nature portfolio

Peer Review File

# Multi-country and intersectoral assessment of cluster congruence between pipelines for genomics surveillance of foodborne pathogens

Corresponding Author: Dr Vitor Borges

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

**Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Borges review:
(1) Line 76-77 has no mention by the authors of the NCBI Pathogen Detection pipeline that is publicly available with over 2 million human pathogens being tracked and over 150,000 clusters of pathogens within 50 SNPS. The NCBI Pathogen detection portal is a SNP based tool that has been developed and implemented for over 10 years, yet the authors appear to not have mentioned or integrated this system into any of their research. This is problematic.
(2) The authors did not compare their data to the NCBI Pathogen Detection pipeline currently being used by US government agencies on a daily basis. The pipeline should at least be cited somewhere in the paper if not employed.
(3) Line 132-134. The authors need to report on data availability and make all of the data discussed in this study publicly available for reproducibility.
(4) Line 240 Authors did not mention or compare the CFSAN SNP pipeline that has been used extensively in the US. This should at least be cited and/or discussed.
(5) Line 244-245 The importance of selecting a close reference is not a new idea and the authors fail to cite the earlier literature supporting this idea. This should be fixed.
(6) Line 329-330 The documentation of polyphyletic serotypes is not a new idea, but the authors do not cite or refer to the existing publications on this topic. Please amend.
(7) Line 471 The inability of most of this data to pass the QC of bionumerics suggests that some of the discrepancies are due to poor quality data which likely affect outcomes from the various pipelines. The authors should include discussion of the issue of QC thresholds for comparison. A higher quality threshold generally improves comparison among alternative clustering methods.
(8) Line 861 the authors do not discuss the published reporting of different thresholds for outbreak detection among different serotypes and species but rather try to present this as new observations, which apparently appears to be only new to these authors. please cite relevant literature. (9) Recommendations for thresholds for common foodborne pathogens have been available for some time but go uncited here. This needs amended.
(9) Line 980 the data presented supports the notion that harmonization and standardization are needed for broad integration among these diverse approaches to cluster detection if the region wishes to share data to solve outbreaks sooner. This topic was not discussed by the authors nor did they cite any literature supporting this obvious notion.
(10) This paper seems extremely over-written and could be cut by about one-third without affecting any of the key messages or data presentations.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Please see attachment

(Remarks on code availability)
I am only reporting here to say that I did verify the existence of the code available in Github. Multiple scripts are available and well-organized. I cannot verify the python code itself, however, as I do not have expertise in this language.

Reviewer #3

(Remarks to the Author)
The authors present an extensive comparative study of bioinformatics pipelines for genomic surveillance of Listeria monocytogenes, Salmonella enterica, Escherichia coli and Campylobacter jejuni, with the aim of comparing the resolution power and the output in terms of identification of clusters. Such studies are pivotal for translating into practice the wealth of knowledge in bacterial genomic analysis which has been produced in the last ten years, and are still widely lacking in the literature, despite the wide amount of genomic sequences produced and published.
The study was conducted in the framework of a European research project, BeONE, taking advantage of the consortium ability and resources to conduct parallel analysis of the same wide set of sequences by using different approaches, e.g. cgMLST, SNPs analysis and wgMLST, and different tools, which have been extensively compared in this study.
The results are very interesting and useful for surveillance, and generally well presented. However, there are some minor points that I would suggest to revise for improving the readability of the text.

General comments:

The dataset of sequences used is very wide. In detail, L. monocytogenes dataset is representative of more than 40 STs, with the isolates well dispersed across the different STs (fig. 2A). The dataset for S. enterica is also wide and dispersed across STs and serotypes, with STs 11, 34 and 19 covering the majority of the strains, reflecting the abundance of isolation and sequencing of the strains most often associated to disease (fig 5A). The dataset of C. jejuni is also wide and well dispersed across the STs, well capturing the variability of the species. E. coli dataset is wide, even if smaller than the others, and mainly represented by ST11 strains, mainly belonging to O157 serogroup (856/2300 strains) (Fig 8A). The authors state that this bias reflects the wider abundance of this serogroup in public databases (lines 486-488). Nevertheless, in Europe O157 is not the most common serogroup isolated from severe disease any longer: “The European Union One Health 2022 Zoonoses Report” describes that the most common serogroup among Haemolytic Uremic Syndrome cases in 2022 was O26 (51.3%) followed by O157 (14.5%). The variability of genomic stability across different E. coli serogroups is thus not represented in the presented analysis and the evaluations made on the thresholds considered in the paper could not be valuable for the other serogroups of E. coli. It would have been worth including more O26 sequences in the analysis, which would have had a big added value for interpretation of results during the real outbreak investigations which currently mainly involve O26 strains in Europe. If this is not possible at this stage, the authors should refer to this problem and make this bias clear in the paper.

One of the main important and useful results of this study is the identification of different ranges of allelic distances for different STs of the same species. This is a very useful information for deciding the “outbreak thresholds” within a species and, as stated at lines 861-863, “it underscores the potential inadequacy of applying a single “outbreak threshold” within a species, as this may lead to an oversizing of potential outbreak clusters when highly clonal WGS derived lineages (often overlapping with STs/serotypes) are involved”. An effort should be made to present such result of variability of thresholds in one table or one graph including all the species and main STs tested, to be included in the manuscript itself (not in additional files). Similarly, it would be helpful to summarize in a table the results of the concordance for outbreak detection obtained for all the species (e.g. for each species, percentage of the clusters detected by all the pipelines by using a static threshold, with exceptions, and difference between the AD thresholds required by the different allele-based pipelines to detect each “outbreak-level” cluster).

Specific comments:

The text is very long and sometimes redundant. For example, the text presented in Additional file 1 is repeated in several sections along the manuscript (e.g. lines 140-146 or 1178-1185). The manuscript could contain less details than the additional file, or alternatively explain the same concept in easier words, citing additional file 1 for more details.

Figures 2C, 5C, 8C and 11C (graphs of block of stability regions): this analysis is not well explained in any section of the manuscript. Which is the parameter on y axis for each pipeline?

Line 184: which table in additional file 4 is cited here? S1.1? Please specify

Line 329: please better explain: which type of clusters (which AD)?

Line 330: please verify consistency with fig. 5D: should Thompson or Typhi be cited here?

Lines 504-508: please better explain the PopPUNK analysis, as the aim is not clear in the text

Line 604: “later” should be corrected in “latter”

Lines 802-804: the use of a dynamic wgMLST is not easy to apply as a standardised method, for example to compare

sequences across Member States at the European level. The use of larger schemas of cgMLST was proved effective also for C. jejuni. Why are you suggesting the use of wgMLST approach instead of extended schemas of cgMLST? Please better explain the possible application for the suggested approach for surveillance or consider revising.

Lines 1002-1004: how were the reads tested for this analysis?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
General comments:

I thank the authors for their attention in responding to comments provided in the previous review. While I still believe that the manuscript is overly long, and shortening the manuscript would improve its clarity and readability, I do respect the amount of work being presented, and understand the authors' preference to keep the analyses together rather than split the paper into multiple reports.
Overall, the work presented in the manuscript appears sound and the authors have addressed major concerns raised in the first review. I do think that moving the extensive methods section to the end of the manuscript improved readability. At this time, I do not have any additional major concerns or comments related to the manuscript beyond those that were already provided.

The authors have provided a comprehensive comparison of clustering results derived from multiple bioinformatic pipelines in use by the European public health surveillance community for different bacterial species. While the work here may not provide the answers to questions like: 'which pipeline is best', their exploration of differences in clustering resulting from the different pipelines are an important addition to the field, and should be highly useful to members of the community practicing in the space of WGS-based infectious disease surveillance and outbreak detection. As the authors have noted several times, it is unlikely that we will ever see a single standardized approach put to use across the surveillance community. It is therefore important that work like this is published so that the community is aware of the differences that result from the use of different pipelines, and can make informed decisions on approaches used for national surveillance.

I do have a couple very minor suggestions:
• Food and waterborne diseases is first defined as "FWDs" but is written "FWD" in several locations in the introduction (e.g., L122, L125) and Discussion.
• Thank you for providing information about specific pipelines in use for different bacterial species in Table 1. If I may make a small suggestion, please define the abbreviations Lm, Se, Ec, and Cj somewhere in the table description or in the footer

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The Authors revised the paper according to the reviewers' comments raised in the first round of the revision process. Many of the raised points were addressed and the text has gained readability. The text has not been shortened, as requested by all the three reviewers, and it remains very long. Nevertheless, the topic is very difficult to be shortened and relevant to be published, so I think it could be published in the current form, to keep all the relevant information in the main body of the manuscript. I do not recommend to move the Materials and Methods section in the supplementary material, as proposed by the authors, as they are an important part of the work and ensure a full understanding of the study.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I was recruited to assess whether the authors have satisfactorily addressed or rebutted Reviewer 1's comments

I do want to point out that for a multicenter comparison of computational pipelines, it is important to ensure and report if the same computational environment has been used across centers. It's well known that the same pipeline deployed in different environments can produce different results (https://www.nature.com/articles/nbt.3820, https://pmc.ncbi.nlm.nih.gov/articles/PMC3842296/). Was there any effort to ensure the reproducibility of the analyses?

Comment 1. I agree with Reviewer 1 that not evaluating computational pipelines that major agencies outside of Europe have widely used is a limitation, especially since such agencies have generated or curated the majority of the publicly available genome data of common foodborne pathogens. This limitation should be acknowledged and justified in the manuscript. I don't think the authors have adequately addressed this limitation by just citing NCBI Pathogen Detection and the CFSAN pipeline.

Comment 2. To my knowledge, NCBI Pathogen Detection's SNP pipeline is not publicly available for local implementation and evaluation. By citing it, the authors adequately addressed this comment.

Comment 3. The response is adequate.

Comment 4. If the CFSAN pipeline is referenced and discussed, so should be CDC's pipeline https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.00375/full, especially since this is one of first and most cited papers that compares SNP and cgMLST.

Comment 5. The response is adequate.

Comment 6. The response is adequate.

Comment 7. The response is adequate.

Comment 8. The response is adequate but a bit blunt. Literature outside of Europe, such as https://academic.oup.com/cid/article/63/3/380/2595027 can be referenced.

Comment 9. The response itself makes sense given the logistic challenges of harmonization and standardization, but it is not adequate because it ignores Reviewer 1's comment to include a discussion on this topic in the manuscript, which I find reasonable.

Comment 10. The response is adequate.

(Remarks on code availability)

## REVIEWER COMMENTS

**Reviewer #1**

(1) Line 76-77 has no mention by the authors of the NCBI Pathogen Detection pipeline that is publicly available with over 2 million human pathogens being tracked and over 150,000 clusters of pathogens within 50 SNPS. The NCBI Pathogen detection portal is a SNP based tool that has been developed and implemented for over 10 years, yet the authors appear to not have mentioned or integrated this system into any of their research. This is problematic.

**Response:** We thank the Reviewer for all the comments and suggestions. We now mention this tool in the Introduction section, as suggested. For clarification, this study was conducted on behalf of a large European consortium, including key institutes involved in the surveillance of foodborne pathogens, and, thus, relied on WGS bioinformatics pipelines routinely used by these laboratories (which cover the most commonly used methods in Europe). This constituted an unprecedented effort in the field, not only due to the large amount of pipelines analysed and main FWD pathogens covered, but more importantly, because it showcased the feasibility of conducting large-scale comparability assessments, while leveraging a new methodology for such evaluations, in line with the recent demand of WHO to ensure method comparability at regional, national and international levels (https://www.who.int/publications/i/item/9789240021242). Said that, although we fully agree that it is of interest to assess the congruence of other tools (namely those with high applicability, like the NCBI tool or CFSAN), we believe that the reviewer can understand that adding more pipelines was impractical (namely due to the time and budget constraints of the European project), and would raise additional concerns on how to select them. Nevertheless, recognizing the relevance of having other pipelines tested with our methodology, we were fully committed to this topic and thus directed big efforts to publicly deliver all the tools needed to conduct such demanding follow-up comparisons (including other pipelines, datasets, etc.) towards the global One Health FWD genomic surveillance envisaged by WHO and other entities. We consider that we reached a good trade-off by overcoming the impossibility to cover "all" pipelines with the provision of means that facilitate follow up studies, even for other pathogens. For instance, the new methodology is already in place in the ongoing EU-USA collaborative effort - *Legionella* International typing group - to implement a new pipeline for *Legionella pneumophila* genomic surveillance (poster presented in the ESGLI 2024 meeting; https://www.escmid.org/fileadmin/escmid/media/study_groups/ESGLI/ESGLI_Newsletter_December_2024_final.pdf).

(2) The authors did not compare their data to the NCBI Pathogen Detection pipeline currently being used by US government agencies on a daily basis. The pipeline should at least be cited somewhere in the paper if not employed.

**Response:** We have now cited the NCBI pipeline. Please also see the previous answer.

(3) Line 132-134. The authors need to report on data availability and make all of the data discussed in this study publicly available for reproducibility.

**Response:** At the time of the first submission of this manuscript, all the data (both reads and assemblies) and the scripts developed and used in this study were already publicly available in public repositories, as stated across the text, namely in the Methods and the Data Availability sections. Moreover, extensive details on the results obtained are available in the Additional files.

(4) Line 240 Authors did not mention or compare the CFSAN SNP pipeline that has been used extensively in the US. This should at least be cited and/or discussed.

**Response:** Please see the answer to the first question. The CFSAN pipeline was already mentioned in the Discussion, but it was now added to the Introduction.

(5) Line 244-245 The importance of selecting a close reference is not a new idea and the authors fail to cite the earlier literature supporting this idea. This should be fixed.

**Response:** The usage of a closely-related reference (e.g., same ST/serotype) is in place in several laboratories, including some of those involved in this study. This was exactly why we included this approach in the study and not to point it out as a "new idea", as the reviewer interpreted. Our data showed that it increases resolution in comparison with a "single reference" strategy, which was described as an expected observation (in the lines the reviewer indicated). Still, we have now smoothed the sentence to avoid further misinterpretations.

(6) Line 329-330 The documentation of polyphyletic serotypes is not a new idea, but the authors do not cite or refer to the existing publications on this topic. Please amend.

**Response:** We are aware of previous literature supporting the existence of polyphyletic STs/Serotypes, which was cited throughout the manuscript (including the lines indicated by the reviewer) (e.g. 10.1099/mgen.0.000906, 10.1128/JB.00969-10 or 10.1111/1462-2920.15111). This previous knowledge was actually the trigger for us to explore this "old" subject with new and vast datasets, rendering fruitful results and observations. We have now reinforced the acknowledgement of previous insights on this topic in the Discussion (lines 856-861).

(7) Line 471 The inability of most of this data to pass the QC of bionumerics suggests that some of the discrepancies are due to poor quality data which likely affect outcomes from the various pipelines. The authors should include discussion of the issue of QC thresholds for comparison. A higher quality threshold generally improves comparison among alternative clustering methods.

**Response:** For clarification, we need to first underline that all datasets were subjected to rigorous QC (see Methods), and that BioNumerics had an outlier behavior for the *E. coli* dataset (for which the other pipelines retained the large majority of the samples). As such, contrary to the reviewer's point of view, our interpretation is that the problem was not the dataset but instead this specific pipeline, which we excluded from *E. coli* downstream analysis to avoid biasing the comparisons. Moreover, we should also clarify that the goal of the study was to provide a realistic picture of how the results obtained by different laboratories in their routine activities can be compared between each other when using the same sequencing dataset. As such, the differential behavior of the different pipelines in QC inclusion/exclusion is per se a result with relevance for the community.

(8) Line 861 the authors do not discuss the published reporting of different thresholds for outbreak detection among different serotypes and species but rather try to present this as new observations, which apparently appears to be only new to these authors. please cite relevant literature. (9) Recommendations for thresholds for common foodborne pathogens have been available for some time but go uncited here. This needs amended.

**Response:** Regarding the last point, we really did not understand what the reviewer meant by "Recommendations for thresholds for common foodborne pathogens have been available for some time but go uncited here", because our outbreak-level analysis precisely relied on thresholds that we selected after consulting vast scientific literature and official recommendations (e.g., ECDC/EFSA outbreak cases definitions) for each species. The selected thresholds were clearly justified and cited in the respective species-specific sections (lines 250-252, 279-285, 405-410, 437-440, 535-540 and 677-680). Regarding the application of different thresholds for outbreak detection among different ST/serotypes, we honestly think that this concept was not pointed out as "new". Instead, our study provided additional evidence to this topic (which has an increasing interest among the community), suggesting that, as other references that were cited (e.g. references 38, 39, 88), the application of a single "outbreak threshold" within a species might be inadequate, as, for instance, it might potentially lead to an oversizing of potential outbreak clusters when highly clonal WGS-derived lineages (or STs/serotypes) are involved.

(9) Line 980 the data presented supports the notion that harmonization and standardization are needed for broad integration among these diverse approaches to cluster detection if the region wishes to share data to solve outbreaks sooner. This topic was not discussed by the authors nor did they cite any literature supporting this obvious notion.

**Response:** Although the harmonization of the methods (i.e., all labs use the same pipeline and parameters) would be the ideal scenario, such goal seems utopic considering the diversity of pipelines under usage worldwide, and the constant advancements in the bioinformatics field with new tools and pipeline components being constantly improving/emerging. In this context, our data supports the alternative idea that a smoother and efficient communication and cooperation can be reached, provided that the different players have a good understanding on how their pipelines compare to the others for outbreak detection and investigation. This notion is the basis of the study, as stressed throughout the text with emphasis on the Introduction (lines 104-107), the first paragraph of the Discussion (lines 774-780) and in the concluding paragraph (lines 986-989). Also, we go further in the discussion of this topic by providing practical situations in which our data shows that the lack of knowledge about the pipelines comparability can negatively impact the investigation of multi-national outbreaks (lines 833-837). These results have already triggered relevant discussions at the European Food Safety Authority (EFSA) regarding the comparison of different cgMLST strategies and the need to revisit case definitions during international outbreak investigations (https://www.efsa.europa.eu/sites/default/files/2024-11/3rd-specific-meeting-on-wgs%2C-scientific-network-for-zoonoses-monitoring-data-minutes.pdf).

(10) This paper seems extremely over-written and could be cut by about one-third without affecting any of the key messages or data presentations.

**Response:** We agree that this is a very large paper. However, for each species, we have explored multiple pipelines (with different sets per species) and conducted multiple analyses: the evaluation of their individual behaviour (e.g., stability regions), congruence with ST/serotype, polyphyly signal analysis, congruence at all possible levels, in-depth outbreak-level clustering exercise, etc. Each species had to be presented separately, covering around 2000 words each (approx. short manuscript), which we consider to be reasonable taking into account such volume of data. In our view, the Introduction (750 words) and Discussion (2500 words) are also not too long considering the complexity of the study. A way to considerably reduce the size of the manuscript would be to move all the Methods section (approx. 3700 words) to the Supplementary Material. We leave this at the consideration of the Editor.

**Reviewer #2**

Importance of the study: The adoption of WGS for food and waterborne disease surveillance has happened at varying rates for countries in Europe (and around the world), leading to variability in the pipelines used for WGS processing and analysis to inform surveillance and outbreak activities. This variability presents a challenge for communicating or sharing results between countries, with particular relevance to multi-national outbreak scenarios. In the absence of adopting a unifying goldstandard approach across the multiple countries, methods for assessing the congruence of WGS results are instead necessary to ensure that genomic results are interpreted consistently for outbreak and surveillance activities.

**Noteworthy results:**

1. The authors present results from the concomitant comparison of several different WGS pipelines for typing important human pathogenic enteric bacteria. It is worth noting that the approach they used contains novel elements, such as the use of a 'congruence score' which combines existing well-established metrics for measuring the concordance of typing methods (e.g., Adjusted Wallace and Adjusted Rand).
2. A major strength of the study is the breadth of data and partnerships involved. The comparison study combines the data and results from pipelines from 11 countries across western Europe. This represents a significant amount of work to collect and reanalyze datasets, and represents a very 'real' scenario as opposed to a more theoretical research study.
3. Overall, the authors report that 'like' agrees with 'like'; in other words, allele pipelines concord well with one another, as do SNP pipelines. Allele pipelines do not always agree with results from SNP pipelines, however.
4. Some pathogen specific differences were noted, especially among C. jejuni, which lacked regions of stability in the typing results when compared to the other pathogens under assessment, likely due to a less-stable genome. Genome level typing for C. jejuni also did not have high concordance with clonal-cluster results; other pathogens showed good concordance between CC typing and genome level results.

**Response:** We thank the reviewer for the positive feedback regarding our study and for acknowledging the significant logistics and computational resources involved.

**Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.**

1. The work presented here attempts to tackle a difficult and complex challenge; namely, how to integrate surveillance and outbreak results from different public health partners consistently when each partner uses a related, but different, toolset.
2. I believe this is the first study to attempt to align results from so many different partners, and test pipelines in an all-vs-all approach, rather than simplifying the analyses to SNP vs allele based, for example. The creation of a 'congruence Unclassified / Non classifié score' that combines the values from Adjusted Wallace and Adjusted Rand is a novel approach of summarizing agreement between methods, and works well at high values, but can obfuscate results at values less than 3, where one cannot ascertain why the congruence fails without knowledge on the directionality of the score. E.g., ST -> CC should have a very good AWAB and a low AWBA due to the directionality and differences in resolution between the methods.
3. Historically, many studies have sought to assess concordance between bacterial typing methods (e.g., PFGE vs MLST). With the adaptation of WGS for bacterial strain typing and incredibly high resolution methods for outbreak detection and surveillance, congruence between pipelines has not been a popular topic. I presume because at such high resolutions that WGS offers, all interpretations are considered to be correct?

**Response:** We again thank the reviewer for the positive comments. We agree that the novel method works well at high values and naturally loses the directionality of the individual AWC. Given the goal of the study, these were consequences that we intentionally tolerated because, for instance, we also accompanied the interpretation of the score with heatmaps (whose assessment of symmetry/asymmetry offers an intuitive way to infer the directionality of the resolution discrepancies) and all fine analysis (e.g., identification of "correspondence points") focused on scores near to 3 (i.e. where AWC is high in both directions), as mentioned by the reviewer. We also agree that the assessment of congruence between pipelines has not been a popular topic and with the reason pointed out by the reviewer. Other reasons might be the difficulty associated with the conduction of such comparisons (as demonstrated in our study) and the general perception that traditional EQA / proficient tests are enough to tackle this important need in the field. However, as our study shows, this perception from the community needs to be revisited because such exercises, which are usually focused on outbreak cluster detection with a limited number of isolates, need to evolve to another scale, not only regarding the dataset size and diversity, but also the magnitude of congruence assessment (as addressed in this study). Moreover, as our data shows that the lack of knowledge about the pipelines comparability can negatively impact the real investigation of multi-national outbreaks, congruence studies should be part of the routine of any efficient One Health FWD genomic surveillance system.

**Does the work support the conclusions and claims, or is additional evidence needed?**

1. The authors provide evidence to support their conclusions in the form of many comparisons of congruence between typing methods.

**Response:** Thank you for the positive appreciation.

**Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?**

1. A major focus of the manuscript is on the detection of outbreaks using WGS pipelines, and whether or not clustering results at the outbreak threshold are congruent.
2. One point to consider, as the authors mention in the discussion, is that outbreaks are not determined by genetic clustering alone, but rather by epidemiologic evidence supporting linked cases. Seeking congruence for isolates clustered at specific outbreak-related thresholds therefore does not achieve the result of determining whether pipelines are accurately detecting outbreak-related clusters. Rather, results generated by pipelines using outbreak thresholds should be compared to known outbreaks using epidemiologic data as the gold-standard.

**Response:** We totally agree, as suggested in the Discussion.

**Is the methodology sound? Does the work meet the expected standards in your field?**

1. The work presented in this manuscript, though verbose, appears sound. I am unable to properly provide a review of the Python code used for the manuscript, however, I can confirm that all scripts are present and available on the public Github.

**Response:** Indeed, all scripts are available.

**Is there enough detail provided in the methods for the work to be reproduced?**

1. Reproducibility of the study appears hampered by the number of different bioinformatic pipelines needing to be run on a rather large dataset. Scripts to facilitate the analysis are publicly available, and datasets have been deposited in appropriate repositories. To facilitate reproducibility, however, I believe the manuscript's aims and focus needs to be streamlined. There are far too many results and figures in the paper, hampering the clarity and conciseness of the approach and description of results.

**Response:** Indeed, due to the large amount of data, pipelines and personnel/laboratories involved in this study, its reproducibility would require massive resources, most likely not available outside a "consortium" context. Understanding this point, as the reviewer underlined, we have put big efforts to publicly deliver all the tools needed for other laboratories to conduct the same type of comparisons (e.g., to compare their pipelines with the same or other datasets), while being detailed in the description of the different scripts (https://github.com/insapathogenomics/WGS_cluster_congruence) (including their role in each type of comparison) and intermediate results (provided in 38 additional files). We understand that this is massive, but honestly do not see a more comprehensive way to present all the data at this stage. Still, as mentioned below, we leave a suggestion to shorten the paper at the consideration of the Editor.

**General Comments/Concerns:**

Generally, the manuscript is well written, with relatively few typos throughout. My biggest concern is that the manuscript length is far too long. At over 15000 words for just the Introduction, Results, Discussion, and Methods, I would suggest that the authors exert some effort to shorten this by focusing only on the main results of the study, or by splitting the manuscript into multiple shorter

papers focusing on 1-2 areas of investigation only. Primarily, the results and figures are far too extensive for a manuscript-style publication. The objective of the paper, as laid-out in the introduction, was to provide an assessment of congruence between typing methods, but the authors present several results for each pathogen assessing the congruence of WGS-based typing methods with results from traditional typing methods (e.g., MLST) including sequence-type and clonal complex. The length and number of experiments and results in this paper makes it very difficult to follow for this reviewer.

**Response:** We agree that this is a very large paper. While splitting it in several papers was an option we initially considered, we ultimately decided to combine all data in the same manuscript to provide a more comprehensive and integrative evidence of the congruence analysis (main goal). Indeed, the complementary investigation (congruence with ST/serotype, outbreak-level analysis, etc.) revealed to be critical to reach important contributions to the community, which were nicely summarized by the reviewer above. Given the multitude of pipelines (with different sets per species) and analyses conducted, each species had to be presented separately, covering around 2000 words each (approx. short manuscript), which we consider to be reasonable taking into account such volume of data. In our view, the Introduction (750 words) and Discussion (2500 words) are also not too long considering the dimension and complexity of the study. A way to considerably reduce the size of the manuscript would be to move all the Methods section (~3700 words) to the Supplementary Material. We leave this at the consideration of the Editor.

**Specific Comments:**

1. (Page 2) In the first sentence, I think the application of multiple methods raises questions about the compatibility of the results, but doesn't necessarily raises doubts about their accuracy.

**Response:** Yes, we agree with the reviewer. To clarify, the sentence highlights that the different pipelines might not be comparable and nothing is mentioned about their accuracy.

2. (Page 3) Please double check the accuracy of acronyms throughout: e.g., WOHA should be WOAH 3.

**Response:** Thank you for noting this typo, which was corrected.

3. (Page 3) "a significant effort is in place to develop and refine surveillance-oriented bioinformatics solutions for the assessment of samples' genomic relatedness and outbreak cluster identification" – Unless you're doing CIDT, it's the genomic relatedness of the isolate cultured from a sample, not the genomic relatedness of a sample itself.

**Response:** We have changed the term "sample" by "isolate" whenever adequate. For simplification, we maintained the generic term "sample" (as commonly used in bioinformatics) to refer to the "samples" reads/assemblies that compose the studied datasets.

4. (Page 5) I'm not sure what 'power and magnitude' are referring to here? Is this statistical power? If so, was this tested? I did not see any power calculations in the manuscript.

**Response:** We have now rephrased the sentence to remove the word "power" (used here figuratively).

5. (Page 5) Im unclear why both HC and MST were tested. The objective of the manuscript was to calculate the congruence between methods in use, not test every possible method.

**Response:** We understand the revierwers's point, but in Europe both methods are widely used. For instance, both public agencies acting in this field (ECDC and EFSA) incorporated different clustering methods in their respective platforms. Therefore, we have felt, as the results corroborate, that it would be important to cover both methods, whenever possible.

6. (Page 6) I'm unclear why this entire section on congruence with traditional typing is included in this manuscript when the objective was to test cluster congruence between methods in use in different countries. The same note applied to all four organisms included in the study.

**Response:** Please see the response to the general comment above. Specifically regarding the congruence with traditional typing, we believe its inclusion is very valuable for the study, leading to a better understanding not only of the overall congruence results, but also of the dataset and study outcomes. For instance: i) it showcases the existence of different levels of genetic heterogeneity among the most prevalent STs, CCs and/or serotypes, which is useful information for routine genomic surveillance and outbreak case definition (note that ST/serotypes are always part of communication during outbreak investigations); and, ii) it provides backwards compatibility with pre-WGS typing Era, thus increasing the confidence of laboratories to pursue the technological transition to WGS surveillance.

7. (Page 6) There are too many references to supplemental tables and files throughout the text. It makes it very difficult to follow.

**Response:** We understand, but we really want to direct the readers to the specific and detailed supplementary information that supports the main results presented in the manuscript.

8. (Page 7) Note that figure S3 shows partitions per threshold, but does not indicate the extent of congruence of partition membership.

**Response:** We assume that the reviewer is referring to Figure 2C (because we do not have any Figure S3), which indeed depicts a first overall view of the number of partitions at each possible distance threshold level. The congruence on cluster composition at all levels is then the focus of Figure 3 (heatmaps, slope plots, etc).

9. (Page 8) Generally references should remain in the introduction, discussion, and methods. The results shouldn't contain references to other works, but should present the results generated by the paper.

**Response:** We agree and we tried to only include references in the Results section when strictly necessary to understand the employed strategy or to acknowledge the concordance with previous works.

10. (Note many of the above comments can be extrapolated to sections on E. coli, Salmonella, and Campylobacter).

**Response:** Ok.

11. (Page 26) "reduce the likelihood of missing isolates that go unreported as they fall outside the case definition threshold in a given national pipeline" – Even if isolates are Unclassified / Non classifié not captured within the threshold, they will still be reported as a reportable illness. During outbreak investigations, closely related isolates will generally still be reviewed as part of iterative threshold improvements as an investigation proceeds.

**Response:** To clarify, the sentence provides a practical example of the impact of our results on multi-country outbreak investigations. Indeed, based on recent ECDC-EFSA outbreak case definitions (e.g. references 38 to 41), countries are initially "invited/requested" to share WGS data of isolates that cluster within a static threshold from representative outbreak reference strain(s) (shared by other country or ECDC/EFSA). After this "isolate" collection, data is then subjected to a centralized analysis (typically at ECDC/EFSA), where other thresholds can indeed be tested as an investigation proceeds (and depending on epi data), as the reviewer mentioned. Still, based on our results, it is likely that a given isolate that falls outside the initial case definition threshold (e.g., 4 ADs) in a national pipeline goes unreported to the ECDC/EFSA, although it could thereafter cluster within the same defined threshold in the ECDC/EFSA centralized analyses. The threshold flexibilization we suggest during the analysis with national pipelines would reduce this possibility. This precise topic has already triggered recent discussions at the European Food Safety Authority (EFSA) regarding the need to revisit case definitions during international outbreak investigations (https://www.efsa.europa.eu/sites/default/files/2024-11/3rd-specific-meeting-on-wgs%2C-scientific-network-for-zoonoses-monitoring-data-minutes.pdf).

12. (Page 27) I do not think that the congruence assessment of WGS based methods with either ST or CC is novel, nor does it add anything to the field.

**Response:** We would like to clarify that traditional typing nomenclature is often, if not always, used for communication during FWD outbreak investigation (even by European agencies) and several laboratories have not yet transitioned to a "WGS only" surveillance and still communicate with traditional typing classifications. Therefore, in our view, a good and deep knowledge of how traditional typing relates to cgMLST clustering is an important topic to the community. As the reviewer mentioned, indeed, previous studies have already explored this topic, but, to the best of our knowledge, never with this insight.

13. (Page 27) This flexible threshold approach I think is already in effect. Rather than determining the threshold flexibility a-priori, however, a threshold range is adapted to best suit the

epidemiological information derived through case interview and other evidence generated through the course of the outbreak investigation.

**Response:** As discussed in the manuscript (lines 839-845), we completely agree that the threshold definition should be contingent on several factors and thus, as seen in various practical examples, it should be adapted according to the epidemiological and microbiological context. The flexibilization concept that we suggest here regards to the WGS-based cluster definition that triggers the "call for genomic data" posed by centralizing entities during outbreak investigations (please see our answer to comment 11).

14. (Page 28) I do not believe that SNP-based methods are dismissed for application to outbreak investigations. Rather, the augmentation of SNP based methods to allele methods are recommended (e.g., by Pulse Net) for closely related or highly clonal bacteria

**Response:** In this sentence, we precisely intended to say that SNP analyses are intuitively a good approach for outbreak investigation that cannot be dismissed. We have now rephrased the sentence trying to make it more clear (lines 880-885).

15. (Page 28) Given the results of this paper, why do the authors not provide recommendations on which schema perform best? I.e., if low-number cgMLST schema for Campylobacter perform poorly, why not recommend a higher resolution schema be used?

**Response:** We understand the reviewer's suggestions. However, we believe that the results and scientific evidence produced by our study (conducted within a research consortium) should be presented as they are (i.e., as findings to be interpreted by readers, like the reviewer did regarding the *C. jejuni* cgMLST schemas) rather than formulated as recommendations, which are typically framed in time and issued by designated competent or advisory bodies.

16. (Page 28) I do not believe methods like ML-based phylogenies are appropriate or useful for short-term local outbreak investigations anyways. Rather, they are typically used for research studies assessing evolution of genes/organisms.

**Response:** We agree with the reviewer that ML methods are typically used (and more oriented) to assess evolutionary history. Still, they are also frequently applied for the detection (and not solely for evolutionary investigation) of outbreaks (e.g. of recent literature includes 10.1038/s41598-020-78808-y or 10.4315/JFP-20-188). As such, while our study covered the most widely used methods for this purpose, we believe it is relevant - and not problematic - to acknowledge that other approaches, like ML, were not explored in this study.

17. (Page 32) I do not see the acronym "QC" described, am assuming that it refers to 'quality control'

**Response:** Thank you for noticing this typo. The acronym is now described.

18. (Page 32) Are the dates and sources or other areas for bias related to the collection of bacterial isolates described somewhere?

**Response:** Yes, data such as year, country, and other metadata were compiled for all species within the BeONE dataset and are available in the respective Zenodo repositories. As the primary focus of the total dataset compilation was on genetic diversity, such information was not aggregated for the public reads.

19. (Page 54/55) Table 1 and Figure 1: I think it would be beneficial to include a table or chart of which pipelines are in use for which organisms.

**Response:** We have now improved Table 1 to include a footnote that facilitates the rapid identification of the pipelines that were used for each species.

20. (Page 56) Figure 2: far too many sub-panels. It makes readability very poor. In particular Figure D and E are very difficult to read or interpret. Im unsure how to interpret the bottom of panel D.

**Response:** Regarding the bottom part of panel D, it indicates the AD threshold at which each pipeline clusters together all samples belonging to a given ST. This visualization allows readers not only to compare the clustering similarity of the pipelines per ST, but also to identify STs with high genetic homogeneity (i.e., with low AD values, as, for example, ST8) or potentially polyphyletic (e.g., ST7 and ST325). As discussed, this information was highly valuable to interpret the next congruence results, while informing future developments in the field, such as the need of refining case definitions or establishing ST- or serotype-specific thresholds. We understand that the whole figure may appear large and complex. However, we really feel it is important to present a first main figure summarising the dataset characteristics, an individual assessment of each pipeline, and their relationship with traditional typing methods before going in-depth into the core topics of congruence analysis and outbreak detection. Also, we aimed to maintain a consistent style and content across the figures for the different species to facilitate readability (e.g., if readers want to identify the regions of highest congruence with ST across species, they can easily compare the corresponding figures for each species). Although we acknowledge the reviewer's suggestion, we consider that removing specific panels or changing this rationale would likely increase entropy rather than clarity.

21. (Page 60) Figure 4: In general, a box-and-whisker plot would be a better visualization than the ridgeline plot in panel A. Again there are too many subpanel, making readability and interpretability poor.

**Response:** We appreciate the suggestion. Still, we believe that the density plots also fit here, being appropriate to provide a clearer view of the overall AD distribution and making it easier to spot some patterns, such as the peaks. The density plot also allows for better comparison between pipelines by showing subtle differences that box-and-whisker plots might miss. We would like to keep the density plots.

22. (Page 62) Figure 5: Same comments as comment 20.

**Response:** Please see response above.

23. Overall, too many figures. Please reduce to the minimal necessary to display the main results of the study.

**Response:** Please see responses above.

**Reviewer #2 (Remarks on code availability):**

I am only reporting here to say that I did verify the existence of the code available in Github. Multiple scripts are available and well-organized. I cannot verify the python code itself, however, as I do not have expertise in this language.

**Response:** We thank the reviewer for checking that all code is available.

**Reviewer #3**

The authors present an extensive comparative study of bioinformatics pipelines for genomic surveillance of Listeria monocytogenes, Salmonella enterica, Escherichia coli and Campylobacter jejuni, with the aim of comparing the resolution power and the output in terms of identification of clusters. Such studies are pivotal for translating into practice the wealth of knowledge in bacterial genomic analysis which has been produced in the last ten years, and are still widely lacking in the literature, despite the wide amount of genomic sequences produced and published.
The study was conducted in the framework of a European research project, BeONE, taking advantage of the consortium ability and resources to conduct parallel analysis of the same wide set of sequences by using different approaches, e.g. cgMLST, SNPs analysis and wgMLST, and different tools, which have been extensively compared in this study.
The results are very interesting and useful for surveillance, and generally well presented. However, there are some minor points that I would suggest to revise for improving the readability of the text.

General comments:

The dataset of sequences used is very wide. In detail, L. monocytogenes dataset is representative of more than 40 STs, with the isolates well dispersed across the different STs (fig. 2A). The dataset for S. enterica is also wide and dispersed across STs and serotypes, with STs 11, 34 and 19 covering the majority of the strains, reflecting the abundance of isolation and sequencing of the strains most often associated to disease (fig 5A). The dataset of C. jejuni is also wide and well dispersed across the STs, well capturing the variability of the species. E. coli dataset is wide, even if smaller than the others, and mainly represented by ST11 strains, mainly belonging to O157 serogroup (856/2300 strains) (Fig 8A). The authors state that this bias reflects the wider abundance of this serogroup in public databases (lines 486-488). Nevertheless, in Europe O157 is not the most common serogroup isolated from severe disease any longer: “The European Union One Health 2022 Zoonoses Report” describes that the most common serogroup among Haemolytic Uremic Syndrome cases in 2022 was O26 (51.3%) followed by

O157 (14.5%). The variability of genomic stability across different E. coli serogroups is thus not represented in the presented analysis and the evaluations made on the thresholds considered in the paper could not be valuable for the other serogroups of E. coli. It would have been worth including more O26 sequences in the analysis, which would have had a big added value for interpretation of results during the real outbreak investigations which currently mainly involve O26 strains in Europe. If this is not possible at this stage, the authors should refer to this problem and make this bias clear in the paper.

**Response:** We thank the reviewer for the positive feedback regarding our work and all the constructive comments made. We would like to clarify that we compiled the datasets of this study in the beginning of the project, and tried to complement the WGS data shared within the BeONE project with the genetic diversity present in public databases at the time, which we accomplished as showcased in Figure 5A for *E. coli*. As the reviewer mentions, O26 serotype is now the most frequent Serotype linked to Haemolytic Uremic Syndrome cases in Europe. However, O157 is still the most prevalent one and, perhaps for this reason, the O26 is not yet predominant in public sequence databases (e.g., currently, O26 is the sixth most frequent Serotype in Enterobase, while O157 remains the first). As such, in our view, the dataset is still a good proxy of serotype diversity in public databases and reflects the contemporary scenario of O157 being the most prevalent serotype among reported human cases. More importantly, we hope the reviewer understands that it is impossible for us to add more samples at this stage and have the whole dataset run again by all partners, as the consortium is now closed. For this reason, and as we fully agree with the relevance of having follow-up studies focused on assessing pipelines performance and congruence in detecting O26 outbreaks, we now refer to this subject in the Discussion of the manuscript (lines 873-876).

One of the main important and useful results of this study is the identification of different ranges of allelic distances for different STs of the same species. This is a very useful information for deciding the "outbreak thresholds" within a species and, as stated at lines 861-863, "it underscores the potential inadequacy of applying a single "outbreak threshold" within a species, as this may lead to an oversizing of potential outbreak clusters when highly clonal WGS derived lineages (often overlapping with STs/serotypes) are involved". An effort should be made to present such result of variability of thresholds in one table or one graph including all the species and main STs tested, to be included in the manuscript itself (not in additional files). Similarly, it would be helpful to summarize in a table the results of the concordance for outbreak detection obtained for all the species (e.g. for each species, percentage of the clusters detected by all the pipelines by using a static threshold, with exceptions, and difference between the AD thresholds required by the different allele-based pipelines to detect each "outbreak-level" cluster).

**Response:** We apologize but we did not exactly understand what the reviewer is suggesting. Regarding the identification of different ranges of allelic distances for different STs of the same species, we provide detailed tables (Additional files 3, 11, 20 and 29, linked to panel D of Figures 2, 5, 8 and 11) with the lowest threshold levels at which all samples of the same ST/CC/serotype cluster together in each pipeline. As the reviewer underlined, we agree that this information, together with other epidemiological and microbiological information (discussed in lines 839-845), is relevant for future outbreak-specific case definitions. Still, as also mentioned (lines 869-878), we believe that "This topic warrants further research in order to determine which threshold ranges render the highest

epidemiological concordance per genogroup (... and ...) The integration of epidemiological data in large-scale WGS studies, for example, through the analysis of epidemiologically confirmed outbreaks, will be also crucial to tackle this subject.". So, if the table the reviewer is suggesting refers to the data used for the "polyphyly signal analysis", then we consider that all data is well compiled in the above mentioned additional files, being impractical (and conflicting with the paper structure) to compile all species in a single file. In contrary, if the reviewer refers to the presentation of ranges of thresholds to be used as "reference" for outbreak detection involving specific ST/CC/Serotypes, we believe that it would be inadequate to make such deterministic inferences dissociated from epidemiological data, and that, instead, a follow-up fine-tuned analysis exclusively focused on epidemiologically verified outbreaks would be the most reasonable approach. Regarding the request for summary tables with the results of the concordance for outbreak detection for all the species, we apologize, but once again we are not sure if we fully understand the request because we present the number of the clusters detected by all pipelines, by at least two pipelines and exclusive of each pipeline in Additional files 7, 15, 24 and 33. Also, we provide the distance threshold at which each cluster was detected by a given pipeline in Additional Files 8, 16, 25 and 34. We hope this is what the reviewer meant.

Specific comments:

The text is very long and sometimes redundant. For example, the text presented in Additional file 1 is repeated in several sections along the manuscript (e.g. lines 140-146 or 1178-1185). The manuscript could contain less details than the additional file, or alternatively explain the same concept in easier words, citing additional file 1 for more details.

**Response:** We understand that the reviewer may have found some redundancy both between species (A) and between the main body and some additional files (B). Both were intentional. Regarding A, we aimed to maintain a consistent style and content across the text and figures for the different species to facilitate readability (e.g., if readers want to identify the regions of highest congruence with ST across species, they can easily compare the corresponding text and figures in each species-specific section). Regarding B, there is, indeed, some redundancy in the phrasing (but not in the content), which was an option we took to better assist readers in linking the main summarized results to the detailed data. The only exception is indeed the Additional file 1, where the introductory paragraphs (regarding the species studied and dataset composition) are an extended version of the small text presented in the manuscript. However, we think that this contextualization (dataset, method rationale and introduction of clustering methods) at the start of Results is critical before starting with the species-specific sections. We hope the reviewer understands our point of view. Regarding the overall size of the article, we agree that this is a very large paper. However, given the multitude of pipelines (with different sets per species) and analyses conducted, each species had to be presented separately, covering around 2000 words each (approx. short manuscript), which we consider to be reasonable taking into account such volume of data. The Introduction (750 words) and Discussion (2500 words) are also not too long considering the dimension and complexity of the study. A way to considerably reduce the size of the manuscript would be to move all the Methods section (approx. 3700 words) to the Supplementary Material. We leave this at the consideration of the Editor.

Figures 2C, 5C, 8C and 11C (graphs of block of stability regions): this analysis is not well explained in any section of the manuscript. Which is the parameter on y axis for each pipeline?

**Response:** The details behind this analysis are described in lines 1172-1180. Regarding the panel C of the figures, we have now clarified in the legend: "Clustering stability regions determined for each pipeline. To better distinguish each region (represented by separated rectangle blocks), different blocks are vertically phased, starting in a different line".

Line 184: which table in additional file 4 is cited here? S1.1? Please specify

**Response:** It is indeed Table S1.1. Thank you for noticing this typo.

Line 329: please better explain: which type of clusters (which AD)?

**Response:** It is at the highest congruence point, as for the other sentences in the paragraph. We have now improved the text for clarity.

Line 330: please verify consistency with fig. 5D: should Thompson or Typhi be cited here?

**Response:** Thank you for noticing this typo. The correct serotype is Thompson as in the text and aligned with the literature. The error was in the figure legend and caused by both serotypes having the same exact number of samples. This was now corrected.

Lines 504-508: please better explain the PopPUNK analysis, as the aim is not clear in the text

**Response:** In this section of Results, we aimed at identifying the threshold levels (and stability regions) with the highest congruence with traditional typing (ST, CC or serotype) and, when possible, with WGS-derived pathogen main lineages, as detailed in Additional file 1. The latter were inferred using PopPunk, which was indeed not clear in this text. We have now improved the phrasing in the Additional file 1.

Line 604: "later" should be corrected in "latter"

**Response:** Done.

Lines 802-804: the use of a dynamic wgMLST is not easy to apply as a standardised method, for example to compare sequences across Member States at the European level. The use of larger schemas of cgMLST was proved effective also for C. jejuni. Why are you suggesting the use of wgMLST approach instead of extended schemas of cgMLST? Please better explain the possible application for the suggested approach for surveillance or consider revising.

**Response:** To clarify, what we suggest (and did in this study) is the usage of a dynamic wgMLST approach for each cgMLST cluster of interest at outbreak level (identified with a static cgMLST schema) in order to gain resolution and better discriminate outbreak isolates (i.e. initial cgMLST with a static schema followed by cluster-specific wgMLST analysis, where accessory loci are dynamically added). As this was not clear in the lines mentioned by the reviewer, we have now clarified the text (lines 804-806).

Lines 1002-1004: how were the reads tested for this analysis?

**Response:** This analysis was based on the Kraken2 results obtained with the AQUAMIS pipeline. We have now clarified this in the new version of the manuscript.

**RESPONSE TO REVIEWERS**

**Reviewer #1** (Remarks to the Author):

I was recruited to assess whether the authors have satisfactorily addressed or rebutted Reviewer 1's comments

I do want to point out that for a multicenter comparison of computational pipelines, it is important to ensure and report if the same computational environment has been used across centers. It's well known that the same pipeline deployed in different environments can produce different results (https://www.nature.com/articles/nbt.3820, https://pmc.ncbi.nlm.nih.gov/articles/PMC3842296/). Was there any effort to ensure the reproducibility of the analyses?

**Response:** We thank the reviewer for accepting revising our responses to the previous Reviewer #1. Regarding this initial comment, we understand the Reviewer's point, which makes total sense in a study aiming to perform a technical comparison (e.g. memory, time, etc.) of different pipelines. Still, the main goal of this study was to assess the comparability of the clustering outputs (and not time, computational cost, etc.) of the pipelines in place in different laboratories mimicking a real life case scenario. Therefore, each pipeline had to be run with the computational environment that the respective laboratory uses in its routine surveillance activities. Said that, all efforts were done to ensure the reproducibility of the analyses, as demonstrated by the exhaustive description of all parameters, the amount of supplementary and source data provided (including the clustering outputs of each pipeline), and the total availability of the developed code.

Comment 1. I agree with Reviewer 1 that not evaluating computational pipelines that major agencies outside of Europe have widely used is a limitation, especially since such agencies have generated or curated the majority of the publicly available genome data of common foodborne pathogens. This limitation should be acknowledged and justified in the manuscript. I don't think the authors have adequately addressed this limitation by just citing NCBI Pathogen Detection and the CFSAN pipeline.

**Response:** We have now stressed that the study is limited to a large European consortium including key institutes involved in the surveillance of foodborne pathogens (Lines 997-1000).

Comment 2. To my knowledge, NCBI Pathogen Detection's SNP pipeline is not publicly available for local implementation and evaluation. By citing it, the authors adequately addressed this comment.

**Response:** Thank you.

Comment 3. The response is adequate.

**Response:** Thank you.

Comment 4. If the CFSAN pipeline is referenced and discussed, so should be CDC's pipeline https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.00375/full, especially since this is one of first and most cited papers that compares SNP and cgMLST.

**Response:** Thank you for suggesting this reference, which was now cited (reference number 17) in several key points in the Introduction and in the Discussion.

Comment 5. The response is adequate.

**Response:** Thank you.

Comment 6. The response is adequate.

**Response:** Thank you.

Comment 7. The response is adequate.

**Response:** Thank you.

Comment 8. The response is adequate but a bit blunt. Literature outside of Europe, such as https://academic.oup.com/cid/article/63/3/380/2595027 can be referenced.

**Response:** Thank you. We have now added this reference and others on the same subject (reference numbers 34,84,89,90).

Comment 9. The response itself makes sense given the logistic challenges of harmonization and standardization, but it is not adequate because it ignores Reviewer 1's comment to include a discussion on this topic in the manuscript, which I find reasonable.

**Response:** We have now reshaped the first paragraph of the Discussion to provide a better discussion on this topic (Lines 783-799).

Comment 10. The response is adequate.

**Response:** Thank you.

**Reviewer #2** (Remarks to the Author):

General comments:

I thank the authors for their attention in responding to comments provided in the previous review. While I still believe that the manuscript is overly long, and shortening the manuscript would improve its clarity and readability, I do respect the amount of work being presented, and understand the authors' preference to keep the analyses together rather than split the paper into multiple reports. Overall, the work presented in the manuscript appears sound and the authors have addressed major concerns raised in the first review. I do think that moving the extensive methods section to the end of the manuscript improved readability. At this time, I do not have any additional major concerns or comments related to the manuscript beyond those that were already provided.

The authors have provided a comprehensive comparison of clustering results derived from multiple bioinformatic pipelines in use by the European public health surveillance community for different bacterial species. While the work here may not provide the answers to questions like: 'which pipeline is best', their exploration of differences in clustering resulting from the different pipelines are an important addition to the field, and should be highly useful to members of the community practicing in the space of WGS-based infectious disease surveillance and outbreak detection. As the authors have noted several times, it is unlikely that we will ever see a single standardized approach put to use across the surveillance community. It is therefore important that work like this is published so that the community is aware of the differences that result from the use of different pipelines, and can make informed decisions on approaches used for national surveillance.

**Response:** We thank the reviewer for the positive feedback regarding our work.

I do have a couple very minor suggestions:
• Food and waterborne diseases is first defined as "FWDs" but is written "FWD" in several locations in the introduction (e.g., L122, L125) and Discussion.

**Response:** Thank you for pointing this out. We have now defined food and waterborne diseases as FWD and make sure that this is consistent throughout the revised manuscript.

• Thank you for providing information about specific pipelines in use for different bacterial species in Table 1. If I may make a small suggestion, please define the abbreviations Lm, Se, Ec, and Cj somewhere in the table description or in the footer

**Response:** Thank you for pointing this out. We have now replaced these abbreviations in the Table by the respective species name.

**Reviewer #3** (Remarks to the Author):

The Authors revised the paper according to the reviewers' comments raised in the first round of the revision process. Many of the raised points were addressed and the text has gained readability. The text has not been shortened, as requested by all the three reviewers, and it remains very long. Nevertheless, the topic is very difficult to be shortened and relevant to be published, so I think it could be published in the current form, to keep all the relevant information in the main body of the manuscript. I do not recommend to move the Materials and Methods section in the supplementary material, as proposed by the authors, as they are an important part of the work and ensure a full understanding of the study.

**Response:** We thank the reviewer for the feedback. As there was no Editorial request regarding the Methods section, as the reviewer suggests, we will maintain this section in the main body of the manuscript.

Round 1 Reviewer 2 attachment:
Importance of the study:

The adoption of WGS for food and waterborne disease surveillance has happened at varying rates for countries in Europe (and around the world), leading to variability in the pipelines used for WGS processing and analysis to inform surveillance and outbreak activities. This variability presents a challenge for communicating or sharing results between countries, with particular relevance to multi-national outbreak scenarios. In the absence of adopting a unifying gold-standard approach across the multiple countries, methods for assessing the congruence of WGS results are instead necessary to ensure that genomic results are interpreted consistently for outbreak and surveillance activities.

Noteworthy results

- The authors present results from the concomitant comparison of several different WGS pipelines for typing important human pathogenic enteric bacteria. It is worth noting that the approach they used contains novel elements, such as the use of a 'congruence score' which combines existing well-established metrics for measuring the concordance of typing methods (e.g., Adjusted Wallace and Adjusted Rand).
- A major strength of the study is the breadth of data and partnerships involved. The comparison study combines the data and results from pipelines from 11 countries across western Europe. This represents a significant amount of work to collect and reanalyze datasets, and represents a very 'real' scenario as opposed to a more theoretical research study.
- Overall, the authors report that 'like' agrees with 'like'; in other words, allele-pipelines concord well with one another, as do SNP pipelines. Allele pipelines do not always agree with results from SNP pipelines, however.
- Some pathogen specific differences were noted, especially among *C. jejuni*, which lacked regions of stability in the typing results when compared to the other pathogens under assessment, likely due to a less-stable genome. Genome level typing for C. jejuni also did not have high concordance with clonal-cluster results; other pathogens showed good concordance between CC typing and genome-level results.

- Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.
    - The work presented here attempts to tackle a difficult and complex challenge; namely, how to integrate surveillance and outbreak results from different public health partners consistently when each partner uses a related, but different, toolset.
    - I believe this is the first study to attempt to align results from so many different partners, and test pipelines in an all-vs-all approach, rather than simplifying the analyses to SNP vs allele based, for example. The creation of a 'congruence

score' that combines the values from Adjusted Wallace and Adjusted Rand is a novel approach of summarizing agreement between methods, and works well at high values, but can obfuscate results at values less than 3, where one cannot ascertain why the congruence fails without knowledge on the directionality of the score. E.g., ST -> CC should have a very good $AW_{A \rightarrow B}$ and a low $AW_{B \rightarrow A}$ due to the directionality and differences in resolution between the methods.

    - Historically, many studies have sought to assess concordance between bacterial typing methods (e.g., PFGE vs MLST). With the adaptation of WGS for bacterial strain typing and incredibly high resolution methods for outbreak detection and surveillance, congruence between pipelines has not been a popular topic. I presume because at such high resolutions that WGS offers, all interpretations are considered to be correct?

- Does the work support the conclusions and claims, or is additional evidence needed?
    - The authors provide evidence to support their conclusions in the form of many comparisons of congruence between typing methods.
- Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?
    - A major focus of the manuscript is on the detection of outbreaks using WGS pipelines, and whether or not clustering results at the outbreak threshold are congruent.
    - One point to consider, as the authors mention in the discussion, is that outbreaks are not determined by genetic clustering alone, but rather by epidemiologic evidence supporting linked cases. Seeking congruence for isolates clustered at specific outbreak-related thresholds therefore does not achieve the result of determining whether pipelines are accurately detecting outbreak-related clusters. Rather, results generated by pipelines using outbreak thresholds should be compared to known outbreaks using epidemiologic data as the gold-standard.
- Is the methodology sound? Does the work meet the expected standards in your field?
    - The work presented in this manuscript, though verbose, appears sound. I am unable to properly provide a review of the Python code used for the manuscript, however, I can confirm that all scripts are present and available on the public Github.
- Is there enough detail provided in the methods for the work to be reproduced?
    - Reproducibility of the study appears hampered by the number of different bioinformatic pipelines needing to be run on a rather large dataset. Scripts to facilitate the analysis are publicly available, and datasets have been deposited in appropriate repositories. To facilitate reproducibility, however, I believe the manuscript's aims and focus needs to be streamlined. There are far too many results and figures in the paper, hampering the clarity and conciseness of the approach and description of results.

General Comments/Concerns:

Generally, the manuscript is well written, with relatively few typos throughout. My biggest concern is that the manuscript length is far too long. At over 15000 words for just the Introduction, Results, Discussion, and Methods, I would suggest that the authors exert some effort to shorten this by focusing only on the main results of the study, or by splitting the manuscript into multiple shorter papers focusing on 1-2 areas of investigation only.

Primarily, the results and figures are far too extensive for a manuscript-style publication. The objective of the paper, as laid-out in the introduction, was to provide an assessment of congruence between typing methods, but the authors present several results for each pathogen assessing the congruence of WGS-based typing methods with results from traditional typing methods (e.g., MLST) including sequence-type and clonal complex. The length and number of experiments and results in this paper makes it very difficult to follow for this reviewer.

Specific Comments: (Apologies but the manuscript I received did not have line numbers included. I have included page numbers instead)

1. (Page 2) In the first sentence, I think the application of multiple methods raises questions about the compatibility of the results, but doesn't necessarily raises doubts about their accuracy.
2. (Page 3) Please double check the accuracy of acronyms throughout: e.g., WOHA should be WOAH
3. (Page 3) "a significant effort is in place to develop and refine surveillance-oriented bioinformatics solutions for the assessment of samples' genomic relatedness and outbreak cluster identification" – Unless you're doing CIDT, it's the genomic relatedness of the isolate cultured from a sample, not the genomic relatedness of a sample itself.
4. (Page 5) I'm not sure what 'power and magnitude' are referring to here? Is this statistical power? If so, was this tested? I did not see any power calculations in the manuscript.
5. (Page 5) Im unclear why both HC and MST were tested. The objective of the manuscript was to calculate the congruence between methods in use, not test every possible method.
6. (Page 6) I'm unclear why this entire section on congruence with traditional typing is included in this manuscript when the objective was to test cluster congruence between methods in use in different countries. The same note applied to all four organisms included in the study.
7. (Page 6) There are too many references to supplemental tables and files throughout the text. It makes it very difficult to follow.
8. (Page 7) Note that figure S3 shows partitions per threshold, but does not indicate the extent of congruence of partition membership.
9. (Page 8) Generally references should remain in the introduction, discussion, and methods. The results shouldn't contain references to other works, but should present the results generated by the paper.
10. (Note many of the above comments can be extrapolated to sections on E. coli, Salmonella, and Campylobacter).
11. (Page 26) "reduce the likelihood of missing isolates that go unreported as they fall outside the case definition threshold in a given national pipeline" – Even if isolates are

not captured within the threshold, they will still be reported as a reportable illness. During outbreak investigations, closely related isolates will generally still be reviewed as part of iterative threshold improvements as an investigation proceeds.

12. (Page 27) I do not think that the congruence assessment of WGS based methods with either ST or CC is novel, nor does it add anything to the field.
13. (Page 27) This flexible threshold approach I think is already in effect. Rather than determining the threshold flexibility a-priori, however, a threshold range is adapted to best suit the epidemiological information derived through case interview and other evidence generated through the course of the outbreak investigation.
14. (Page 28) I do not believe that SNP-based methods are dismissed for application to outbreak investigations. Rather, the augmentation of SNP based methods to allele-methods are recommended (e.g., by Pulse Net) for closely related or highly clonal bacteria
15. (Page 28) Given the results of this paper, why do the authors not provide recommendations on which schema perform best? I.e., if low-number cgMLST schema for Campylobacter perform poorly, why not recommend a higher resolution schema be used?
16. (Page 28) I do not believe methods like ML-based phylogenies are appropriate or useful for short-term local outbreak investigations anyways. Rather, they are typically used for research studies assessing evolution of genes/organisms.
17. (Page 32) I do not see the acronym “QC” described, am assuming that it refers to ‘quality control’
18. (Page 32) Are the dates and sources or other areas for bias related to the collection of bacterial isolates described somewhere?
19. (Page 54/55) Table 1 and Figure 1: I think it would be beneficial to include a table or chart of which pipelines are in use for which organisms.
20. (Page 56) Figure 2: far too many sub-panels. It makes readability very poor. In particular Figure D and E are very difficult to read or interpret. Im unsure how to interpret the bottom of panel D.
21. (Page 60) Figure 4: In general, a box-and-whisker plot would be a better visualization than the ridgeline plot in panel A. Again there are too many subpanel, making readability and interpretability poor.
22. (Page 62) Figure 5: Same comments as comment 20.
23. Overall, too many figures. Please reduce to the minimal necessary to display the main results of the study.