Whole Genome Sequencing in Food Safety: Evidence, Interpretation, and Limits

Science Meets Policy: From Regulation to Practice – Turning WGS into Action for Foodborne Outbreak Investigation

NH Collection Giustiniano, Rome, Italy, 02/03 September 2026

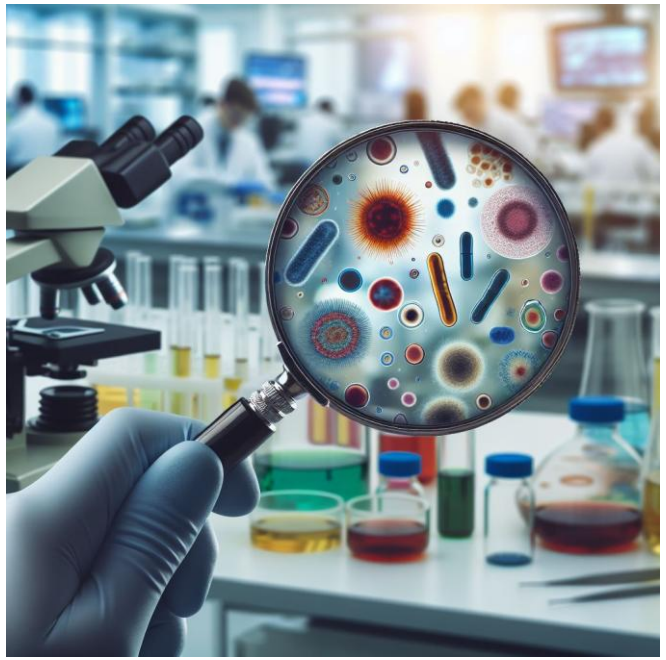
João André Carriço, PhD

AGENDA

- What is WGS in the context of this presentation
- What WGS can and cannot answer
- What is the meaning of thresholds
- Root cause Triangle
- Limitations of WGS
- Database and Governance limits
- Concluding remarks

WGS: WHEN ZOOMING INTO THE DETAILS MATTERS

Detection



Presence/Absence:
Growth Media
RT PCR

Characterization

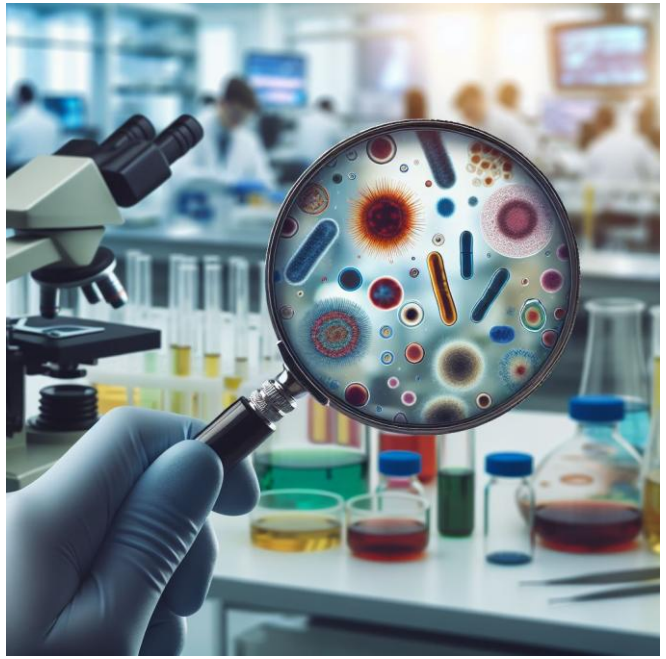
Characteristics :
Serotyping
RT PCR (Specific)

Genomic Analysis

Genomic characterization :
Full strain characterization:
-genomic variation for bacterial typing
-gene presence of Virulence Factors

WGS: WHEN ZOOMING INTO THE DETAILS MATTERS

Detection

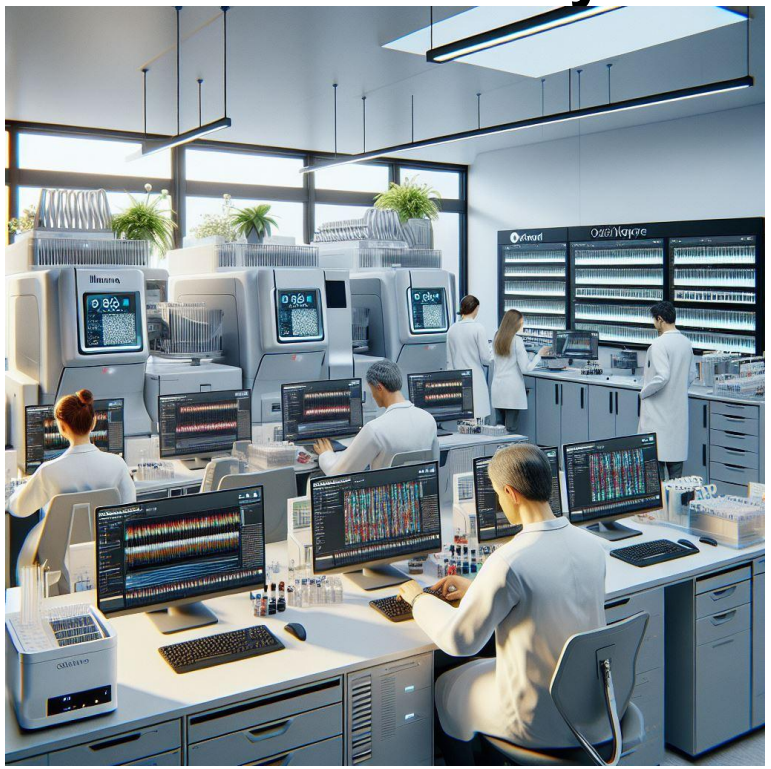


Presence/Absence:
Growth Media
RT PCR

Characterization

Characteristics :
Serotyping
RT PCR (Specific)

Genomic Analysis



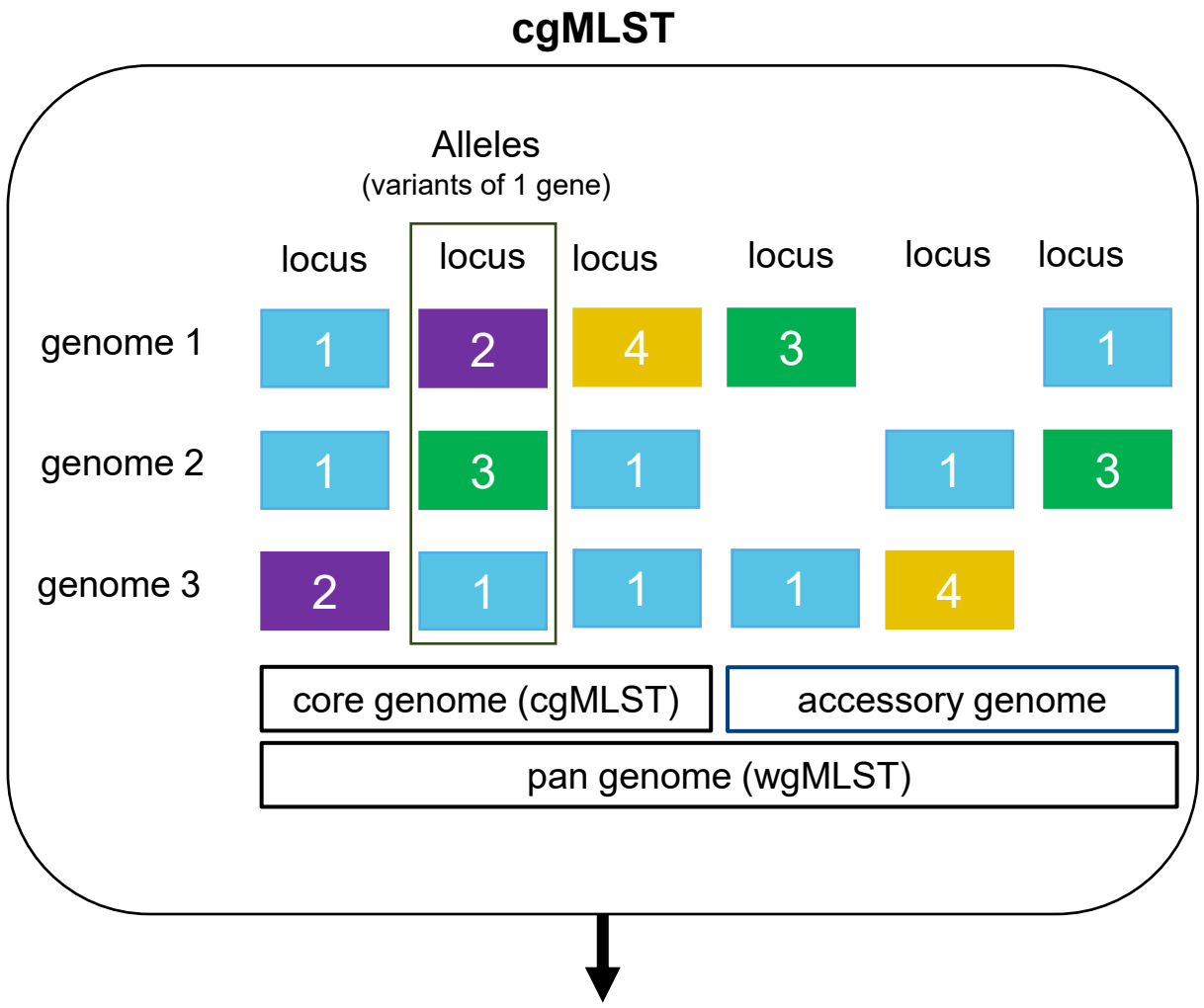
Genomic characterization :
Full strain characterization:
-genomic variation for bacterial typing
-gene presence of Virulence Factors

WGS IN THE CONTEXT OF THIS PRESENTATION

A colony-derived, mostly complete representation of one cultured isolate

- One selected colony does not capture all diversity in the original sample.
- Draft genomes commonly contain multiple contigs and may omit or poorly resolve repeats, plasmids and prophages.
- “Whole genome” means genome-wide data, not necessarily a closed, error-free chromosome.
- Focus is on strain relatedness and not on specific genes or genetic islands of interest (i.e. *stx* in STEC, SPI-1/2 in *Salmonella enterica,* bcrABC for *Listeria monocytogene. etc)*

CORE GENOME MLST : GENE-BY-GENE COMPARISON



Distance between isolates:
number of loci with different allele

- A **predefined schema** assigns an allele number to each conserved locus.
- Distance is the number of loci with different allele calls.
- Portable and efficient for surveillance when laboratories use the same schema and quality rules
- One allele difference may represent one nucleotide or several nucleotides.

Strength: portability

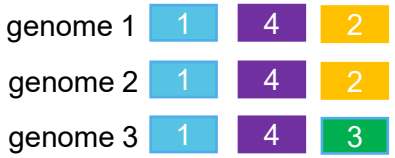
Risk: schema dependence

Core genome only

REFERENCE-BASED SNP ANALYSIS

compare nucleotides (wgSNP)

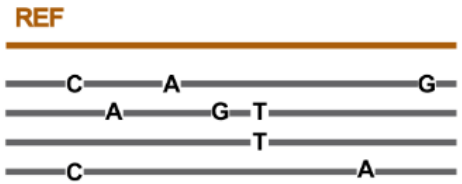
Set of genomes which are highly similar based on cgMLST



Select 1 reference genome from genome set

Map cleaned reads onto reference (for all genomes in the selected set)

Determine positions with a different nucleotide compared to the reference



single nucleotide polymorphisms (SNPs)

Distance between isolates:
number of positions with different nucleotide

- Provides nucleotide-level resolution within the genomic space that passes filtering parameters (Quality and recombination)
- Choice of reference genome, coverage, masking and recombination treatment shape the final SNP matrix.
- Pairwise counts should be interpreted with tree topology, branch structure and appropriate background isolates.

Strength: resolution

Risk: pipeline dependence

Always inspect the phylogeny

CGMLST AND SNP ANALYSIS ARE COMPLEMENTARY

Dimension	cgMLST	Reference-based SNP
Unit	Allele at a defined locus	Nucleotide position
Portability	High within one schema	Requires harmonised reference + pipeline
Resolution	High	Often highest within a lineage
Missing data	Uncalled loci	Uncovered or masked sites
Best use	Routine surveillance	Focused cluster investigation
Main risk	Universal allele cutoff	Universal SNP cutoff

Use cgMLST to screen broadly;
use focused SNP analysis to interrogate a selected lineage.

WHAT WGS CAN AND CANNOT ANSWER

WGS can support

- Candidate common-source clusters
- Repeated-strain detection
- Facility, line or zone comparisons
- Links to clinical, food or regulatory databases
- Virulence and AMR gene screening
- Exclusion or prioritisation of source hypotheses

WGS cannot independently establish

- Direction of contamination
- Exact introduction date
- Continuous persistence
- Direct transmission
- Food-to-case causality
- Identity of an unsampled source

ALL METHODS INFER DISTANCES

Observed distance is a function of biology + time + sampling + method

- **Organism / lineage**
- **Mutation + recombination**
- **Duration of contamination**
- **Population size and bottlenecks**
- **Colony selection**
- **Sampling rate**
- **Background diversity**
- **Analysis Pipeline** (schema/reference/parameters)

Thresholds for defining strain relatedness are operational screening rules, not biological verdicts.

WHAT DOES “0 DIFFERENCES” MEAN?

It supports

- No detected difference in the successfully compared genomic space
- Possible recent common ancestry
- A common-source hypothesis remains plausible

It does not prove

- Biological identity
- Direct transmission
- Direction of contamination
- Origin in the implicated facility
- Absence of unsampled intermediates

**Zero is a property of the comparison,
not a certificate of causality.**

FROM OBSERVATION TO ATTRIBUTION

“Genomically highly related”	→	Laboratory observation
“Consistent with recent common ancestry”	→	Biological interpretation
“Supports a common contamination source”	→	Integrated process interpretation
“Supports attribution to a food or process”	→	Multidisciplinary conclusion

“A matching human and food profile does not, by itself, establish the food as the source.”

ON THE CURRENT REGULATION FAQ

EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR HEALTH AND ANIMAL WELFARE

Brussels, 26 March 2025

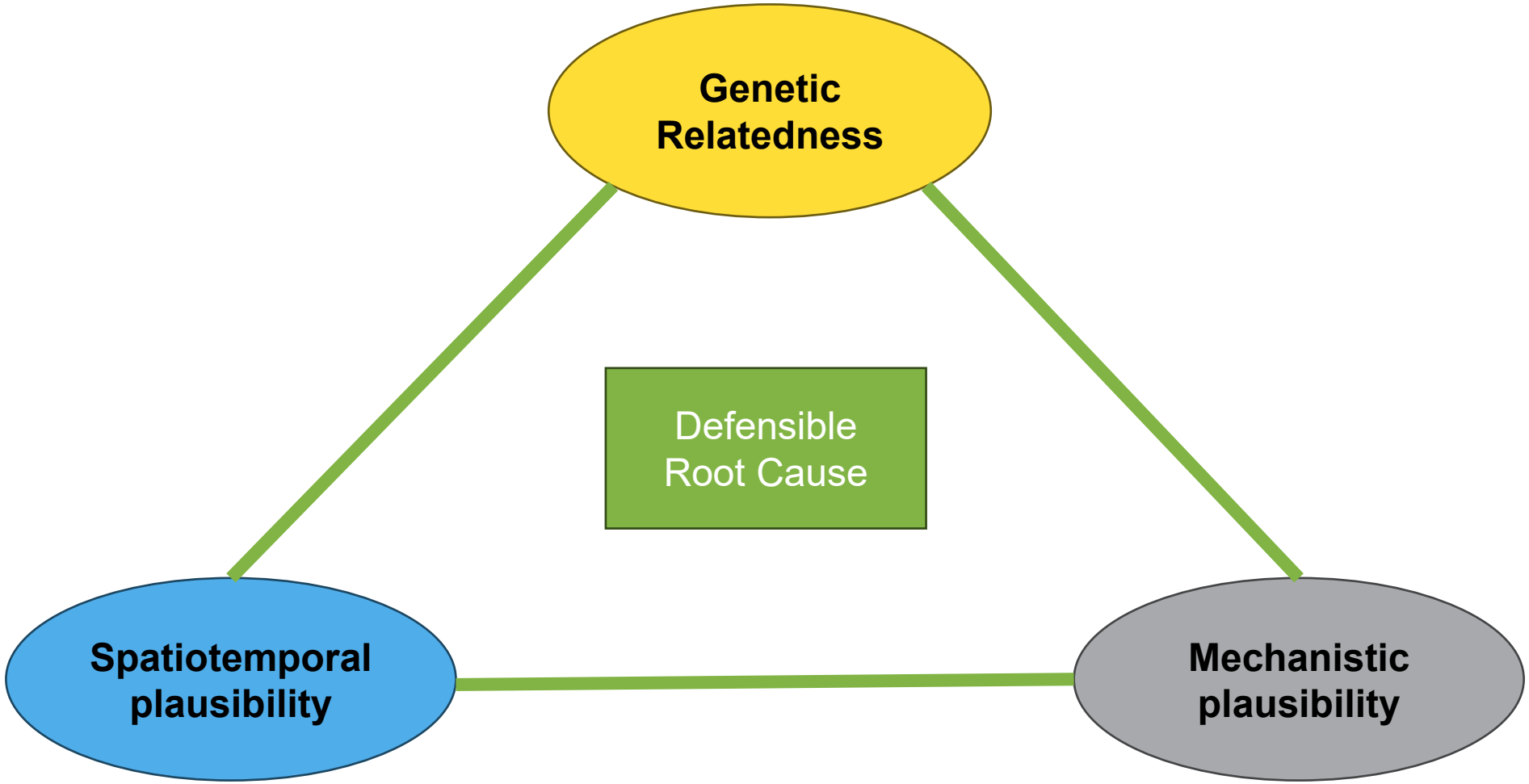
Whole genome sequencing within the frame of foodborne outbreak investigations (implementation of Regulation (EU) 2025/179)

Frequently asked questions

6. WHAT WILL HAPPEN IF THE WGS PROFILE OF A HUMAN ISOLATE DURING AN OUTBREAK CORRESPONDS WITH THE WGS PROFILE OF A CERTAIN FOOD ISOLATE PRESENT IN THE EFSA WGS SYSTEM? WILL THE FOOD BE CONSIDERED AS THE SOURCE OF THE OUTBREAK?

A genetic match of WGS profiles between a human isolate and a food isolate alone does not confirm the food as the outbreak source. Additional epidemiological investigations, such as patient interviews, exposure history, and traceback analysis, are required to establish a link. The WGS data from human and food isolates must be combined with epidemiological investigations to confirm the food source. Common WGS profiles between human isolates and a matrix without (initial) epidemiological evidence, may be sufficient to be considered as a suspicion requesting further investigations e.g. when the production period of a food corresponds to the consumption period of the human cases.

ROOT CAUSE TRIANGLE



SOME PRACTICAL CASES (EVIDENCE)

FALSE POSITIVES DUE TO LAB CONTAMINATION

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Whole genome sequencing used in an industrial context reveals a *Salmonella* laboratory cross-contamination

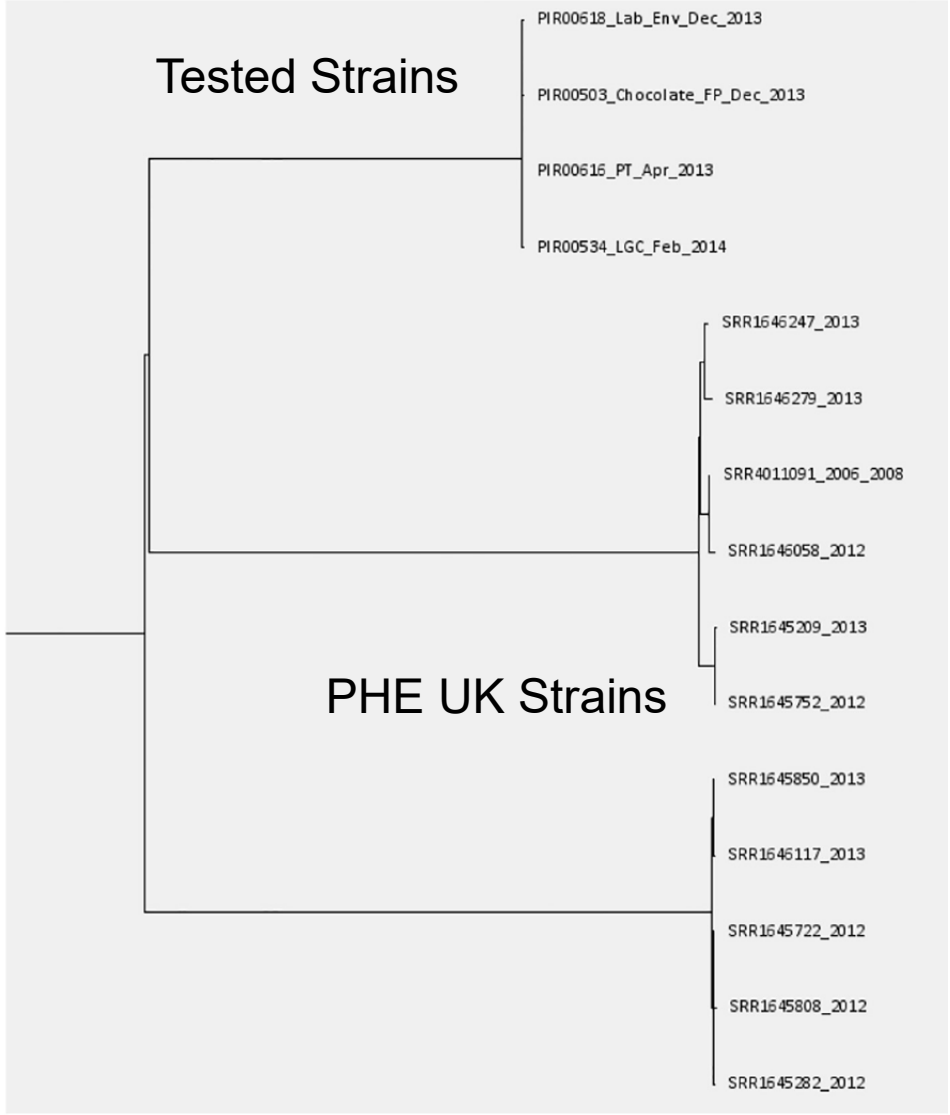
Katia Rouzeau-Szynalski[a,*], Caroline Barretto[a], Coralie Fournier[b], Deborah Moine[b], Johan Gimonet[a], Leen Baert[a]

[a] Nestlé Research, Vers-chez-les-Blanc, 1000 Lausanne 26, Switzerland
[b] Nestlé Research, EPFL Innovation Park, 1015 Lausanne, Switzerland

S. Hadar isolates involved in the suspected laboratory cross-contamination event.

Isolate code	Origin of *S.* Hadar isolate
PIR00616	Proficiency test (PT) sample spiked with strain NCTC 9877
PIR00618	Laboratory environmental sample (from the thermocouple in the incubator)
PIR00503	Finished product sample (chocolate)
PIR00534	Reference strain NCTC 9877, originally used in the PT

[a] Strain acquired from LGC bacterial reference collection.



SNP analysis: CFSAN Pipeline v1.0.0
Reference : PacBio sequencing and assembly of NTCT 9877

IMPORTANCE OF TREE TOPOLOGY

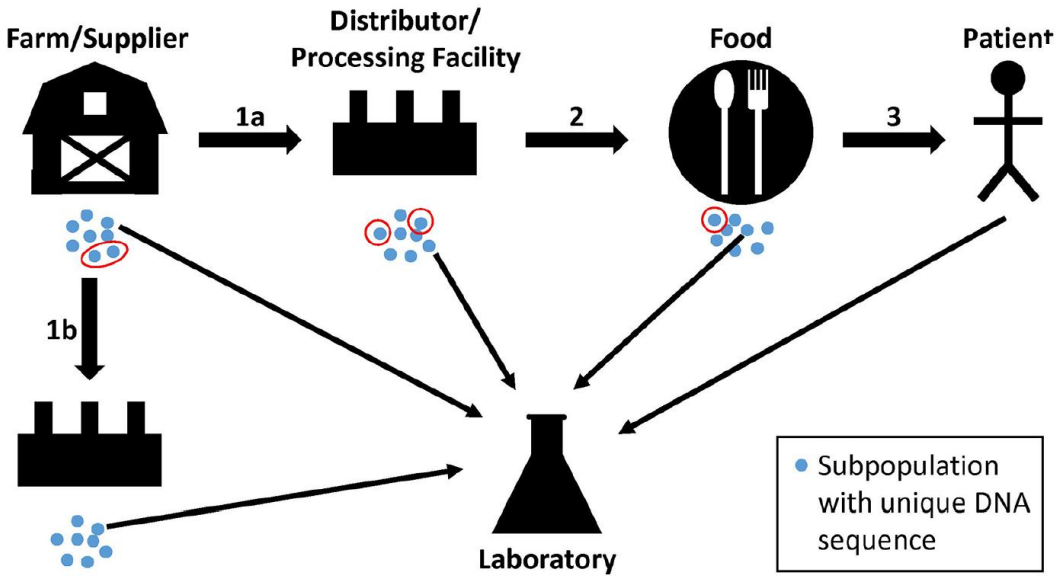
frontiers in Microbiology

TECHNOLOGY REPORT
published: 10 July 2018
doi: 10.3389/fmicb.2018.01

Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations

Arthur W. Pightling, James B. Pettengill, Yan Luo, Joseph D. Baugher, Hugh Rand and Errol Strain*

Biostatistics and Bioinformatics, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, United States



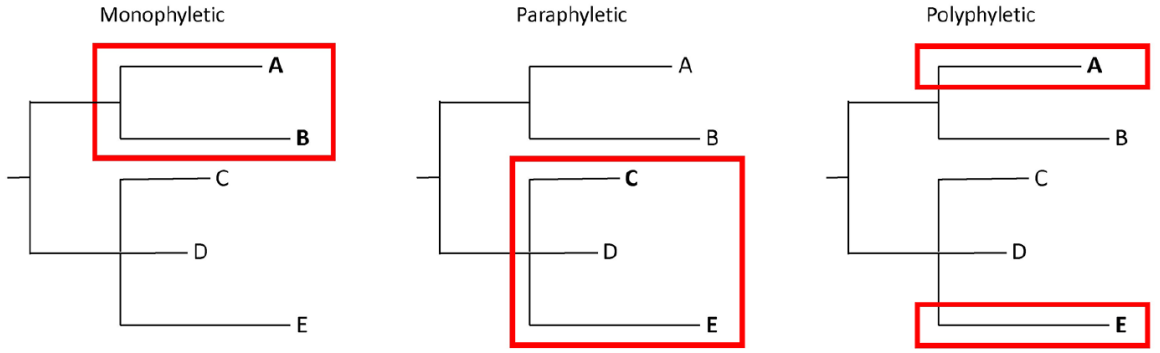
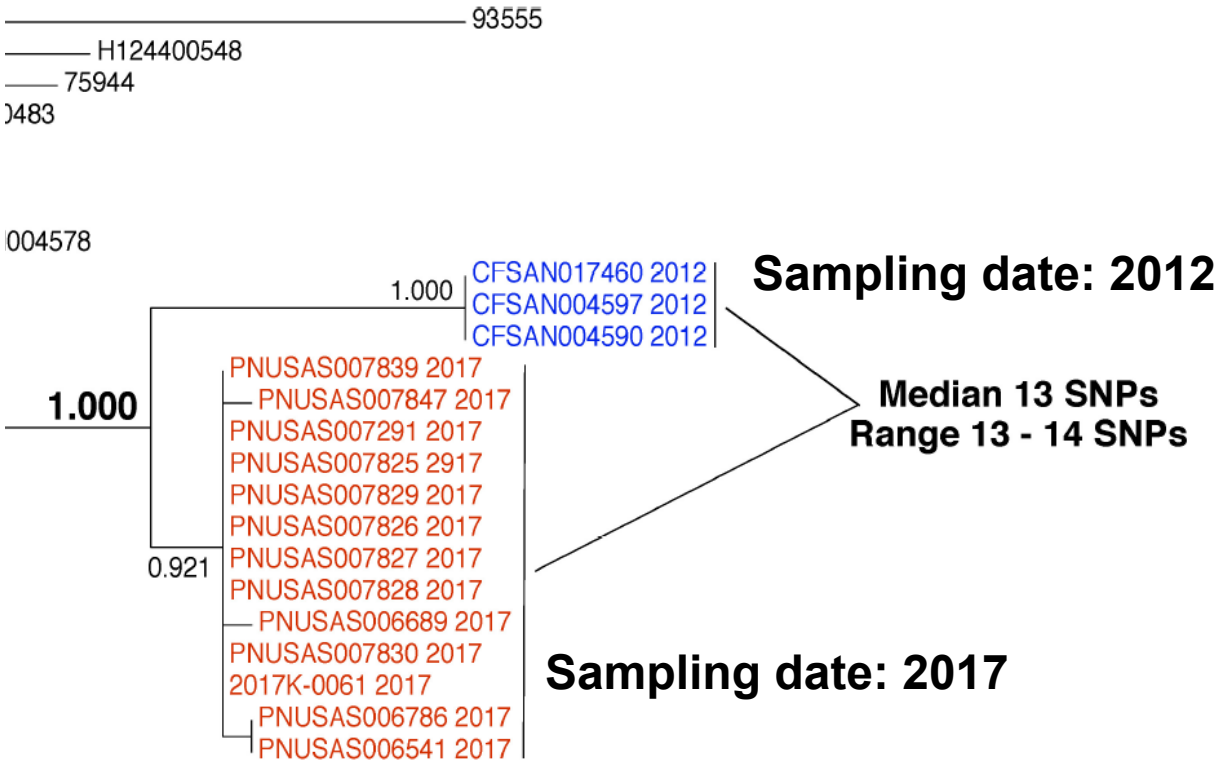


TABLE 2 | Conditions used to determine whether whole-genome sequence analyses support a match between two or more genomes.

	Supports	Neutral	Does not support
SNP distance	<21	21–100	>100
Bootstrap support	>0.89	0.80–0.89	<0.80
Tree topology	Monophyletic	Paraphyletic	Polyphyletic

IMPORTANCE OF TREE TOPOLOGY

This evidence indicates that these isolates arose from the same source. However, inspectors collected the environmental isolates in 2012 and the patients were sickened in 2017. Also, there was no epidemiological signal for the commodity produced by this facility. Finally, no traceback evidence linked the patients to the food product. Therefore, the connection between these environmental and clinical isolates is unclear. A conclusion that the isolates originated from the same source by WGS analysis can be made but the epidemiological and traceback evidence do not link them. The tree topology offers more information that suggests the isolates might not be linked. The environmental and clinical isolates form distinct subgroups, rather than being interspersed. These data show that although the clinical isolates originated from the same source of contamination, that source may not have been the food processing facility. With SNP thresholds, it is possible that investigators could have erroneously concluded that product from the food processing facility caused the illnesses.



om *Salmonella enterica* isolates collected from a food processing facility and closely related clinical isolates. A SNP matrix was generated for 24 isolates with the CFSAN SNP pipeline v0.8.0 (Davis et al., 2015). The SNP matrix was phylogenetically analyzed with GARLI v2.01.1067 (Zwicki, 2006), using the K80 and HKY nucleotide substitution models. Branch labels indicate support values for 1000 bootstrap replicates. Bootstrap values less than 0.800 are not shown. Blue labels indicate environmental isolates and red labels indicate clinical isolates. Black labels indicate isolates that are not the focus of this example but are closely related.

PERSISTENCE AND REINTRODUCTION

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

FOOD MICROBIOLOGY

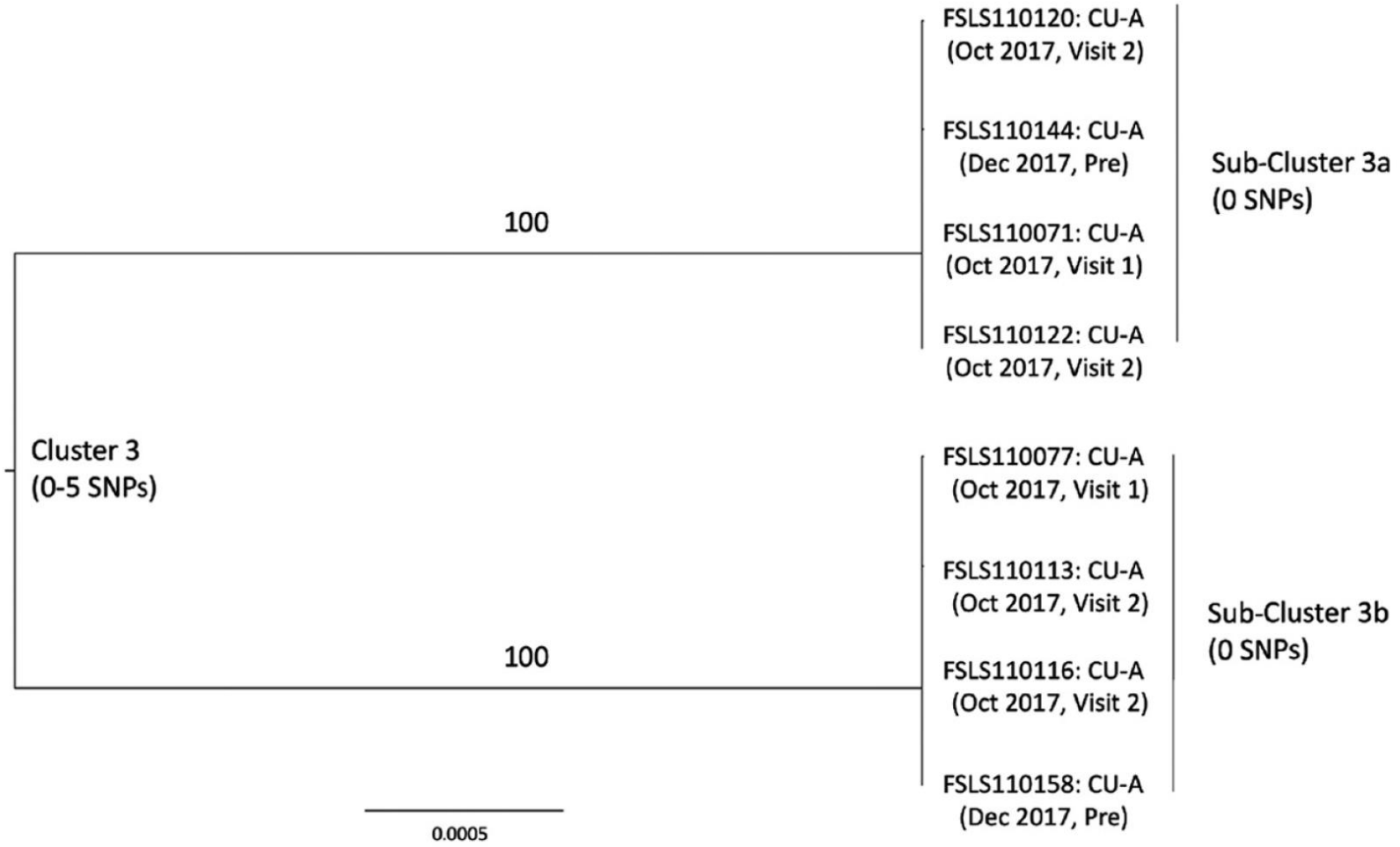
Whole-Genome Sequencing-Based Characterization of *Listeria* Isolates from Produce Packinghouses and Fresh-Cut Facilities Suggests Both Persistence and Reintroduction of Fully Virulent *L. monocytogenes*

Genevieve Sullivan,[a] Renato H. Orsi,[a] Erika Estrada,[b] Laura Strawn,[b] Martin Wiedmann[a]

[a]Department of Food Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, New York, USA
[b]Department of Food Science and Technology, Eastern Shore Agricultural Research and Extension Center, Virginia Tech, Painter, Virginia, USA

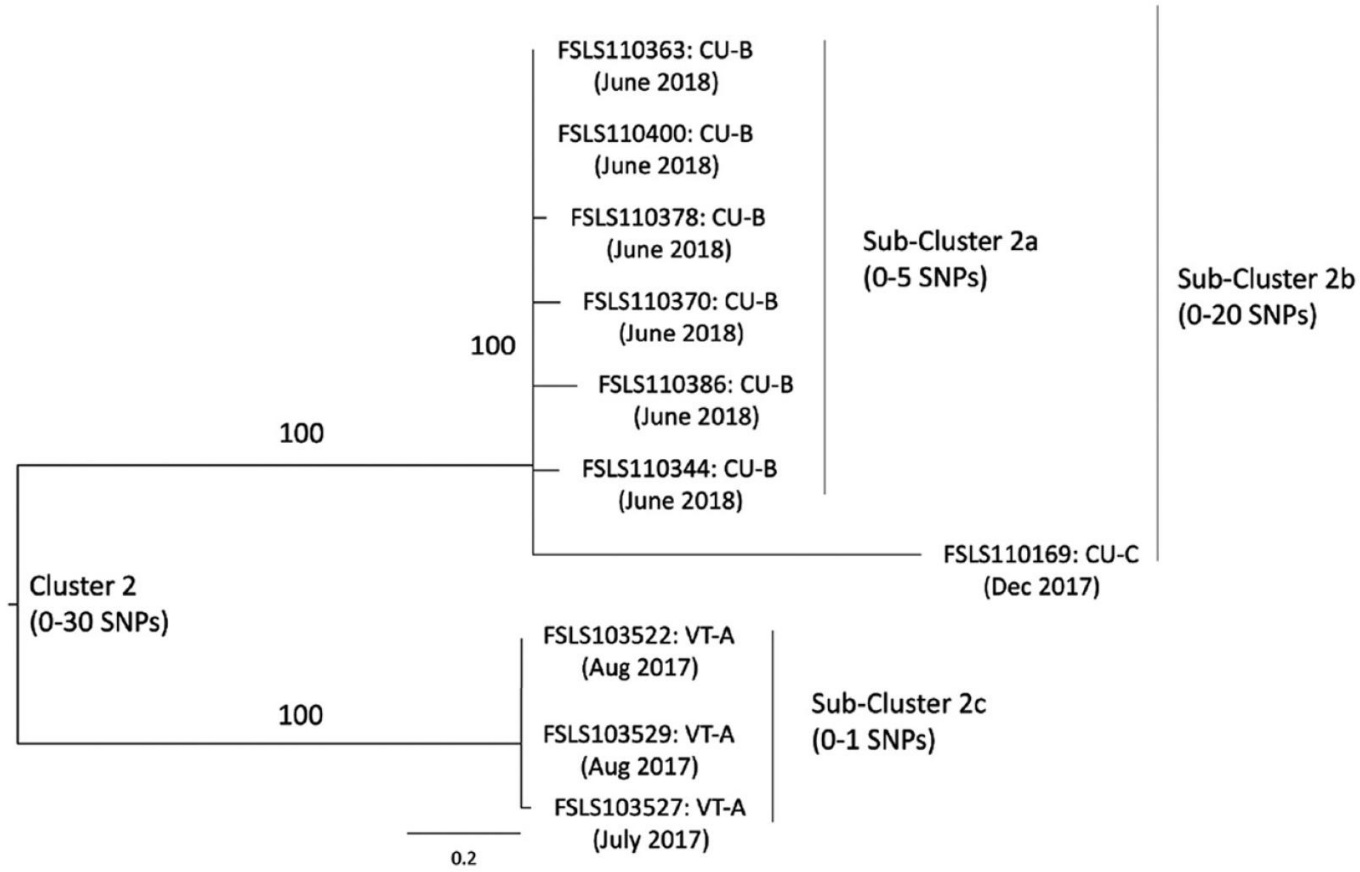
- 169 L. monocytogenes + 107 other Listeria isolates (13 packinghouses and 3 fresh-cut facilities)
- Analysis done with CFSAN SNP pipeline (v1.0.1)
- Highly related isolates, defined within the study as fewer than ten hqSNP differences, were recovered at least two months apart in seven of 16 operations;
- Two operations had related strains more than one year apart. Pre-production and production detections supported survival through sanitation in some settings.

PERSISTENCE AND REINTRODUCTION



Persistence

PERSISTENCE AND REINTRODUCTION



- **Highly related isolates were found in separate, independently owned operations**
- **2 Packing houses in the same US Region**
- **Potential common upstream source (Causing reintroduction) ?**

LIMITATIONS OF WGS

SAMPLING LIMITATIONS

A negative sample is not proof of absence

Contamination may be localised, intermittent, below detection, removed by cleaning, present in a historic lot or located in an unsampled niche.

A positive match is not proof of direction

Product may contaminate drain; drain may contaminate product; both may share an upstream source; or an unsampled reservoir may connect them.

Hypothesis-driven spatial and temporal sampling beats convenience sampling.

BIOLOGICAL LIMITATIONS

Mutation is not one clock

Rates vary by species, lineage, environment and genomic region. Recombination, selection and bottlenecks complicate time inference.

Persistence is not a genome-only claim

Repeated close isolates may reflect a resident niche, repeated supplier introduction, shared reservoir or reintroduction.

One colony hides diversity

A selected colony is a single representation of a potentially mixed population. Multi-colony sampling may change the inferred cluster.

SEQUENCING AND BIOINFORMATICS LIMITATIONS

SEQUENCING

- Contamination
- Same species
- Different species
- Low / Uneven coverage
- Barcode bleaching
- Homopolymer/Methylation errors
- Metadata errors
- Time to result

BIOINFORMATICS

- Assembly fragmentation
- Handling missing cgMLST loci
- Reference selection for SNP analysis
- Handling recombination
- Repeat masking
- Plasmid/ prophage treatment
- Diferent Software and versions
- Database bias

DATABASE AND GOVERNANCE LIMITATIONS

- Absence of a match may mean that the source was never sequenced or cannot be accessed.
- A close match may represent a widespread clone, not a direct link to the named facility or country.
- Industry decisions must account for confidentiality, ownership, regulatory discovery, retention, legal interpretation and brand impact

Governance questions

Who owns the sequence? And the isolate?

Who may search it?

Which metadata are disclosed?

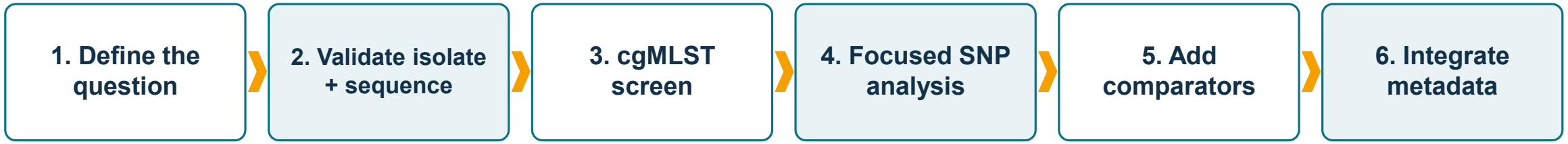
How long is it retained?

What decisions may follow a match?

Representativeness determines confidence.

CONCLUSIONS

A DEFENSIBLE WGS DECISION WORKFLOW



Competing hypotheses

H1 persistent niche
H2 repeated raw-material introduction
H3 laboratory cross-contamination
H4 widespread background clone
H5 independent events

Metadata map

Date + exact location
Hygienic zone + line
Product + lot + supplier
Sanitation + production sequence
Personnel and equipment movement

Analyse the question, not just the sequences.

FIVE TAKE HOME MESSAGES

1. **WGS is a high-resolution comparison method, not an automatic source-attribution test.**
2. **cgMLST and SNP analysis are complementary,** but their distances are method-dependent.
3. **No universal SNP or allele cutoff separates related from unrelated isolates.**
4. **Interpret similarity with phylogeny, time, place, sampling, epidemiology and process knowledge.**
5. **The best investigations compare alternative hypotheses and communicate uncertainty.**

WGS narrow the plausible explanations.
Root-cause analysis selects the explanation that best fits all evidence.

PIONEERING DIAGNOSTICS