This document is provided as an advanced copy of CPS materials under review and may be updated. You may not further distribute, post, or publish without the permission of the Center for Produce Safety.

Center for Produce Safety STEC Seasonality Project:
Romaine Lettuce Seasonal Risk in the California Central Coast Region
Prepared by Trevor V. Suslow, Professor of Cooperative Extension and Research, Emeritus, University of California, Davis | 2022

CPS Issue Brief 4: Genomics and SNPs
A Primer for Understanding Whole Genome Sequencing Applications

Overview Statement: Whole genome sequencing (WGS) has become a familiar term and acronym across the food and produce supply chain at a level and frequency unprecedented even five years ago. The outcomes of WGS applications within outbreak and environmental investigations, and the pervasive use of WGS in modern research efforts directed at produce safety and farmscape ecological studies, find their way into everyday "coffee-shop" and workshop talk among growers and handlers.

Specific to CPS needs, a desired resource from this STEC Seasonality Project was a somewhat deeper overview of how to understand and appreciate these WGS outcomes, without getting lost in the details of methodology. This Issue Brief was developed largely from the perspective of the efforts to organize and assemble root cause information applicable to evaluating research proposals and outcomes on this topic, applying WGS approaches to filling knowledge gaps and practical solutions to produce safety issues.

Although this Issue Brief is longer than intended, a key concern was that oversimplifying the background would lead to a loss of understanding. We hope it is a relatively easy read and provides the level of insights helpful to appreciation of the topic in relation to the CPS mission.

Purpose: This CPS-internal document provides a lay-technical overview of WGS as it relates, in particular, to STEC Issue Brief 1: Hypothesis Risk Matrix, which lays out an assessment of research opportunities and prioritization challenges. Useful background is presented in relatively plain language, helpful and applicable to reviews of CPS proposals and cited methodology and literature. This brief is intended to increase familiarity with the key approaches and applications of WGS to achieve a better understanding of diverse and everyday communications impacting the produce and allied industry sectors. Some examples include public agency communications on outbreak and environmental investigations, planning for root cause analysis, and other current, emerging, or potential produce industry applications.

The approach here is to largely exclude the often-mind-numbing jargon, bioinformatic methodologies, and computational details (terms **bolded** are explained briefly in sidebars). Although these details are critical to high confidence and comparability of outcomes, they are not essential to formulating necessary questions and recognizing pitfalls in discussions of research proposals. To facilitate readability and clarity, much of this Issue Brief presents bulleted information rather than a detailed narrative laden with supporting citations. Several resources are provided for those interested in reviewing some current journal papers related to this topic. Most are highly specialized technical papers and a heavy read for those without detailed training, but some are readable for diverse food safety professionals. It is hoped that this Issue Brief will assist the reader in gliding over the complex details of methods and WGS analysis to extract the key learnings applicable to STEC Seasonality and other CPS proposed research priorities.

Introduction: Whole genome sequencing (WGS) is firmly established in the everyday chatter within the produce supply chain. WGS has been a truly transformative technology in produce safety, with valued benefits but also challenges to resolve from potential and realized direct and indirect collateral consequences on regional and farm-level land-use decisions. General knowledge and understanding of both the outcomes of WGS applications to produce safety and the terminology associated with methodologies has greatly improved across the CPS Technical Committee over the past three years. A significant level of trust is required in acceptance of results generated by WGS comparative analysis and associated applications, such as **microbiome** and

microbiome – in this use, the combined genomic information for all culturable and non-culturable microorganisms in an environment… may be broken into sub-parts in some uses (e.g., bacteria, fungi, phage, viruses)

metagenomic – in this use, a term broadly describing the characterization of the multiple and mixed organism microbiomes in an environment

metagenomic analysis, now so pervasive in research proposals. (These applications are not covered in this Issue Brief but planned for a future supplement). In general, public agencies operate under detailed standardized procedures, often referred to as pipelines, but even this varies between agencies in the U.S. and internationally. This is generally much less standardized in academia, public research labs, and across commercial labs. Achieving another level of shared insights on methodology capabilities and limitations is anticipated to foster more critical assessment of experimental design and validity of results during CPS proposal review. At the same time, it is hoped that this brief will provide an additional resource for effective communication as industry explores and considers its own applications of the broad and rapidly evolving techniques with a foundation in WGS platforms.

So how does this fit into the CPS STEC Seasonality Project? Loosely, to be sure. However, WGS-based information linking historical and current produce outbreak clinical cases, clinical cases associated with other foods, and diverse environmental isolates has been the driver for these efforts at resolving questions laid out in the STEC Project Risk Matrix (STEC Issue Brief 1). WGS-based isolate relatedness predictions have been the catalyst for much of the efforts to develop a root cause investigation strategy to identify the sources of apparent seasonality patterns in product contamination and the specific reoccurring sub-group of *E. coli* O157:H7 (EcO157), namely REPEXH02. REPEXH02 is a recognized cluster of WGS subtypes associated with outbreaks attributed to romaine lettuce produced within the California Central Coast region. As described in STEC Issue Brief 1, with so little other essential information regarding production locations, seasonal timing and practices, and distribution specific data available to this project, the data accessible from WGS analysis could only support broad investigative speculation in root-cause hypothesis generation. At the same time, the much more limited WGS profile diversification demonstrated for the main Central Coast regional EcO157 REPEXH02 of concern (for example, as compared to REPEXH01, which includes the subtypes associated with the Romaine outbreak in early 2018) has stimulated several well-founded but unresolved questions. Many of these questions are deeply rooted in seasonality patterns (See Figure 4) and practices across several co-regional ag industries. A number of these questions will ultimately be answered using evolutionary genomics, metagenomics, and a broader and long-term effort to resolve modes of dispersal from unequivocal domesticated animal and wildlife reservoirs, confirm or exclude presumptive sources, and identify unknown environmental sources of this specific and other clinically relevant Shiga toxin-producing *E. coli*.

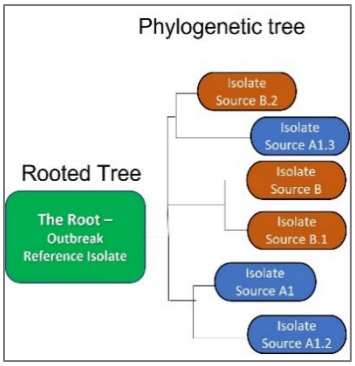
Impact of WGS databases on seasonality root cause research design:

Concerns have been identified for building a family of valid hypothesis-generation tracks, needed to prioritize and fund research projects to address these possibilities, based largely on WGS relatedness mapping of clinical, product, and environmental isolates. These concerns and uncertainties have a foundation in the general lack of standards and inherent opaqueness of multiple algorithms and computational statistics used in this field. Progress has been hampered by widespread reluctance to share sequence and metadata in public repositories, and an instable comparability across suitable bioinformatic tools and workflows, other than those established in federal public health agencies. Challenges to conducting effective analysis, due to uncertainties and biases impacting the validity of isolate comparisons have been identified. This reduced confidence is based on many **draft genomes** uploaded to public databases and the absence of standards for quality assurance and curating functions. Publication of draft genomes have led to divergent conclusions based on isolate relatedness, which may be as limited as 61% genome coverage. The details may be provided, or accessible with sufficient skill and patience, but most individuals only view the various graphic mapping results of these family trees (**phylogenetic** imaging schemes). Attempts to layer many factors associated with seasonal patterns for outbreaks on romaine, and positive detections on diverse leafy greens, have been both aided and confused by apparent WGS relationships and how different groups interpret sameness of isolates in relation to a common source.

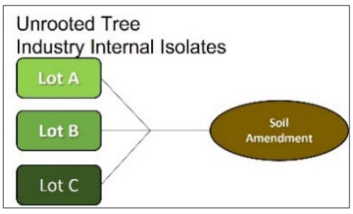
To appreciate how WGS is being applied to resolve this dilemma, it is important to have a lay-technical understanding of the basics of the variation and how isolate relatedness mapping is developed. Not the deep technical details of how this happens, but more attuned to the influences that occur along the way to a seemingly absolute result. Keep in mind that it has been said that phylogenetic trees are as much prone to being a hypothesis as a roadmap to identifying a common food vehicle, specific geographic source, seasonal or regional production or harvest practice, or even alternate animal hosts.

draft genomes – in this use, isolate sequence information that is incomplete, improperly aligned with missing information, and recognized as having low accuracy for phylogenetic placement on a tree

phylogenetic – in this use, a descriptive or graphical vehicle to show genetic relatedness, lineage, and diversification groupings



Root WGS selected as standard reference to suggest ancestry and source relationships



Unrooted approach often used for internal source-tracking and root cause investigation without suggesting matches to ancestry in public databases

Selected WGS basics for reviewers and industry decision-makers: Numerous written narratives and reviews, PPT overviews, and animated video resources are available to obtain a lay or lay-technical appreciation for the foundations in DNA structure and process steps underpinning WGS results. The deeper you go, the more complex and complicated it gets, and quickly. As in any field, it is very easy to become lost and overwhelmed with the acronyms and jargon. Here, we have attempted to greatly simplify a basic understanding without losing too many of the essential steps in following where a proposed activity is going or what is behind an emerging and preliminary or final result communicated.

clonal – in this use, a group of cells or isolates that are essentially identical to a common parental source

SNP (Single Nucleotide Polymorphism) – in this use, a single nucleotide variation (the building blocks of DNA information, ATGC) at a specific base pairing A:T or G:C location. This variation becomes a marker for differences among isolates in phylogenetic relatedness analysis.

wgMLST (whole genome Multilocus Sequence Typing) – in this use, similar to cgMLST but all sequences and sources of potential gene allelle information and variation is retained and included in the analysis

cgMLST (core genome Multilocus Sequence Typing) – a relationship scheme based on a defined and fixed array of genome-wide alleles

- The genome of an organism is a set of instructions from highly conserved to those characterized as hyper-variable and often strongly influenced and subject to modifications by external stress factors.
- An easy way to imagine the basis for WGS is to think of the genome of an isolate (predominantly a population of **clonal** cells from cultures but single-cell genome sequencing is also possible) as a bound Owner's Manual that is largely the same for the type and model but will change over time. The Manual is subject to copying errors and may have small inserts with specialized information or non-essential information that is not necessarily or may be reversibly bound to the main pages. To compare the various Owner's Manuals among isolates, they must first be sent through an enzymatic or mechanical force "shredder," breaking the entire Manual and any loose inserts in small pieces (more later on this) to read every letter in the proper sequence in which they exist. Once the "sequence reads" are available, sophisticated computational algorithms (software) reconstruct and "exactly" align the standard reference Owner's Manual with all others under investigation and compare for mismatches across multiple words or every letter of every word.
- There are various ways to explore the relatedness and degree of differences among the Manuals of interest. Currently, the two most related to public health, food safety, and found within many CPS proposals and reports are high quality **SNP (Single Nucleotide Polymorphism)** analysis and **wgMLST** or **cgMLST (whole genome** and **core genome Multilocus Sequence Typing)**.

- hqSNP protocol establishes a key Owner's Manual as the reference book (the reference genome) to guide the reassembly of all the sequence reads into an aligned order to analyze for differences
- wg and cgMLST assess similarities and differences across a library of Owner's Manuals to reassemble the new outbreak or environmental isolates added to the same database

hqSNP:

- Provides greater resolution in relating isolates by the closeness of genetic lineage to their relationships in an outbreak and combined with epidemiology and traceback data to a potential product, production location, and environmental and/or host source.

SNP-calling – in this use, a computational process of genome sequence alignment and determination of the presence of a SNP in a specific location

- Obtaining quality sequence reads and aligning the entire Owner's Manual is a complex and potentially variable process, entirely dependent on the algorithms used and influenced by the operator's training, skill, and subjective decision-making. There are no current standards or required proficiency tests or certifications in the general research and commercial services industries. Due to the complexity and the popularity of the powerful and immensely information-rich technology outputs, a plethora of user-friendly software tools for sequence alignment and data interpretation have emerged.
- The process of assembling (mapping) the raw sequence reads from all the shredded bits back into one of a group of related Owner's Manuals depends on several factors beyond the detail appropriate here. However, once mapping is completed, the next step is **SNP-calling**, a term which has become a regular feature of communications and research proposals or reports, and is based on a set of user-defined parameters. These may also be proprietary factors of convenience or simplicity for the operator embedded in free-ware or commercial bioinformatic tool sets. The result is to assign quantitative SNP differences as a measure of relatedness, graphically represented as a tree, a circle, color-coded cluster diagrams, heatmaps, of other forms of highly condensed illustration of complex datasets, for all Owner's Manuals. Most in the industry are now at least visually exposed or familiar with phylogenetic trees (See Figure 2).
- A key feature of most research proposals is to describe the coverage and quality of the sequence reading of the shredded pieces. More coverage allows for better accuracy and more confidence in alignments and SNP-calling.
- The alignment and sequence read management pipelines (software, algorithms, process and protocols) are nonequal and arguably increasingly subjective in the hands of under-trained personnel accessing a multitude of "user-friendly" online systems.

- The unbound or bound inserts within an individual or cluster of Owner's Manuals mentioned above may be included or excluded depending on an operator choice or masked and decision-making embedded in the software. This may greatly impact a reported analysis of relatedness.
- A true challenge for the industry is a desire or interest to tap into the power of SNP analysis, but not use a reference Owner's Manual (reference genome) that would link isolates obtained and characterized from their product or facility to an outbreak. In this case, the WGS process may be conducted against a selected internal reference among many environmental isolates. This may be sufficiently informative for source-tracking to an input (e.g., soil amendment), farm, product lot, or environmental swab location(s) but is prone to sequence alignment limitations during mapping and analysis uncertainties.
- Advantages of hqSNP:
    - Provides greatest degree or resolution on relatedness (however requires a closely related and fully characterized reference genome and improves as the database of representative isolates increases).
    - Comparisons of new isolates to a highly related reference isolate can produce a phylogenetic tree that reflects the emerging epidemiology very closely, and may help guide traceback in the case of an emerging outbreak or source-tracking for root cause analysis.
    - Reflects evolutionary origins and diversification among a lineage of isolates. May be suggestive of environmental or host-associated sites of EcO157 amplification (growth), which would tend to drive SNP differences.
    - From a research perspective, understanding the drivers and rates of SNP diversification will be an important tool in understanding environmental sources of reoccurring EcO157 subtypes.
    - SNP position within a genome (the specific page-paragraph-sentence-word in the Manual) can be determined, but is rarely done due to resource limitations. A classic example for delving into SNP position to postulate evolutionary lineages, rates of diversification, and sources of acquiring virulence and hypervirulence traits for *E. coli* O157:H7 is found in Manning et al. 2008.

- Challenges to strict acceptance of a SNP-based source tracking and attribution:
  - Different sequencing methods, different algorithms, different operator-imposed or software-assigned filtering of sequence data, and reference genome as the root of the tree will create different relatedness outcomes.
  - The differences in the phylogenetic tree may be relatively minor but can be highly charged in speculative attribution to source.
  - Research studies using different reference WGS or **unrooted trees** for industry internal comparisons (to avoid connection to a clinical case or outbreak) are likely to report unsubstantiated relationships.

unrooted trees – in this use, a phylogenetic tree not developed from a standard reference isolate but developed across isolates, generally in a more limited collection, not rooted to a past clinical or recent outbreak sequence

allele(s) – one (or multiple variants) type of gene variant that is differentiated by a mutation(s) which can be localized at the same place on a segment of genetic sequence information (DNA)

locus / loci – in this use, a definable location on a properly aligned and oriented cellular genome sequence

nomenclature – a standardized scheme defining the body of rules for naming (e.g., a sequence type and all hierarchical/descendant clones)

<u>wgMLST and cgMLST:</u>

- MLST schemes are useful for a more rapid comparison of Owner's Manuals by focusing on key characterized sentences (genes) in regions among the pages of hundreds to thousands of related Manuals in a large reference database built over time.
- wgMLST and cgMLST schemes are built from a recognized foundational Owner's Manual and a large set of highly related and characterized Manuals (multiple isolate genomes) to define a collection of **alleles** around a set of standard **loci**.
- In general, wgMLST gives better resolution of comparability between and among isolates, but cgMLST is typically very similar and less complex to arrive at predictive groupings, especially quickly separating clearly genetically distant isolates.
- cgMLST is often used as a pre-clustering tool in a scheme to prioritize subsequent WGS and SNP analysis in an emerging outbreak.
- wgMLST and cgMLST analysis of similarities and differences can be assigned a unique naming type, not possible with SNP analysis. The stability of the informatically assigned allele-classification system allows for a consistent identity system (**nomenclature**) for communication among groups analyzing isolates. This requires a well-curated and managed database with standards for quality and confidence in the uploaded information. Within the CDC system, this is a different and higher standard for database management than is typical for the widely used

- NCBI database uses an automated system to assign and organize uploaded WGS sequences into a phylogenetic tree.
- Comparisons of gene alleles can be used to group isolates into a **clonal complex** (common recent ancestor), a strain grouping, or highly related clones. (Remember that EcO157, like other bacteria, multiply by simple binary fission; one cell duplicates its DNA, enlarges, and splits in two with each new cloned cell receiving one copy of the genomic DNA). DNA copying errors during growth may occur spontaneously at a low rate (1 in 100,000 to 100,000,000 events, depending on the strain and external stressors), and allelic and SNP differences from a close ancestor will accumulate over time. It has been frequently reported, too, that SNP differences may develop during the isolate purification subculturing that must occur prior to WGS processing (See Figure 1 and 3).

clonal complex – in this use, a grouping of isolates that are highly related to a common parental source

NCBI BLAST – Basic Local Alignment Search Tool hosted at NCBI (National Center for Biotechnology Information) which provides many tools for diverse sequence alignment needs

- The presence or absence of mutational alterations of an allele at genomic loci or SNP differences among the clonal cluster identified as most responsible for serious and recurring multistate outbreaks hold many unresolved clues for intelligently addressing the STEC Seasonality dilemma.
- Here again, the process of allele-calling requires skill and process standards to combine both differences and similarities into a relationship (phylogenetic) tree. Judgements are made on what information is to be discarded as outliers, noise, or defective reads and what is to be included. Cross-pipeline (allele finder algorithm) comparisons are prone to misinterpretations. CDC has developed a very robust and standardized pipeline for these analyses.
- cgMLST is generally a faster route to ruling in or ruling out isolate relatedness based on core genes common to a species subtype (*E. coli* O157:H7), while wgMLST provides greater resolution of fine degrees of diversification but may also result in challenges to cross-institution comparisons due to inclusion of the unbound or mobile genetic elements among isolates.
- In many research proposals, the allele calling is based on an **NCBI BLAST** algorithm. Details are not important here, but it is worth noting that a multitude of platforms are used and may give slightly different, but critically important, results regarding any location, product, or environmental attribution.

Sequencing provides the raw data for WGS alignment and analysis:

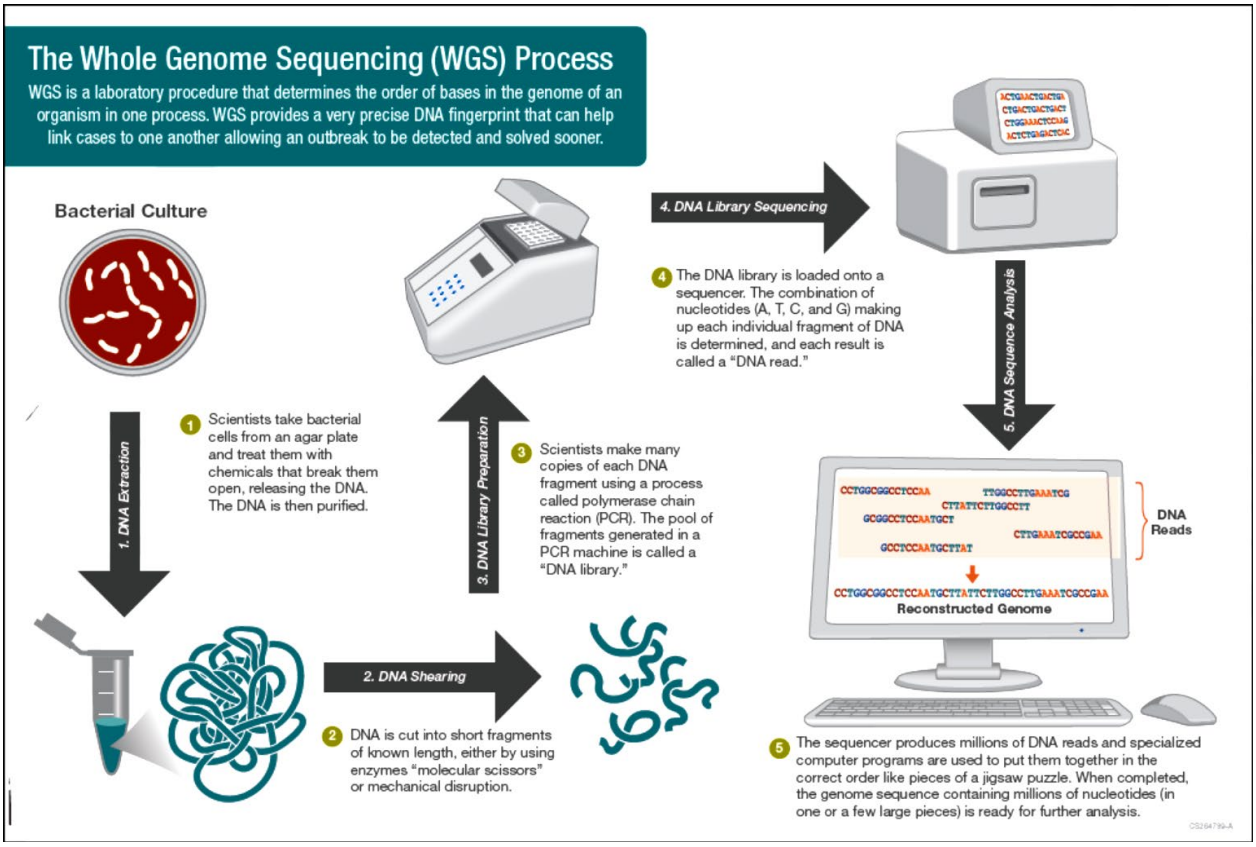


Figure 1. A simplified graphic illustration and short narrative of the basic steps for sequencing is available from the CDC. Acknowledgement to Steven Stroika, Enteric Disease Laboratory Branch/DFWED/NCEZID/CDC.

The CDC also provides open access to their standardized genome sequencing protocols to promote and ensure cross-lab comparability and a uniform naming system (nomenclature). These details are not essential to understand the outcomes, but a brief overview of the sequencing choices may be helpful in reviewing research proposals and reports. One can get quickly lost in the abundance and complexity of comparative studies and most choices are economic and institutional availability.

Currently, the key sequencing choices for microbes relevant to current CPS proposals include the following:

- Illumina – short reads, low error rates
    - MiniSeq – fastest turn-around (10-24h); smaller data output
    - MiSeq – fast turn-around (generally 24-48h) and somewhat lower cost than NextSeq. Speed is generally not a priority in research studies but is often chosen to increase numbers of samples run at a sacrifice of detail. Common choice for 16S microbial community and metagenomic studies
    - NextSeq (largely replaces HiSeq referred to in several CPS proposals) – 10X more short sequence reads yield greater detail. Cost difference not as great a factor. Necessary for more complex shogun metagenomics, increasingly noted in proposals.
- PacBio – tends to be the standard for more complex, detail requiring applications, long reads, higher error rate
- Oxford Nanopore – lower cost of system; comparable to PacBio third-generation capabilities, long reads, higher error rate

Quality proposals discuss their approach to **bootstrap analysis**: In many research reports and proposals, the term **bootstrapping** is tossed in but not always explained or quantitatively defined. Bootstrapping is an essential process in building a WGS relationship and constructing the most accurate phylogenetic tree possible. Keeping it simple and within the context here, bootstrapping is a statistical method for randomly resampling a data set, generally 1,000 times, to develop a level of confidence in assigning an exactness estimate to build the relatedness among isolates on the tree. This is a very important data management step that should be described in every WGS study to characterize isolates. In public health, the FDA CFSAN has an established and published system for bootstrapping criteria, which most researchers follow when working with rooted and unrooted relatedness trees. The number of SNP differences among individual isolates within a branch and between seemingly highly related branches may be misleading without understanding the range and median distances between these branches. Isolates on two close branches may be only 0-2 SNPs different between clustered strains but may be 25-35 SNPs different between the two groups. Strong bootstrap analysis (expressed as a fraction of 1.0; >0.85 as a typical standard) gives evidence of exclusion of a group of isolates from others in the database and a reasonable, though not absolute, indication of a common source. This must, of course, be supported by epidemiology and other environmental source-tracking or traceback data.

bootstrapping – a highly repetitive statistical analysis sampling and substitution of a large sample by randomly selected subset samples to test for and predict standard errors and confidence intervals for sequence alignment

Building and reading a phylogenetic tree:

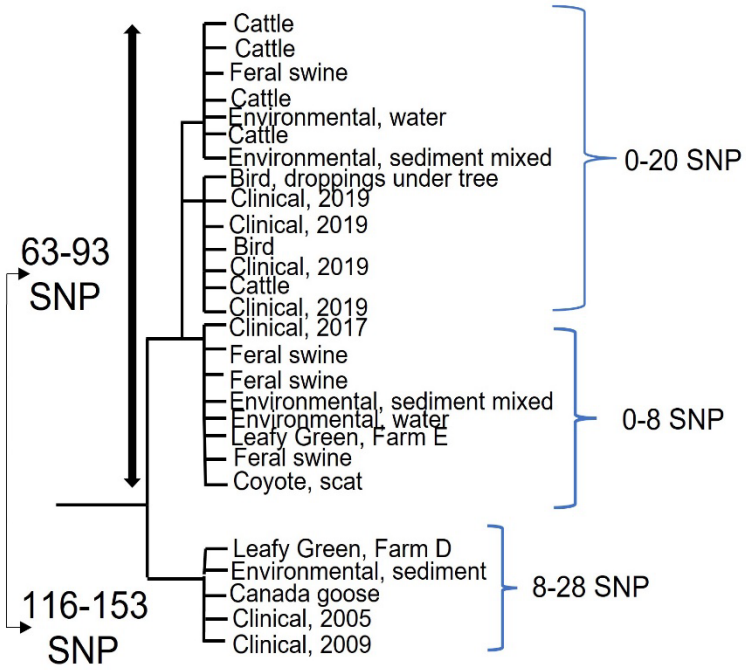


Figure 2. Simplified model of a phylogenetic tree (visual depiction of the evolutionary descent or relatedness of organisms – single bacterial isolates in this example) representing four **clusters** of related isolates of *E. coli* O157:H7. **SNP** analysis displayed in this manner is often interpreted and communicated in very different ways. As a rule of thumb, isolates with 20 or fewer SNP differences are considered highly related. However, this is not a hard and fast rule and, while the focus here is on EcO157, not adhered to in every case with different foodborne pathogens. It is important to understand that the specific position of these SNP changes in the **genome** is rarely characterized. Other evidence has shown that some specific positional SNP changes may be more meaningful than others, such as host adaptation, insights not reflected or extractable on this kind of tree. From a broad genomic (complete set of genetic code in an organism) perspective, the different branches of the phylogenetic tree would be considered highly related. Aligned in this manner for graphic display, isolates from an unspecified "farm" location (which may be associated with a broad geographical region) may appear highly related to clinical and environmental sources. Industry interpretation, without the benefit of the SNP distances annotation provided on the right side of the figure, may give individual source relatedness more significance than is warranted. Within the cluster (fine bent arrows, 116-153 SNP) and sub-clustered groupings (bolded two-headed arrow, 63-93 SNP), from the SNP range values provided on the left side of the illustration, a greater span

clusters – in this use, a broader grouping of isolates recognized as linked to more than one clonal complex

genome – in this use, the complete set of genetic material present in an organism

of diversification is apparent. Careful analysis of the SNP distances provides an understanding of a greater degree of diversification over time, but subject to significant uncertainty as to what timeframe(s) are represented. This spread of SNP changes may be within a localized or regional environment or may be the result of introduction or repeated introduction from a distant source where diversification is occurring. The knowledge gaps in drivers for genomic stability and diversity remain, and have proven to be, problematic in using phylogenetic data to intelligently and strategically design industry-solutions based root cause investigations and environmental source-tracking in the absence of paired, publicly available traceback information. Used appropriately, these WGS cluster relatedness trees are helpful tools and may provide directional starting points for root cause investigations but, acted on alone, will not provide actionable answers.

As discussed above, another approach to aligning and comparing different isolates is most commonly by generating cgMLST relationships between a large body of characterized isolates with recent isolates (e.g., clinical samples indicating an emerging outbreak; surveillance or environmental sample isolates; research studies or root cause investigative source-tracking). A very simplified example of the process workflow is presented below. cgMLST is often recommended for industry applications by unrooted phylogenetic analysis, as all comparisons are internal and not reliant on a public database.

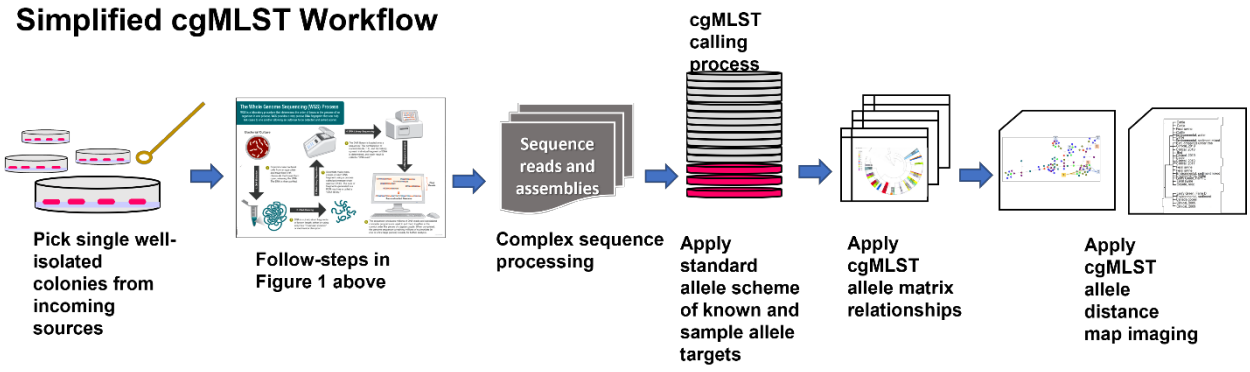
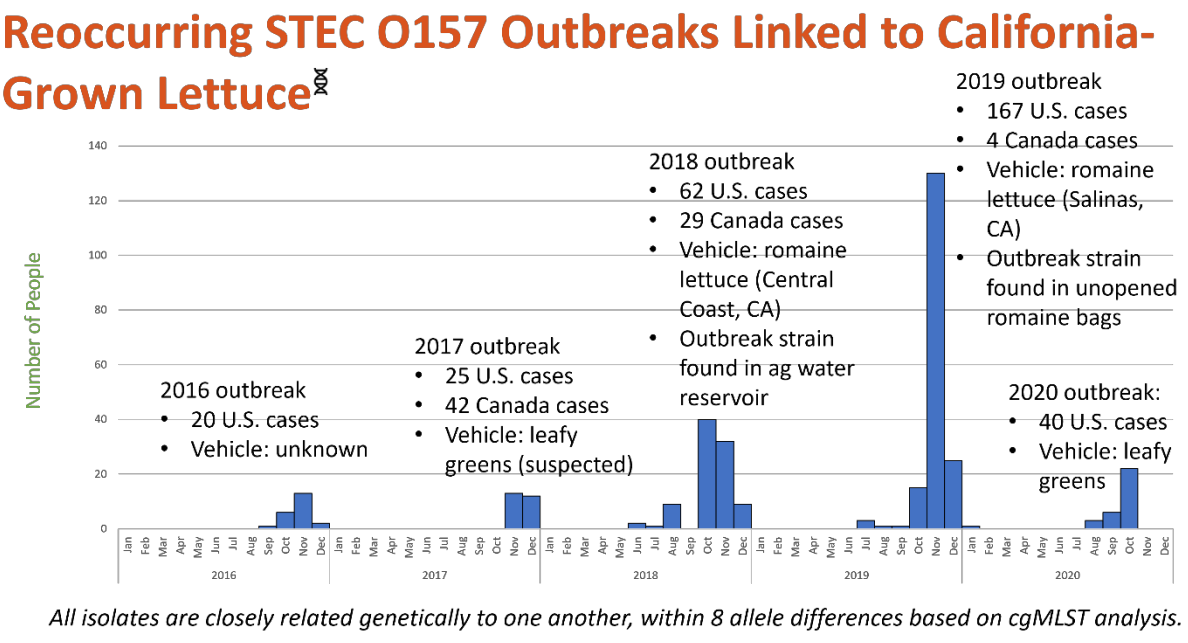


Figure 3. cgMLST is utilized by state and federal agencies, such as the CDC, to rapidly assess potentially linked clinical cases or surveillance samples to a large database of archival profiles and those being collected in "real-time" from reported and characterized clinical cases using a standardized protocol. An outcome example relevant for this STEC Seasonality project is shown below.



All isolates are closely related genetically to one another, within 8 allele differences based on cgMLST analysis.

CDC data are considered preliminary and subject to revision (version provided June 2021)

clade – a grouping of an ancestral type and all the related descendants

Figure 4. A CDC-provided graphic illustrates the approximate seasonal pattern of illnesses of closely related isolates, collectively referred to in this CPS Issue Brief as REPEXH02. As described in the narrative, core-genome MLST (cgMLST) groupings are built from a recognized foundational set of highly related and characterized strains (multiple isolate **genomes**) to define a collection of **alleles** around a set of standard **loci**. Among all isolates that CDC considers to be part of the REPEXH02 strain, they differ from one another by a maximum of 8 allele differences among ~2500 core genome loci analyzed. It is important to be aware that differences detected between alleles for the group may be the result of a single or multiple changes. The use of cgMLST is generally faster and provides more directional information (inclusion or exclusion of cases) for application within the public health system during rapidly emerging outbreaks. For industry internal applications, cgMLST may be the preferred choice, as the reference isolate may be entirely internal and not subject to linkage to a clinical case. cgMLST techniques provide a standardized and more relational approach to identifying and "naming" isolates than are SNP-derived groupings. Subsequent SNP analysis provides an added level of resolution within a cgMLST group. The REPEXH02 group includes isolates that span the California Central Coast region, which is more northerly, and the CA South Central Coast. Further differential clustering, conducted by SNP analysis, within this *E. coli* O157:H7 **clade** from clinical cases, detection in product implicated in the outbreaks, and diverse environmental samples, indicates a secondary level of diversity between the two regions. Close inspection of the dates of the earliest recognized onset of illnesses, spanning June to mid-October, provide critical indicators which remain to be fully aligned with discriminatory details of production locations, practices, harvest timing, processing, and

distribution. The shift in onset dates and peak dates between years are noteworthy data points. While very helpful in a high-level overview, the efforts during this CPS STEC project clearly indicate that overly simplistic approaches to root cause seasonality and environmental source-tracking, based on cgMLST clade-cluster relationships (and treating all onset dates as a continuum of a single event) is likely to mask and impede research and industry efforts at solutions. While not resolved within this 1.5-year effort, extensive discussions and interactions among many stakeholders were achieved. Multiple parallel efforts to continue to search for the knowledge to provide meaningful and sustainable regional solutions is continuing.

Challenges in interpreting a phylogenetic tree:

Looking at a large section or sub-section (cluster clade) as isolates are posted during an outbreak, randomly from surveillance testing, or postings from an academic survey leads to interesting speculation and sometimes intriguing hypothesis generation for root cause investigations. However, it is an unreliable exercise in the absence of an epidemiologic and site-knowledge foundation. Current research has shown that a lineage of foodborne pathogens may have a wide geographic range, with little diversification (SNP or cgMLST based analysis) and no clear linkage to a specific vehicle or regionally defined environmental source. Sometimes, these phylogenetic tree relationships do lead to the discovery of previously unrecognized routes of transference. Movement of cattle interstate and intrastate has been suggested to explain the occurrence of highly related subtypes of the REPEXH02 subtypes of *E. coli* O157:H7 along the CA Central Coast and extending to the South Coast region of California. However, confirmatory details of this hypothesis could not be adequately resolved during the course of the STEC Project. The role of human activities in the dispersal or range extension of this phylogenetically grouping has also been suggested and explored but a without definitive resolution.

Applications to the CPS STEC Seasonality Project:

In response to a series of interactions over the timeframe of the CPS STEC Project, FDA CFSAN provided a numerical assessment and high-level characterization of diverse samples tested during the environmental investigation phase (C. McDermott, Stakeholder Engagement Specialist, U.S. FDA CFSAN Communications). A recent publication provides the background and many details of outbreak investigations and analysis from 2018 and 2019 (Waltenberg et al. 2021).

Among the questions covered in several rounds of dialogue were: What is the breakdown by source type and timing for the 800 samples collected in fall 2019? Soil, air, water, tissue, roadside – how many on-ranch samples, from how many specific ranch locations? Can the sample locations be regionally mapped out or mapped to a greater degree of specificity?

Some key bulleted take-aways are given below:

- Sample numerator (positives) and denominators cannot be compared across sites, as the sampling conducted was variable and dependent on several situational factors of opportunity during investigations. FDA subject matter experts exercise flexibility in targeting sample type and sites, as incoming information may change over the course of the assignment.
- With these caveats, these are the overview breakdown of results:
    - 809 samples from three "farms" (different ranch number and sizes) and public land samples
    - 737 on-farm samples; none matching recurring outbreak isolate (STEC isolates found on each "farm")
    - A descriptive cataloging of STEC genetic profiles not matching EcO157 or non-O157 EHEC pathogens of concern is pending, but a diversity of types is well-recognized within the region from research studies, industry testing programs, and other FDA leafy greens testing assignments
    - 72 public land samples with 25 STEC isolates recovered and 1 matching the recurring REPEHX02 outbreak subtype most associated with the Central Coast region
    - Limited sampling but no positive detections in central Salinas Valley production region, at the time
    - In South Valley, with the greatest number of on-farm and public land samples taken, there was only one detection of the reoccurring outbreak isolate recovered, from a cattle feces composite. The greatest number of STEC positives were recovered from wildlife scat (70 samples of scat with 60 taken from public access points). Subsequent year sampling detected another REPEXH02 isolate in a nearby off-farm location.
    - One STEC was reported from the 20 romaine product samples obtained

Concluding Remarks: Whether outbreak-related, public and private research, or modern tools for industry systems for investigation and preventive control improvements, WGS will continue to develop as a powerful and increasingly cost-reasonable set of technologies. It was not possible or reasonable to cover all applications in this Issue Brief. A few resources that discuss this in some details are provided below. What seems clear, from a general industry and non-WGS specialist food safety educator and professional perspective, is that a high degree of trust and confidence is required, due to the technical and interpretive complexities, as we apply the data being generated to, often, highly sensitive and impactful policies, standards, and direct economic consequences for individuals, a region, and a broad produce commodity or category. A pragmatic approach is to understand that the diverse applications of WGS may provide very clear

and reproducible outcomes and insights but remain prone to overly limited experimental design, consequential misinterpretation, and premature decision-taking when other needed data and context do not exist.

From its inception, the well-intentioned efforts of this STEC Seasonality Project to support broad and collaborative efforts across the produce supply chain most responsible for romaine lettuce and other leafy greens supply and marketing, and the research and regulatory branches of public health agencies, experienced variable levels of success in engagement. It was not surprising that challenges were encountered in assembling and aligning the critical data sets needed to assist and support a collective root cause analysis and investigative effort to develop short and long-term research priorities for CPS and to share with other funding entities. Regardless, the process was worth the effort and several parallel group efforts have and are continuing to pursue similar goals.

Ultimately, the product of this effort is the collection of five Issue Briefs, which are intended to contribute to a better understanding of the intricacies of the issue and the compelling need for a collaborative and cooperative systems-based and integrated multidisciplinary approach of near-term and long-term targeted research.

Acknowledgements: Many individuals were involved in this component of the CPS STEC Project but special acknowledgement is appropriate to DeAnn Davis, Drew McDonald, Jim Brennan, John Gurisi, Tony Banegas, Sam Myoda, Jennifer McEntire, Mark Shakespeare, CDC Division of Foodborne, Waterborne, and Environmental Diseases and Preventive Controls, FDA CFSAN, FDA ORA, Natalie Krout-Greenberg and members of the CA Ag Neighbors Dialogue Group.

Resources

Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. 2019. **The use of next generation sequencing for improving food safety: Translation into practice.** Food Microbiology. 79:96-115. https://doi.org/10.1016/j.fm.2018.11.005

Denamur E, Clermont O, Bonacorsi S, et al. 2021. **The population genetics of pathogenic *Escherichia coli*.** *Nat Rev Microbiol* 19:37–54. https://doi.org/10.1038/s41579-020-0416-x

Deneke C, Uelze L, Brendebach H, Tausch SH, Malorny Burkhard. 2021. **Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST.** Frontiers in Microbiology 12:874-886. https://www.frontiersin.org/article/10.3389/fmicb.2021.649517.

Fitzgerald SF, Lupolova N, Shaaban S, Dallman TJ, Greig D, Allison L, Tongue SC, Evans J, Henry MK, McNeilly TN, Bono JL, Gally DL. 2021. **Genome structural variation in *Escherichia coli* O157:H7.** Microb Genom. 7(11). doi: 10.1099/mgen.0.000682.

Jia M, Georgnaras I, Martin JN, Belk KE, Yang H. 2021. **Comparative whole genome analysis of *Escherichia coli* O157:H7 isolates from feedlot cattle to identify genotypes associated with the presence and absence of *stx* genes.** *Front Microbiol*. 12:647434. doi:10.3389/fmicb.2021.647434

Mageiros L, Méric G, Bayliss SC, et al. 2021. **Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*.** *Nat Commun* 12:765. https://doi.org/10.1038/s41467-021-20988-w

Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. 2008. **Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks.** Proc Natl Acad Sci USA. 105(12):4868-4873. doi: 10.1073/pnas.0710834105.

Murphy R, Palm M, Mustonen V, Warringer J, Farewell A, Parts L, Moradigarav D. 2021. **Genomic epidemiology and evolution of *Escherichia coli* in wild animals in Mexico.** mSphere 6:e00738-20. https://doi.org/10.1128/mSphere.00738-20.

Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. 2018. **Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations.** *Front Microbiol*. 9:1482. doi:10.3389/fmicb.2018.01482

Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. 2021. **Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV).** Genome Res. 31:159-169. doi: 10.1101/gr.266932.120.

Reid CJ, Blau K, Jechalke S, Smalla K, Djordjevic SP. 2020. **Whole genome sequencing of Escherichia coli from store-bought produce.** Frontiers in Microbiology 10: 3050. https://www.frontiersin.org/article/10.3389/fmicb.2019.0305

Stanton E, Wagner S, Florek KR, Kaspar CW. 2020. **Genome sequences of 14 Escherichia coli O157:H7 strains isolated before and during the time frame of the 2018 multistate outbreak associated with romaine lettuce.** Microbiol Resour Announc 9:e00458-20. https://doi.org/10.1128/MRA.00458-20

Uelze L, Grützke J, Borowiak M, et al. 2020. **Typing methods based on whole genome sequencing data.** *One Health Outlook* 2:3. https://doi.org/10.1186/s42522-020-0010-1

Waltenburg M, Schwensohn C, Madad A, Seelman S, Peralta V, Koske S, et al. 2021. **Two multistate outbreaks of a reoccurring Shiga toxin-producing Escherichia coli strain associated with romaine lettuce — United States, 2018–2019**. *Epidemiology and Infection*, 1-22. doi:10.1017/S0950268821002703