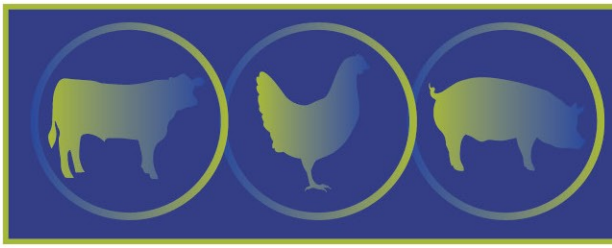

Whole Genome Sequencing

Introduction to the Interpretation of Whole Genome Sequence Data in Food Safety

Introduction

Whole genome sequencing (WGS) for bacterial foodborne pathogen characterization is here to stay. Developments in WGS platforms have made it possible to sequence the entire genome of bacteria at prices comparable to common molecular subtyping methods. These data can provide near-perfect discrimination of bacterial isolates. While common molecular subtyping methods only interrogate small parts of the genome (e.g., for **pulsed-field gel electrophoresis [PFGE]** restriction sites, MLST sequence of ~7 loci of ~500 **basepairs [bp]**), WGS approaches make it possible to interrogate more than 99% of the genome, which translates to approximately 2.8 million and 4.8 million basepairs in *Listeria monocytogenes* and *Salmonella enterica*, respectively.

What's more, as sequencing technologies and data analytics continue to mature, WGS will provide results at costs and timeframes cheaper/faster than traditional subtyping. U.S. Governmental agencies (CDC, FDA, and USDA-FSIS) are beginning to build large, shared databases (e.g. GenomeTrakr) to store WGS data generated from foodborne pathogen isolates collected from routine surveillance or human disease cases to compare records between isolates and use these comparisons to inform regulatory action.

While the field is beginning to coalesce around common WGS sequencing platforms and basic data analysis approaches *(1)*, the application of those platforms and approaches to bacterial food safety has not yet matured. The purpose of this article is to:

1. Introduce whole genome sequencing platforms and analytical approaches to practitioners who have not yet encountered these data in their work;
2. Describe dimensions of WGS analysis for which there are still significant ambiguity in the scientific approaches; and
3. Summarize some outstanding challenges to the application of these methods to bacterial foodborne pathogen subtyping.

We will introduce whole genome sequencing technologies and common analyses, then discuss the application of these methods to regulatory action and different foodborne pathogens.

Sequencing technologies

Sequencing technologies used for WGS can be subdivided into two categories; (i) **short-read technologies**, which produce sequence reads up to 500 bp (e.g., Illumina, IonTorrent), and (ii) **long-read technologies**, which produce reads longer than 1000 bp and often lengths over 70,000 bp (e.g., Pacific Biosciences, Oxford Nanopore). At the time of writing this document (May 2016), the two sequencing platforms most commonly used in WGS are Illumina and Pacific Biosciences. Illumina sequencers (e.g., MiSeq, NextSeq, HiSeq) are popular because of speed, throughput and high accuracy of the data produced by these sequencers, allowing bacterial genomes to be sequenced at low costs (between $50 and $100 per bacterial genome). The per bacterial genome costs of Pacific Biosciences sequencers are considerably more expensive (>$800), making it cost prohibitive for WGS-based typing.

Short-read technologies are best suited for high-throughput applications due to high accuracy and low costs per base sequenced. They are the main technologies used for routine WGS by government agencies and are used for whole genome analogs to nucleotide-based subtyping schemes, such as whole-genome **Single Nucleotide Polymorphism (SNP)** or **Multi Locus Sequence Typing (MLST)**

analysis. Gaps in genome sequencing (see 'data analysis – genome assembly') prevent interrogation of genome-scale events, such as genome rearrangements or differences in PFGE patterns.

Long-read sequencing technologies can complement short-read technologies, at a higher cost and lower throughput. Their main advantage is that longer reads can often be assembled *de novo* into a complete genome, either alone or in combination with short-read data. In principle, long-read data could be used to directly calculate PFGE pattern profiles for comparison to existing databases.

Data analysis

The field of computer science called **bioinformatics** is used to analyze WGS data. This involves algorithm-, pipeline- and software development, analysis, transfer and storage/ database development of genomics data.

A typical WGS workflow contains the following steps; (i) quality control and data grooming, (ii) genome assembly and/or variant calling, and (iii) post-assembly analysis. Current academic reviews, such as *(1)*, give more detail on these steps than what follows below.

Data quality control and data grooming

Quality control of WGS data involves multiple aspects, but some of the most important involves **read quality** (e.g., how may sites of a 300 bp read fall below a specific quality threshold), **fold coverage or sequence depth** and **putative contaminants**. Read quality is usually dealt with by data grooming, i.e., removal of low-quality regions of the individual reads with specialized bioinformatics tools. The second aspect involves fold coverage. WGS data typically consists of hundreds of thousands of short sequence reads representing fragments of the genome. Fold coverage or sequence depth refers to the median or average number of reads that cover each nucleotide in a genome. Too low coverage will influence the accuracy of downstream analyses, as will too high coverage. A commonly overlooked aspect of data quality is contamination with a non-target organism, which can be a laboratory-introduced contamination or an organism that is co-isolated. This may not pose problems for some downstream analyses, such as reference-based assembly for outbreak detection, but may be problematic for gene detection-based applications such as WGS-based screening of antibiotic-resistance genes.

Genome assembly and/or variant calling

The main objective of WGS analysis is the identification of genomic differences between bacterial strains. Since the raw data of WGS technology are bacterial genome sequence fragments of various size (from 100s-10,000+ bp), a fundamental question is how to use those fragments to determine genomic differences, referred to as genomic variants. Bioinformatics pipelines are tools that identify these variants, and are generally referred to as 'variant' callers. Genomic variants include (i) single nucleotide polymorphisms (SNPs) or **single nucleotide variants** (**SNV**s), which indicate a single nucleotide substitution difference between genomes, (ii) **insertions** and **deletions** of nucleotide/s (commonly referred to as **indels**), and (iii) **genomic rearrangements**.

One approach to detect variants is to first **assemble** genomes ***de novo*** and then use **whole genome alignment**-based methods to compare two or more strains. *De novo* assembly of short-read sequences generally yields so-called **draft genome sequences**, genome sequences that still contain gaps. These gaps are generally caused by the presence of repetitive sequences (e.g., rRNA sequence clusters) in the genome. Recent bioinformatic advances in the assembly strategies of long reads from Pacific Biosciences sequencing technologies now make it possible to produce *de novo* **closed genome sequences** (i.e., sequences without gaps).

A second approach is the **reference mapping approach.** In this approach, reads are aligned (mapped) against a (preferably closed) reference genome. After mapping, variants are called from the consensus of the mapped reads. Reference mapping-based approaches are very popular because they are computationally inexpensive and are fast compared to *de novo* assembly. A limitation of this method is the reliance on a reference genome. Especially if a closely related reference genome is absent, mapping against a distantly

related genome may lead to problems with variant calling, and unique regions not found in the reference sequence will not be included in downstream analysis.

In addition to *de novo* assembly and reference mapping-based methods, reference-free *de novo* variant calling methods exist. These methods do not require reference sequences and are faster than *de novo* assembly-based methods.

Post assembly analysis

One underappreciated subtlety in WGS analysis is how to interpret the genetic variants between strains as biologically relevant measures of strain difference. This problem has two facets: (i) determining which differences matter and how to count them, and (ii) visualizing the differences in a way that can guide action.

Most WGS analyses use SNPs as the primary measure of genetic distance, although other methods include whole-genome multilocus sequence typing (MLST) and gene presence/absence tables. Concerning SNP differences, one could count: core SNPs, at sites present in all strains in the analysis, pan-genome SNPs, at any site present in 2 or more strains, or SNPs present in some percentage of the strains. SNP sets can also be filtered to exclude sites likely introduced by recombination. The point is that the statement 'Isolate A differs from isolate B by < X SNPs' should later specify what SNPs are being counted. Once a SNP set is determined choices must be made about how to interpret the distances. The simplest is to count SNP differences between sampled isolates.

In practice, relatedness of bacterial strains is commonly displayed in the form of a **phylogenetic tree.** These trees represent a plausible 'path' of relatedness where the observed differences are mapped onto hypothetical divergences from common ancestors. Generally only SNPs from the core genome are used to infer phylogenetic trees. Phylogenetic inference methods can be subdivided into four different classes based on the underlying method used to identify the best trees; parsimony, maximum likelihood, Bayesian and distance methods. To assess confidence in the resulting phylogenetic trees, branch-support statistics can be inferred using statistical methods such as the bootstrap for parsimony, maximum likelihood and distance methods and posterior probabilities for Bayesian methods.

The next challenge is to visualize the differences in a way that can guide action, such as identifying a plausible outbreak or source of contamination. When distances are counted, a reasonable visualization approach is to plot, or report, the differences between groups of strains. For example, one can plot the number of pairwise SNP differences between bacterial isolates within or between outbreaks *(2)*, or individual food establishments *(3)*. Another common choice is to build a phylogenetic tree that displays the calculated evolutionary model of the isolates as a series of splits from a root state. In these, clusters of isolates near the 'tips' of the tree are more closely related to each other than isolates elsewhere in the tree. As a hybrid approach, one can combine SNP counting and phylogenetic tree production to find clusters of isolates and then report the differences between the number of SNPs. The CDC, FDA, and FSIS all use some variation on this hybrid approach to assess if newly sequenced isolates are highly genetically related to (by evolutionary or simple SNP distance) to clusters of isolates already characterized.

Current regulatory use

The U.S. government has been routinely gathering WGS information on foodborne pathogen isolates since 2013 as part of the GenomeTrakr project *(4)*. In that year, the CDC and FDA began a partnership with the goal to sequence all *Listeria monocytogenes* isolates collected from clinical, food, and environmental sources. In addition, selected states began to sequence clinical, food, and environmental *Salmonella enterica* isolates to improve discrimination within that common serotype. Sequencing activities have since expanded to include *Campylobacter*, and STEC pathogens.

Once sequenced, government agencies now routinely use bioinformatics pipelines (details below) to improve the discrimination of foodborne pathogens beyond the previous gold-standard

techniques, such as PFGE, at comparable time frames. As with previous molecular subtyping approaches, groups of similar isolates are used to inform epidemiologic and regulatory actions. But the increased sensitivity of the approach leads to a few novel outcomes, highlights being:

1. Increased sensitivity is allowing for more and smaller foodborne disease outbreaks to be identified and linked to a causal food source ('solved').
    a. In the year before WGS was implemented for *Listeria*, September 2012-2013, 2 *L. monocytogenes* outbreaks were solved, with a median of 6 cases per outbreak. In the following two years, 5 and then 9 outbreaks were solved, with a median of 4 and then 3 cases per outbreak in Sept. 2013-2014 and 2014-2015.
2. Routine WGS of isolates from regulatory activities are linking products to foodborne disease outbreaks.
    a. The Spring 2016 outbreak of listeriosis in leafy greens was first putatively linked to a Dole plant in OH when WGS showed a close match between a cluster of human illness and an isolate collected during routine sampling of retail products by the Ohio Dept. of Agriculture *(5)*.
3. Even sporadic cases of foodborne disease can now be linked to specific products.
    a. WGS linked an *L. monocytogenes* isolated from retail lettuce testing in Canada to an isolate from a listeriosis patient who probably consumed the product *(6)*. Note this report was published in the same month as the Dole *Listeria* recall mentioned above.
4. Cases are being retrospectively added to current outbreaks based on WGS data.
    a. During the Blue Bell outbreak in 2015, 6 cases of listeriosis with illness occurring as early as 2010 were added to the outbreak case count *(7)*.

The WGS sequence data is deposited to the public National Center for Biotechnology Information (NCBI) where it is freely available to the public. Additional non-sequence-related information (i.e. strain source, geographical location, etc.) for each strain is also available; however, confidential information such as patient identification or food manufacturer source of the isolate is not.

Bioinformatics pipelines

In bioinformatics, a **pipeline** is a standardized set of analysis programs run on standardized input data. Three government agencies, NCBI, FDA, and FSIS, have all developed their own bacterial WGS analysis pipelines for analyzing newly sequenced isolates, identifying genetic differences, and interpreting those differences (summarized in **Table 1**). In addition, many academic researchers have developed pipelines for similar analysis, some of which are publically available.

Across the field, pipelines to translate WGS raw data to biological insights are still developing. While they all address the core problem of translating raw read data into markers of genetic distance, e.g. SNPs, they all differ in the exact algorithms used, the availability of the pipeline and its inputs, and the interpretation of the results. In particular, different pipelines might result in different SNP sets that identify differences between isolates, so two isolates are unlikely to differ by the same number of SNPs as analyzed by the NCBI, FDA, and FSIS. And the systematic comparison of these pipelines, ideally by independent academic research, which would help understand the implications of these differences is complicated by the lack of full transparency of the pipelines. While the FDA/CFSAN pipeline has been peer-reviewed *(8)* and is available for use by other researchers, both the NCBI and FSIS pipelines are not publicly available. However, the SNPs distance outputs from the NCBI pipeline are available for download and updated as new strains are added. Increasing transparency will improve the ability of academic and industry bioinformaticians to replicate, and then extend, the government results for their own use.

Foodborne pathogens

While the basic WGS approach is universal across foodborne pathogens, the analysis and interpretation of those results are not likely to be universal across *L. monocytogenes, Salmonella*, *Campylobacter*, and the STEC that are currently being sequenced (**Table 2**).

At the most simplistic level, each organism has a different genome size (ranging from 1.7 mb to 5.6 mb). WGS workflows that require individual nucleotides to be sequenced a given average number of times may need to be adjusted by species, for example, by adjusting the number of isolates that can be sequenced in a single sequencing reaction.

At a more complicated level, the population structure, rates of evolution, and genomic features differ between and within species. For example, *L. monocytogenes* is a relatively clonal organism, with minimum genomic variation between isolates and a relatively small genome, making it relatively easy to perform high-throughput sequencing and then identify a small number of SNPs that differentiate organisms in a biologically meaningful way. In contrast, *Salmonella* has a much more variable genome that is significantly larger, which means the standard approaches to identifying SNP differences between isolates tend to find many more SNPs. And even within species, organisms can evolve at different rates. Organisms persisting within a relatively cold food processing facility are likely to grow more slowly and accumulate mutations at a slower rate than organisms passing through a warm-blooded host, for example. Finally, organisms might present different regions of recombination, plasmid or prophage presence, or phenotypic traits such as antimicrobial resistance. It is unclear what implications all of these sources of genetic variation have for the current major use of WGS in food safety and the identification of closely related isolates for outbreak identification and source traceback.

Besides grouping strains based on SNP differences, methods are available to predict relevant strain attributes directly from WGS data, such as patterns of antimicrobial resistance and virulence gene presence/absence (9). Several databases are available for virulence and antimicrobial-resistance genes to screen WGS sequence. However, databases are only as good as the data they contain, so results need to be carefully scrutinized to make sure erroneous assignments are removed from the final results. Also, just because a genome has a predicted virulence or antimicrobial-resistant gene doesn't mean that it will be functional in the host. For example, a strain's virulence phenotype may not match virulence predicted from WGS analysis.

Conclusion

While WGS-based methods will allow for significant improvements in surveillance and foodborne disease outbreak detection, in the vast majority of outbreak investigations, strong epidemiological data are still essential to conclusively identify the source of a foodborne disease outbreak. In addition, interpretation of SNP differences remains a challenge as, at least theoretically, genetically identical isolates can be found in different locations. (It may take >2,000 generations for bacteria to accumulate a single SNP, a time frame that also allows for efficient dispersal through fomites or vectors). In addition, isolates may accumulate SNPs quickly and it has been shown that SNP differences can occur during lab passage or passage through a human host.

WGS analysis is complicated. To be able to carry out meaningful work under all those complications, U.S. government agencies have developed bioinformatics pipelines that make some standardized assumptions to lead to reproducible results. Those pipelines are being used by public officials to track foodborne disease outbreaks and perform source attribution. The food industry will need to consider the major assumptions in government analyses as it develops pipelines appropriate to their food safety operations.

Academics should play a role in the application of these WGS technologies to industry challenges – asking which areas in food production/processing can WGS results help make informed decisions. Simple random "we can look for it" sampling and sequencing is not as beneficial as targeted applications. For example, if industry is making major reconstruction/remodeling of a plant, WGS-informed surveillance could help ensure a perturbed environment does not introduce a resident, previously hidden pathogen (*Listeria* behind a wall being removed, for example). In this instance, WGS could identify how any new isolates relate to previous environmental or outbreak strains, which could inform control measures. This would be a proactive application of WGS-enhanced environmental control.

At the same time, the conclusions drawn by all parties are only as good as the reference databases and data analyses available. The current GenomeTrakr database has a wide variation in coverage by organism and isolate source; methods to compare new isolates to the database are not yet standardized. Expansion of reference databases and improved bioinformatics pipelines could considerably improve the ability to interpret the results of sequencing new organisms.

Further research on short-term (or micro-) evolution and population genomics of foodborne pathogens, including in-food associated environments, is a considerable need. This research will create the knowledge to facilitate improved interpretation of WGS data in the context of foodborne disease outbreak investigation and environmental control of foodborne pathogens.

Glossary

Whole Genome Sequencing (WGS): The process of using a modern DNA sequencing platform (such as an Illumina or PacBio sequencer) with the goal of sequencing the majority of an organism's genome.

Pulsed-Field Gel Electrophoresis (PFGE): A molecular subtyping technique for bacteria involving the comparison of the DNA fragments that result when the organism's DNA is digested (i.e., cut) at specific sequences. Fragments are separated by size and compared to a reference database of variation.

Basepair (bp): The basic information unit in a genome sequence/DNA molecule; **see nucleotide.**

GenomeTrakr: A public bacterial genome sequence database initiated by the FDA, and hosted by the NCBI, that stores the data for foodborne pathogens sequenced by U.S. and international food safety efforts.

Short-read sequencing technologies. DNA sequencing technologies that produce sequence reads, currently up to 500 bp. These include Illumina and IonTorrent sequencers.

Long-read sequencing technologies. DNA sequencing technologies that produce reads longer than 1000 bp and often lengths over 70,000 bp. These include Pacific Biosciences and Oxford Nanopore sequencers.

Single Nucleotide Polymorphism (SNP): A difference between DNA sequences in the identity of a single nucleotide (an A, T, G, or C). For example, two sequences of AATAA and AAGAA differ in a single SNP in the 3rd position.

Multi Locus Sequence Typing (MLST): A molecular subtyping technique for bacteria involving the sequencing of usually 7 slowly evolving genes and comparing those sequences to a reference database of variation ('allelic types') for each gene. The allelic types for all genes are then combined to result in a single MLST subtype.

Bioinformatics: The field of computer science used to analyze complex biological data, such as WGS data, and translate that into interpretable information.

Read quality: A parameter related to the raw data of DNA sequencing that reports how likely an error existed in the determination of the individual base.

Fold coverage or **sequence depth:** A parameter related to the raw data of DNA sequencing that reports how many times an individual base in a genome has been sequenced. For example, 100 fold average coverage means that each individual base in the genome has been sequenced 100 individual times, on average.

Putative contaminants: In the context of WGS, this refers to the portion of the raw data of DNA sequencing that may have come from sequencing DNA contaminating the sample.

Single nucleotide variant (**SNV**): synonym for SNP.

Insertions and **deletions:** In the context of a WGS, this refers to the presence or absence of stretches of DNA that can range from a single nucleotide to regions encompassing multiple genes. Commonly referred to as **indels**.

Genomic rearrangements: Genomic events where a large portion of a DNA sequence has been moved to another position in the genome or has been inverted.

Genome assembly: The process of inferring an organism's true genome from the raw data of DNA sequencing.

***de novo*:** In the context of WGS, this refers to bioinformatics analyses that consider only the data generated by a sequencing procedure, without comparing that data to a reference genome.

Whole genome alignment: A bioinformatic method that compares the genomes of two or more strains in an attempt to find genomic regions that are shared or differ among the strains.

Draft genome sequences: Genome sequences that still contain gaps.

Closed genome sequences: Genome sequences that do not contain gaps. These attempt to represent the full genome, in its proper order.

Reference mapping: Reads are aligned (mapped) against a (preferably closed) reference genome.

Phylogenetics: The study of evolutionary relatedness of organisms. Relatedness of bacterial strains is commonly displayed in the form of a **phylogenetic tree.**

Pipeline: In bioinformatics, this is a standardized set of analysis programs run on standardized input data.

Tables

Table 1. Key features of U.S. government and academic WGS analysis pipelines.

Pipeline Source[1]	NCBI Pathogen Detection Pipeline (FDA, FSIS, CDC use)	FDA CFSAN	USDA FSIS	Academic Research
Used for	Food outbreak and trace back investigation	Food safety trackback among FDA regulated foods	Food safety trackback in the meat industry	Understanding foodborne pathogen ecology, transmission, and control. Developing improved bioinformatics
Interpreted output	Clusters of 2 or more closely related isolates	Can use matrix for pairwise distance and phylogenetics	Not publicly available	Phylogenetic trees, statistically supported biological claims
Basic output	Phylogenetic trees from high-quality SNPs: all SNPs called against chosen reference	SNP matrix: from SNPs common to all strains aligned to a chosen reference	Not publicly available	Varys. Usually high-quality SNPs + phenotypic information
Basic approach	Partition strains, call SNPs w.r.t. reference, phylogenetic analysis	Partition strains, call SNPs w.r.t. reference, phylogenetic analysis	Not publicly available	Highly varied.
Metadata availability	Limited to NCBI biosample files, e.g. year, state, human/food/environment, serotype	NCBI biosamples	NCBI received: source type, year, state, subtyping	As supplemental materials or by request
Key Software	Custom pipeline, unpublished but documented. NCBI Genome Workbench viewer	Custom pipeline, published. Uses open-source tools	BioNumerics, CLC Genomics, and Geneious commercial software	Vary. Usually open-source tools + custom Unix or Python scripts
Results available	Yes, by FTP. Updated when new strains are submitted to NCBI.	No. Used internally	No. Used internally	Published paper often links to results or raw data + code
Pipeline available	No	Yes, GitHub (command line interface)	No	Varies. If so, from GitHub, author websites, or request

[1]NCBI Pathogen Detection Pipeline *(10)*. FDA CFSAN *(8)*. Academic from Holt Lab Blog, https://holtlab.net/2016/01/17/microbial-genomics-methods/, and *(1)*.

Table 2. Species specific features relevant to WGS analysis for food safety.

Foodborne Pathogen	***Listeria monocytogenes***	***Salmonella***	***Campylobacter***	***STEC / Shigella***
Median genome size (kb)	3.0 mb	4.9 mb	1.7mb	5.6 mb
Strains in GenomeTrakr (Q1,2016; *[4]*)	8,000+	35,000+	1,800+	10,000+
Routine WGS in public health	FDA/CDC Since 2013. FSIS since 2014	FDA/CDC Since 2013. FSIS since 2014	FSIS since 2015	FSIS (STEC) since Dec. 2014
Relative ease of WGS analysis	Easier	Medium	Medium	Hard
Why?	Highly clonal, meaning limited HGT/ recombination ∴ few SNP differences tend to be informative	Large genome that can contain prophages	Chromosome can contain a large repetitive region	Multiple prophages are integrated into the chromosome that are almost identical
Demonstrated impact	Identifying more, smaller outbreaks. Linking human cases to routine regulatory inspection isolates. Retrospective cases identification	Improved resolution with hard to sub-type serovars, e.g. Enteritidis	Major examples still to come	Major examples still to come
Key questions	How many SNPs differences are needed to determine how close iso-lates are to one another?			

References

1. Deng, X., H. C. den Bakker, and R. S. Hendriksen. 2016. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu Rev Food Sci Technol*. 7:353-74.
2. Leekitcharoenphon, P., E. M. Nielsen, R. S. Kaas, O. Lund, and F. M. Aarestrup. 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS ONE*. 9:e87991.
3. Stasiewicz, M. J., H. F. Oliver, M. Wiedmann, and H. C. den Bakker. 2015. Whole genome sequencing allows for improved identification of persistent *Listeria monocytogenes* in food associated environments. *Appl Environ Microbiol*. 81:6024-6037.
4. FDA. Date, 2015, GenomeTrakr Network. Available at: http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm. Accessed 5/5/15.
5. CDC. Date, 2016, Multistate Outbreak of Listeriosis Linked to Packaged Salads Produced at Springfield, Ohio Dole Processing Facility (Final Update). Available at: http://www.cdc.gov/listeria/outbreaks/bagged-salads-01-16/index.html. Accessed 6/1/2016.
6. Jackson, K. A., S. Stroika, L. S. Katz, J. Beal, E. Brandt, C. Nadon, A. Reimer, B. Major, A. Conrad, C. Tarr, B. R. Jackson, and R. K. Mody. 2016. Use of Whole Genome Sequencing and Patient Interviews To Link a Case of Sporadic Listeriosis to Consumption of Prepackaged Lettuce. *Journal of Food Protection*. 79:806-809.
7. CDC. Date, 2015, Multistate Outbreak of Listeriosis Linked to Blue Bell Creameries Products (Final Update). Available at: http://www.cdc.gov/listeria/outbreaks/ice-cream-03-15/index.html. Accessed 7/6/2015.
8. Davis, S., J. B. Pettengill, Y. Luo, J. Payne, A. Shpuntoff, H. Rand, and E. Strain. 2015. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Computer Science*. 1:e20.
9. Inouye, M., H. Dashnow, L. Raven, M. B. Schultz, B. J. Pope, T. Tomita, J. Zobel, and K. E. Holt. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*. 6:90.
10. NCBI. Date, 2016, Methods description for ftp://ncbi.nlm.nih.gov/pathogen/; . Available at: ftp://ftp.ncbi.nlm.nih.gov/pathogen/Methods.txt. Accessed 6/1/2016.

Writers:

Leads:

Matt Stasiewicz, mstasie@illinois.edu

Henk den Bakker, Henk.C.den-bakker@ttu.edu

Others:

Jim Bono, jim.bono@ars.usda.gov

Martin Wiedmann, mw16@cornell.edu

Reviewers:

Zaid Abdo, Zaid.Abdo@colostate.edu

Steve Ricke, sricke@uark.edu

Funded in part by the Beef Checkoff.