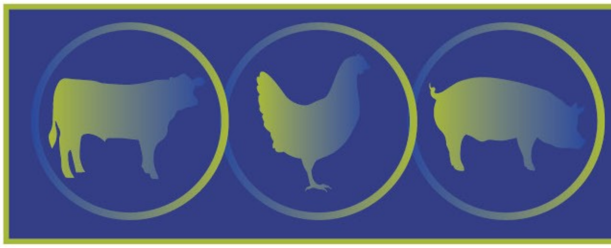

Whole Genome Sequencing (WGS) 101

Understanding the genetics of bacteria has been a priority of the industry for many years. Genetic typing reveals whether bacteria have the genes for certain traits, such as antibiotic resistance, or have the genes that make them pathogenic, i.e., capable of making people sick. Genetic typing also can show how closely related bacteria are to one another, which can help determine the source of pathogenic bacteria. Genetic typing has commonly been done using pulsed-field gel electrophoresis (PFGE), a method that uses enzymes to break the bacterial DNA into pieces of different lengths. The pieces of DNA are then separated on a gel and the different lengths form different bands on the gel. PFGE banding patterns tell us how similar bacteria are to one another. If two bacteria have the same banding pattern, they are considered identical and may have the same source. But PFGE testing doesn't provide as much information as the whole DNA genome sequence. Thus, as the ability to sequence the whole genome has become more cost effective over the last decade, researchers and public health agencies are beginning to use whole genome sequencing (WGS) technology for genetic typing of bacteria, including pathogens relevant to food safety. A commonly used definition for WGS is the process of using a modern DNA sequencing platform with the goal of sequencing the majority of an organism's genome.

In contrast to humans and other organisms that have multiple linear chromosomes, the genome of most bacteria form a single circular chromosome. Four possible DNA molecules, called nucleotides or bases (A, G, T, C), are strung together to code for genes. As an example, an *E. coli* chromosome is about 4.7 million bases which code for about 4,300 genes. To put this in perspective, human genomes have more than 3 billion bases and 20,000 genes. All genomes also contain regions of "non-coding" DNA because they serve functions other than describing the sequence of a gene. Bacteria also contain plasmids, which can be thought of as mini-chromosones, that can move from one bacterium to another and often carry genes coded for antibiotic resistance and virulence. Together the chromosome and plasmids, if any, constitute a bacterial genome. To sequence the whole genome means to identify all the DNA bases from the chromosome and plasmids in the correct order, and then, by comparing to a previously sequenced genome, to identify what genes are present and what they code for that give the bacteria specific characteristics.

Whole genome sequencing is currently done with two main technologies. Short-read technology gives continuous bases of sequence up to 300 bases. Long-read technology gives thousands of bases of sequence. Short-read technology is popular because it's faster and lower cost. Long-read technology is more expensive, so it is not used for routine WGS-based typing. However, long-read technology can recover the complete genome and in the correct order. Short-read technology provides a majority of the genome; however, when put together, regions are missing or can't be resolved, so gaps are introduced that split the genome into smaller pieces. Once the initial sequence is obtained, data analysis and quality control software help obtain the final sequence.

The main objective of WGS analysis is the identification of genomic differences between bacterial strains. The main difference of interest is when there is a single nucleotide (A, G, T or C) change between two genomes when compared at the same location. These are called single nucleotide polymorphisms (SNPs). The number of SNPs indicates how closely related two bacterial strains are. The number of SNPs required before two strains are considered "different" is still under debate because the use of WGS for the purpose of genetic typing is still evolving as more is learned about its advantages and limitations. Data analysis programs that consider the number of SNPs and their location in the genome are used to draw relatedness trees that provide a visual of how closely related different strains are. These trees are usually referred to as phylogenies.

While WGS-based methods should allow for significant improvement in foodborne disease outbreak detection and source traceback compared to PFGE, strong epidemiological data (patient interviews linking illnesses with consumption of a certain food) are still essential to conclusively identify the source of a foodborne disease outbreak. Rigorous comparisons have yet to be conducted to assess how often strains with identical WGS can be found in widely separate locations and products that are unrelated, which emphasizes the continued need for epidemiological data in outbreak investigations. Further research on how to decide if two strains are the same is a critical need, especially in food-associated environments, in order to facilitate improved interpretation of WGS data in the context of foodborne disease outbreak investigations.

Writers:

Tommy Wheeler, tommy.wheeler@ars.usda.gov

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

Martin Wiedmann, martin.wiedmann@cornell.edu

Reviewers:

Zaid Abdo, Zaid.Abdo@colostate.edu

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

BEEF ® Funded in part by the Beef Checkoff.

BEEF INDUSTRY BIFSCo FOOD SAFETY COUNCIL MEMBERSHIP

pork checkoff®

FOUNDATION FOR MEAT & POULTRY RESEARCH & EDUCATION

USPOULTRY® U.S. POULTRY & EGG ASSOCIATION