabat A, Friedrich A, Ramirez M, Markers ESGfE. 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. 2013;18(4):20382.
3. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335-336.
4. Mead PS, Slutsker L, Dietz V, et al. Food-related illness and death in the United States. Emerging infectious diseases. 1999;5(5):607.
5. Scallan E, Hoekstra RM, Angulo FJ, et al. Foodborne illness acquired in the United States--major pathogens. Emerging infectious diseases. 2011;17(1).
6. Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM. Foodborne illness acquired in the United States—unspecified agents. Emerging infectious diseases. 2011;17(1):16.
7. Navas-Molina JA, Peralta-Sánchez JM, González A, et al. Chapter Nineteen - Advancing Our Understanding of the Human Microbiome Using QIIME. In: Edward FD, ed. Methods in Enzymology. Vol Volume 531: Academic Press; 2013:371-444.
8. Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods. 2013;10(1):57-59.
9. Morris Jr JG. How safe is our food? Emerging infectious diseases. 2011;17(1):126-128.
10. Frenzen PD. Deaths due to unknown foodborne agents. Emerging infectious diseases. 2004;10(9):1536.
11. Phillips CV, LaPole LM. Quantifying errors without random sampling. BMC Medical Research Methodology. 2003;3(1):9.
12. Powell M, Ebel E, Schlosser W. Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. International Journal of Food Microbiology. 2001;69(3):209-215.
13. Diaz-Sanchez S, Hanning I, Pendleton S, D'Souza D. Next-generation sequencing: The future of molecular genetics in poultry production and food safety. Poultry science. 2013;92(2):562-572.

ACKNOWLEGEMENTS

I would like to thank Dr. Michael J. Rothrock Jr. and Dr. Kelli L. Hiett for their support and training at the USDA Richard B. Russell Research Center. I would also like to thank Dr. Ming Zhang for reviewing my capstone project and Dr. Travis C. Glenn for introducing me to next generation sequencing technologies.