WGS and GenomeTrakr Q&A

This question and answer document is intended to provide additional information which may be helpful in understanding FDA's whole genome sequencing program and GenomeTrakr. It is meant to complement FDA's other whole genome sequencing and GenomeTrakr web pages, with minimal redundancy of the information presented on the other pages, so please be sure to visit all of FDA's pages on this topic.

This is a working document and FDA may update it periodically. If you have questions that are not addressed in this document or suggestions for additional questions to include, please contact FDA at FoodWGS@fda.hhs.gov.

Sections:

WGS Background Information, Including Terminology

1. What is a pathogen?

A pathogen is any organism that can cause a disease. The term is frequently used to describe harmful bacteria, viruses, or fungi.

2. What is a foodborne pathogen?

A foodborne pathogen is a pathogen that is transmitted through food. Illnesses resulting consumption of a food contaminated with pathogens are foodborne illnesses. Common foodborne pathogens include *Salmonella, Listeria monocytogenes, E. coli*/*Shigella*, *Campylobacter*, and *Vibrio parahaemolyticus*.

3. What is a pathogen subtype?

Within every species of pathogen, there are subgroups of the pathogen to which any single organism of the species may be categorized, based on similarity in genetic composition. These subgroups are known as *subtypes* of the species of pathogen. The subgroup to which a pathogen belongs is its subtype.

4. What is pathogen subtyping?

Pathogen *subtyping* is the process of identifying the subgroup to which a pathogenic organism belongs, based on its genetic composition. Subtyping is important in food safety because it allows us to identify which strain of an organism is responsible for an illness or outbreak of illnesses. This, in turn, can help public health officials and members of industry determine the root cause (the source and precise set of conditions) that led to a food contamination event.

5. What is an isolate?

An *isolate* refers to a microorganism (in this case a pathogen) that has been separated from a natural and potentially mixed population of living microbes. Hence it has been "isolated" from the environment and microbial population from which it was originally collected.

6. What are the differences between "food," "environmental," and "clinical" pathogen isolates?

These labels refer to where the pathogen samples were collected. *Food isolates* were collected from food samples; *environmental isolates* were collected from the environment

in which a food was grown, processed, packed, or held; and *clinical isolates* were collected from people who became ill.

7. What is a gene? What is a base pair? What is a genome?

A *gene* is a segment of DNA that provides a blueprint for a cell to make a specific protein that performs a specific task. For example, a bacterium may have a gene that produces a protein that allows the bacterium to grow in the presence of a certain antibiotic.

DNA contains four types of chemical bases; adenosine (A), cytosine (C), guanine (G), and thymidine (T). Adenine is always paired with thymine and guanine is always paired with cytosine. Each pairing is called a *base pair*. The number of base pairs and order of base pairs in a gene determine the type of protein the gene will build.

A *genome* is the complete set of DNA for an organism, including all of its genes. It contains all the information required to build and maintain the organism. A *Salmonella* genome contains about 4,500 genes and has approximately 4.8 million base pairs. (In comparison, a human genome contains roughly 20,000 genes and 3 billion base pairs.)

8. What is whole genome sequencing (WGS)?

Whole genome sequencing (WGS) is the process of determining the order of the four chemical bases (A, G, T, and C) for the entire genome of an organism.

9. What is the advantage of using whole genome sequencing to subtype pathogens?

Some subtypes are easily differentiated from each other using a variety of laboratory methods, but other subtypes are so similar that traditional laboratory methods cannot differentiate them. For example, pulsed-field gel electrophoresis (PFGE), which for years was considered the gold standard method for pathogen subtyping, is unable to differentiate some strains of *Salmonella* spp. In contrast, whole genome sequencing has the power to differentiate all subtypes of pathogens, no matter how similar they are. Its ability to differentiate between even closely related organisms allows outbreaks to be detected with fewer clinical cases and provides the opportunity to stop outbreaks sooner and avoid additional illnesses.

10. How do you sequence the entire genome of a foodborne pathogen?

Currently no method exists that can identify the sequence of all the base pairs for the genome of a pathogen in a single continuous read of the base pairs. To get around this limitation, researchers break a pathogen's DNA into fragments (typically 250-300 base pairs long), sequence the fragments, and then use mathematical algorithms to assemble the fragment sequences. The assembled fragment sequences reveal the pathogen's genomic sequence.

11. How can we be sure the DNA fragment sequences are assembled in the correct order?

Before an isolate is sequenced it is allowed to "grow out" or "multiply" in a process known as *culturing*. That way, when it comes time to sequence the isolate, DNA from multiple cells of the pathogen are available to be sequenced. The DNA from multiple cells is then broken apart at different places and the fragments are sequenced. Next, the sequences from the fragments are compared to find overlapping runs of the base pairs. The overlapping areas are then aligned. By identifying and lining up the overlapping areas we can determine the correct order of the sequence fragments. (If we were just lining them up end to end, without overlapping identical areas, we would never be sure they were in the correct order). Still, there might be some sequence fragments that could potentially fit in more than one place along the genome. The parts of the genomic sequence where more than one fragment could possibly fit are areas of uncertainty. These areas of uncertainty are referred to as "gaps" in the sequence. In other cases, "gaps" could result from segments of DNA not receiving sufficient "coverage" during sequencing. If there is insufficient coverage, there is less certainty about the chemical bases in that section of DNA. Some gaps are to be expected and are a result of today's technology limitations.

12. What is a "draft" genomic sequence?

Assembled genomic sequences that contain gaps (areas of uncertainty) are called ***draft*** genomic sequences. By definition, all draft genomic sequences contain gaps.

13. Can sequence gaps be closed?

Yes. Sequence gaps can be closed with additional sequencing runs, but this takes extra time and increases the cost. In virtually all instances, the information provided in a

sequence that contains gaps is detailed enough to be relied upon to inform public health decisions.

14. What is a "closed" genomic sequence?

A *closed* genomic sequence is an assembled sequence that contains no gaps.

15. What is "coverage" and what level of coverage is expected?

Coverage refers to the average number of times (x) a base pair, in its given position in the genome, is sequenced during analysis. Strains sequenced for uploading to GenomeTrakr are currently recommended to have at least 30x coverage. This means that each base pair position, on average, is sequenced 30 times. At 30x coverage there are lots of points of overlap in the fragment sequences, which allows us to have high confidence that the base pairs were correctly read and that the fragment sequences are being assembled in the correct order.

16. What do differences in genomic sequences mean?

The number of differences in the sequence of base pairs, between two genomic sequences, is an indicator of how closely related two pathogen strains are genetically. The fewer differences, the more closely related the pathogen strains are. This information can be used to build a phylogenetic tree.

17. What is a phylogenetic tree?

A *phylogenetic tree* is a diagram that shows the evolutionary relationships among a group of organisms based on their genetic characteristics. Somewhat similar to a family tree, which shows how individuals are related to their siblings, parents, grandparents, aunts and uncles, cousins, etc., a phylogenetic tree can be built for a pathogen. The arrangement of the tree's limbs, branches, and twigs indicate the evolutionary relatedness of different pathogen strains.

The National Center for Biotechnology Information (NCBI) uses computer algorithms to analyze the genomic sequences that have been uploaded to the GenomeTrakr database

and to generate a phylogenetic tree for each species of foodborne pathogen. The more closely strains are related to each other, the closer they will be on the phylogenetic tree and the more likely they are to share a recent common ancestor. Each time a new strain is identified, a new branch or twig is added to the phylogenetic tree. As more strains are identified, the appearance of the tree will change, just as a person's family tree would change if a previously unknown relative is found.

Knowing how closely pathogen strains from different isolates are related to each other can help investigators define the scope of an outbreak and identify likely sources and routes of contamination tied to an outbreak.

18. What constitutes a "match" between pathogen isolates?

If the genomic sequences from two pathogen isolates share a recent common ancestor on the phylogenetic tree **and** differ by only a small number of base pairs, those two isolates are considered to be genetically similar enough to be a "match." What constitutes a small number of base pair differences may vary from one pathogen species to another, but as a general rule, two isolates whose sequences have fewer than 20 differences out of 3-5 million base pairs are considered a match.

19. Why is identifying "matches" between pathogen isolates important?

Matches reveal which isolates have the highest degrees of genetic similarity. Because we know the person, product, and location from which each isolate was collected, we can: determine how many reported illnesses were caused by the same strain of the pathogen; plot the locations of the illnesses caused by that strain of the pathogen, and; know which foods, food production areas, and food holding areas have been shown to harbor that strain of the pathogen. These pieces of information are powerful tools which can help identify outbreaks and determine the size of an outbreak, how a pathogen got into the food supply, and how it moved through the food supply.

20. If a draft genomic sequence contains gaps, how can it be relied upon to determine a pathogen isolate's position in a phylogenetic tree or to make a comparison to the genomic sequence from another pathogen isolate?

Typically, about 1-3% of the genomic information in a draft sequence is considered to be uncertain. Those portions of the sequence are considered to be gaps in the sequence. This means that 97-99% of the genomic sequence, which is several million base pairs, is available for determining a pathogen isolate's position in a phylogenetic tree. (The areas of uncertainty in draft genomic sequences are not utilized for comparative analyses.) Comparing such a large number base pairs provides a high degree of certainty when determining where a pathogen isolate fits in a phylogenetic tree or how it compares to another pathogen isolate.

More specifically, there is exponentially more information being compared using sequence data from WGS than using banding patterns from Pulsed-field Gel Electrophoresis (PFGE), the previous "gold standard" for pathogen subtyping.

For example, a *Salmonella* genome contains about 4.8 million base pairs. This means that a *Salmonella* draft genome would have 4.65 – 4.75 million base pairs (97-99% of the genomic information) available for comparison. In contrast, a typical 2 enzyme PFGE analysis of a *Salmonella* isolate surveys approximately 9,600 base pairs to generate 1,600 bands, of which the 7-15 most defined are used to test for matches against other isolates. It is the increased number of pieces of information (4.65 – 4.75 million base pairs) that whole genome sequencing makes available for comparison that allows it to differentiate organisms with a precision that other technologies do not allow. Thus, even though a draft genome has areas of uncertainty, the volume of data that is available for comparison is so great that there is a very high degree of certainty when determining genomic relatedness.

In instances, when a closed genome is required for analyses, an isolate can always be re-sequenced using different WGS instrumentation. Because closing a genomic sequence is more costly and takes longer than assembling a draft genomic sequence, and because draft genomic sequences already provide a very high degree of certainty, genomes are usually closed on an as needed basis.

21. How rapidly can a pathogen's genome change over time through normal cell division?

The number of changes to a pathogen's genome through cell division, over a period of time, can be influenced by a number of factors, including the growth conditions in the pathogen's environment. When a pathogen cell replicates, there may be no changes in base pairs or there may be a very small number of changes. There have been a number of instances in which pathogen isolates from the same outbreak, collected two months apart, had only five or six differences in their base pairs out of 3-5 million base pair

positions. However, there are also instances when changes have been observed over very short periods of time. For example, FDA performed a controlled growth and sequencing study, in which up to two changes in base pairs were observed during normal replication of isolated cells left overnight in a growing medium in the laboratory. Similar results have been reported in studies by other scientific bodies.

In an environment with ideal conditions for pathogen growth, cells will replicate more frequently than in an environment with less ideal growth conditions. Since each replication is a moment when base pair changes could occur, environments where replications are more frequent will provide greater opportunity for changes in base pairs. However, it is unlikely that such changes would significantly alter that isolate's placement in a phylogenetic tree.

22. Could two strains of a pathogen each develop random genetic mutations or other genetic changes, which result in the two pathogens converging on a common genomic sequence and appearing to be a match, when in fact they are from different recent lineages and not a true match?

It is extremely unlikely that such an event would occur. While many strains of *Salmonella*, *E. coli*, or *Listeria monocytogenes* can be closely related within their species, these relationships are not random nor are they due to random or inexplicable genetic events. When a pathogen cell divides into two cells, each daughter cell can accrue unique genetic differences that immediately distinguish the cells. These genetic differences are changes in individual base pairs along the pathogen's genomic sequence. The more cell divisions that each of those two daughter cells and their own daughter cells go on to have, the more likely that differences can be found between the original daughter cells and the most recently divided descendants. Over many generations, enough differences could accrue that the cells descended from those original daughter cells might be very distinct from each other. However, a phylogenetic tree would still be able to map out how the different descendants were related to one another.

Strains that do not share a recent common lineage will have many differences in base pairs and those differences are likely to appear at many different points along the genomic sequence. When a genetic mutation or random genetic event takes place, it only affects a small section of DNA. So even if a mutation resulted in convergence in one part of the genomic sequence, differences in other parts of the genomic sequence would remain. This means the evolutionary history of the isolates could still be determined.

23. What kinds of data are stored in the GenomeTrakr database?

GenomeTrakr contains the genomic sequences for thousands of pathogen isolates, along with the corresponding metadata for each isolate sequence.

24. What is metadata?

To understand how pathogens are related to each other, how they move through the environment, and how they can get into the food supply, researchers need to know more than just the genomic sequences of the pathogens. Researchers need to know where the pathogens were collected; when they were collected; whether they were from food, environmental, or clinical isolates; and a variety of other details, such as the food or foods in which the pathogens were found, whether any people or animals became ill from eating that food, and who collected the pathogens. These contextual details are called *metadata* and are uploaded into the GenomeTrakr database along with each genomic sequence.

Metadata are essential pieces of the puzzle in foodborne illness outbreak investigations. They can be used to determine whether a particular geographic area has had multiple reports of illness associated with the same pathogen subtype, or if a pathogen subtype previously associated with one geographic region is now appearing in multiple geographic regions. These can be clues as to how a foodborne pathogen moves through the food supply chain.

25. Where do the genomic sequences and the accompanying metadata in the GenomeTrakr database come from?

The genomic sequences and information that accompanies each sequence come from public health laboratories, university laboratories, hospital laboratories, industry laboratories, and independent laboratories. These laboratories are located in the U.S. and around the world. A list containing the names and locations of many of the contributing laboratories can be found on FDA's GenomeTrakr Network web page.

Using WGS and GenomeTrakr

26. Do I have to register to view and use GenomeTrakr data?

No registration is required. GenomeTrakr data are publicly accessible and can be viewed by anyone with web access. Genomic sequences in the database can be compared using web-based software packages. Updated phylogenetic trees are produced daily, so you can see the closest genetic relative for each pathogen strain in the database.

27. How frequently does FDA use GenomeTrakr data?

FDA uses GenomeTrakr data on a daily basis. Virtually every day, the genomic sequences from new isolates are uploaded to the database by any number of sequencing labs around the world. As the sequences are added, FDA checks to see if they "match" or are similar to sequences already found in the database. Certain matches may have regulatory implications which could prompt FDA to gather additional information. For example, a match between a sequence from a clinical isolate and a sequence from an isolate collected from a food facility might prompt FDA investigators to go back to the food facility to assess the processing conditions and to conduct additional testing. Matches that do not have immediate regulatory implications still provide information that helps researchers, public health officials, and members of industry better understand where the pathogens have been found and how they move through the environment. This information can be highly informative in designing food safety systems.

To read more about how FDA uses WGS for regulatory purposes and how public health officials and others can use the technology for proactive applications, please visit:

- How FDA Uses Whole Genome Sequencing for Regulatory Purposes
- Proactive Applications of Whole Genome Sequencing Technology

28. How can GenomeTrakr data help scientists develop better and faster diagnostic tools to detect bacterial contamination?

There are two ways that GenomeTrakr data can help scientists develop faster and better diagnostic tools to detect bacterial contamination. First, having a database of the thousands of different subtypes and lineages of *Salmonella, E. coli, Listeria*

monocytogenes, and other pathogens allows us to see just how diverse these species of pathogens really are.

Knowing the diversity within each species is important because it makes us more aware of rare or infrequently observed strains of pathogens that might have otherwise gone unnoticed. When developing new detection tools, scientists can ensure the new methods are sensitive enough to detect these rare or infrequently observed strains. More sensitive and specific detection tools will permit better identification and quicker removal of contaminated foods from the food supply.

The second way that GenomeTrakr can aid in the development of future diagnostic tests is through the application of "metagenomics." Metagenomics is the sequencing and study of genetic material recovered directly from the food or environment in which it resides. With metagenomic testing there isn't a need to culture and "grow out" a pathogen isolate before sequencing it. Instead, the DNA from any and all bacteria collected can be sequenced immediately. The time saved by not having to culture a pathogen before sequencing it can increase the speed with which investigations of foodborne outbreaks can be conducted. The GenomeTrakr database is providing the foundation and reference information for the design of metagenomic tests for foods.

29. Are there economic advantages to using the information in GenomeTrakr to help with foodborne illness outbreak response and traceback investigations?

There are a number of economic advantages to using GenomeTrakr to help with pathogen identification, outbreak response, and traceback investigations. Here are a few:

a) Reduced illness costs: The information provided by WGS and GenomeTrakr gives investigators the ability to detect foodborne illness outbreaks sooner and to speed the investigations into the root causes of the outbreaks. This has the potential to reduce the number of illnesses and deaths that might otherwise result from an outbreak. Fewer illnesses and deaths improve the public health and reduce the economic costs from missed days of work, hospital stays, medical appointments, and other treatments necessary to recover from illnesses, etc.

   The larger the GenomeTrakr database grows, the more informative it becomes, and the greater the speed in which links between pathogen isolates will be made.

b) Better resource allocation: The increased precision in pathogen identification and the ability to leverage the metadata in GenomeTrakr to reveal the geographic locations historically associated with genomic matches and closely related strains of the isolate in question can help investigators more quickly identify the strongest leads to pursue in outbreak investigations. This allows FDA and other public health officials to better target where to deploy outbreak investigators. Public health agencies have a finite number of resources and investigators, so being able to quickly and accurately identify the food facilities most likely to have played a role in a foodborne illness outbreak or a food contamination event means: 1) fewer food facilities will likely have to be visited, 2) fewer investigators will be needed, and 3) the root causes of the food contamination event will be determined more quickly. These efficiencies combine to reduce the number of resources an agency must spend, on average, to investigate foodborne illness outbreaks and food contamination events.

c) Reductions in food waste and lost inventory: Being able to more rapidly and precisely identify which foods should be part of a recall reduces the likelihood of safe foods being unnecessarily removed from the market. This reduces both food waste and costs to industry and consumers.

d) Eliminates need for other tests: WGS is more discriminatory and informative than other subtyping methods, and it provides information about virulence, antimicrobial resistance, and a historical reference to pathogen emergence. Traditionally, it required many tests to gather this information. This means that WGS can replace multiple tests, saving both time and money.

30. Can the food industry benefit by incorporating WGS into its in-house testing and checking the results against the data in GenomeTrakr?

Although there is no requirement to use WGS, companies may benefit from proactively incorporating WGS into their food safety and sanitary control plans. By performing WGS on any pathogen(s) isolated during their sampling efforts, companies will have detailed information about the pathogen(s) detected. They can then compare this to the genomic information publicly available in GenomeTrakr. For example, if testing of an incoming ingredient indicates the presence of a pathogen, comparison of the pathogen's genomic sequence to information in GenomeTrakr may reveal information about the root source of the ingredient's contamination in the company's supply chain. It may also identify which preventive or sanitary controls may have failed and need to be corrected.

31. How does GenomeTrakr relate to other large genome sequencing projects, such as the Global Microbial Identifier project?

The Global Microbial Identifier initiative began at a 2011 meeting in Brussels to discuss how WGS activities worldwide could be coordinated to improve public health. GenomeTrakr is one of several international efforts that support that goal. Because data from GenomeTrakr are free and publicly available, researchers, public health officials, and members of industry can access it and use if for many purposes, including creating new tools and methods to improve food safety. For more information on FDA's role in international whole genome sequencing efforts, please visit FDA's International WGS Efforts web page.

32. How long does it take to sequence a pathogen's genome, and how much does it cost?

Using current methods and equipment, it takes approximately one week to sequence a pathogen isolate and to analyze the results. Excluding the initial cost of the sequencer, sequencing costs less than $50 per isolate. This is the cost of the reagents (chemicals) needed for a single run of the sequencer.

Information for Collaborators and Prospective Participants

33. How can I contribute sequences to the GenomeTrakr database?

Any laboratory with a genome sequencer can collect the draft genomic sequence of a pathogen isolate and upload it to the database. Instructions for contributing genomic sequences and their corresponding metadata to the GenomeTrakr database can found on the NCBI's "How to Submit Data for Real-Time Analysis" web page or by contacting FDA at FoodWGS@fda.hhs.gov.

34. What metadata should I include with each sequence?

Each food or environmental isolate should include a unique strain ID, the species binomial and any serotype information if available, the name of the lab that collected the isolate,

the collection date, the collection location (country or country and state if in the U.S.), and the isolation source (ex. generic description of "food" or "environmental"). Additional metadata is useful and welcome, but not required. A description of the metadata that should be included with each uploaded genomic sequence can be found on NCBI's "BioSample Attributes" web page.

35. What quality assurance (QA) and quality control (QC) steps are there for submissions?

The first step to help ensure the quality of the genomic information uploaded to the GenomeTrakr database is meeting the recommended 30x coverage for all sequences. Beyond that NCBI does additional quality assurance and quality control checks on the submissions. For contributors who send their data to FDA for uploading, FDA conducts its own quality assurance and quality control tests on the sequencing data. Once uploaded to the GenomeTrakr database, these submissions will also be checked by NCBI. Hence, they undergo multiple QA and QC checks. The QA and QC checks by NCBI and FDA can detect things like mislabeled or contaminated samples.

36. Do labs that upload their data to GenomeTrakr get feedback on their submissions?

Labs that submit genomic information to FDA for uploading to GenomeTrakr will receive feedback if FDA's quality assurance and quality control checks reveal a problem. The lab can re-sequence the isolate and resubmit the genomic information.

37. Are there meetings for researchers and food safety professionals interested in participating in the GenomeTrakr program?

Yes. FDA has been holding a workshop for GenomeTrakr contributors every other year since 2013. The first workshop was held in Baltimore, MD, the second in Washington, DC, and the third in College Park, MD.

A list of upcoming events at which FDA will either be 1) presenting information on foodborne pathogen whole genome sequencing, its applications, and its positive public health impact, or 2) otherwise participating in discussions related to foodborne pathogen whole genome sequencing, is available on FDA's WGS Events web page. The information presented by FDA may include but is not limited to experiences and observations drawn

from its foodborne pathogen whole genome sequencing research program and the GenomeTrakr network.

May 11, 2018
Version 1