BMC Bioinformatics

RESEARCH

Open Access

Managing false positives during detection of pathogen sequences in shotgun metagenomics datasets

Lauren M. Bradford[1], Catherine Carrillo[3] and Alex Wong[1,2*]

*Correspondence:
alexwong@cunet.carleton.ca

[1] Department of Biology, Carleton University, Ottawa, Canada
[2] Institute for Advancing Health Through Agriculture, Texas A & M University, College Station, USA
[3] Ottawa Laboratory (Carling), Canadian Food Inspection Agency, Ottawa, Canada

Abstract

Background: Culture-independent diagnostic tests are gaining popularity as tools for detecting pathogens in food. Shotgun sequencing holds substantial promise for food testing as it provides abundant information on microbial communities, but the challenge is in analyzing large and complex sequencing datasets with a high degree of both sensitivity and specificity. Falsely classifying sequencing reads as originating from pathogens can lead to unnecessary food recalls or production shutdowns, while low sensitivity resulting in false negatives could lead to preventable illness.

Results: We used simulated and published shotgun sequencing datasets containing *Salmonella*-derived reads to explore the appearance and mitigation of false positive results using the popular taxonomic annotation softwares Kraken2 and Metaphlan4. Using default parameters, Kraken2 is sensitive but prone to false positives, while Metaphlan4 is more specific but unable to detect *Salmonella* at low abundance. We then developed a bioinformatic pipeline for identifying and removing reads falsely identified as *Salmonella* by Kraken2 while retaining high sensitivity. Carefully considering software parameters and database choices is essential to avoiding false positive sample calls. With well-chosen parameters plus additional steps to confirm the taxonomic origin of reads, it is possible to detect pathogens with very high specificity and sensitivity.

Keywords: Pathogen detection, Metagenomics, Shotgun sequencing, Salmonella

Background

Foodborne illnesses are a global public health issue, causing an estimated 600 million incidents of illness and 420 thousand deaths worldwide each year [1, 2]. In order to prevent consumers from becoming ill, it is essential to detect foodborne pathogens in the production chain.

Culture-based microbiological methods for pathogen detection, which rely on selective enrichment and isolation on agar plates, have been in use for more than a century [3]. Although these methods are sensitive, they are time- and labour- intensive and require labs staffed by expert personnel. In recent years, there has been increasing interest in using culture-independent diagnostic tests (CIDTs) for diagnosis and surveillance

of pathogenic organisms of concern. CIDTs include PCR-based methods as well as high-throughput sequencing of either marker genes (e.g 16 S rRNA or virulence-related genes) or metagenomes via shotgun sequencing [3–9].

Shotgun sequencing, wherein all DNA in a sample is sequenced, provides metagenomic data that can be used to detect the presence of pathogens. This type of sequencing avoids the amplification biases that plague phylogenetic metabarcoding [10] and produces datasets containing the full breadth of genetic material [11]. Accordingly, these datasets can also provide information on genes conferring virulence [12] and antimicrobial resistance [3, 13]. Metagenomic data can be used for serotyping of pathogens [14]. It may even be possible to produce metagenome-assembled genomes (MAGs) of pathogens for use in multi-locus sequence typing (MLST) and other analyses [3]. Furthermore, metagenomic datasets can be searched for multiple pathogens during diagnostics or for routine monitoring during food production, although culture enrichment, which requires prior knowledge of possible pathogens-of-interest, is usually still essential to detect organisms at low abundance [3].

While these factors make metagenomics via shotgun sequencing an enticing option for pathogen detection, there are downsides. The pure culture isolates produced by microbiological methods can be used for downstream analyses including drug-resistance phenotyping and whole-genome sequencing (WGS) for source attribution [3]; by definition, CIDTs bypass this step [15]. Furthermore, the depth of sequencing required and associated cost must be considered. Detecting low-abundance organisms in samples with overwhelming numbers of reads from the host, food matrix, and/or other microbes is a major barrier [16].

Trustworthy taxonomic classification of each sequencing read is an ongoing challenge, and many bioinformatic tools have been and continue to be developed to address this issue [17–23]. Metagenomic read classification algorithms primarily rely on identifying species by comparing them to the closest matches in existing databases. However, this approach poses challenges when dealing with species that have limited representation in public repositories, especially when compared to pathogenic species. Additionally, certain sequences exhibit high conservation between species, creating a risk of misclassifying non-pathogens as related pathogens.

Falsely identified reads (that is, sequencing reads erroneously classified as coming from the pathogen of interest) can lead to false positive calls of samples, which presents a particular problem in the field of pathogen detection. In the context of food production, these could cause economic loss from unnecessary recalls or factory shutdowns.

Various strategies have been proposed to eliminate false positives, such as setting a high threshold for the number of pathogen-derived reads required for a sample to be considered "positive" [24]; manually curating reference databases and using stringent software settings [25]; or confirming reads putatively classified as the pathogen-of-interest by comparison to species-specific regions (SSRs) [26]. Generally, a trade-off must be made between specificity (in this case, the reduction or elimination of false positives) and sensitivity (being able to detect pathogen-derived sequencing reads at very low relative abundance).

In this paper, we investigate trade-offs between sensitivity and specificity in metagenomic detection of pathogens, using *Salmonella* as a test case. Our goal is to identify the

effects of parameter choices and databases on this trade-off, since software defaults may not be ideal. We focus on two commonly used software tools, Kraken2 [27] and MetaPhlan 4 [21]; however, the lessons from our analysis are extendable beyond these specific tools. Kraken [27] and its updated version, Kraken 2 [27] use k-mer based methods and are among the most highly cited metagenomic classifiers. There are a range of premade reference databases available for Kraken 2, and it also allows the production of custom databases. With well-chosen databases and parameters, Kraken2 achieves high precision and recall [28]. However, the addition of a confirmation step using SSRs could be applied to the outputs of various classifiers. We compared the detection sensitivity of k-mer based Kraken 2, with and without an SSR-based confirmation step, against Metaphlan4 [21], which promises high specificity by mapping reads against a database of clade-specific marker genes.

Many other tools and pipelines are available for analysis of shotgun sequencing datasets, and benchmarking comparisons of all of them is beyond the scope of this project. However, we wish to emphasize the importance of testing a chosen tool with simulated datasets in which the origin of each read is known, before trusting the outputs of those tools used on real samples.

We use *Salmonella* as a model for the broader problem of pathogen detection in metagenomic datasets. Non-typhoidal serovars of *Salmonella*, which cause potentially life-threatening gastrointestinal illess, are one of the most common contributors to foodborne illness in Canada [29] and are one of the top 4 causes of diarrhoeal diseases globally [30]. However, the findings of this study could be adapted and extended to other pathogens.

Results

We tested reliability and lower limits of detection of simulated *Salmonella* reads in shotgun sequencing datasets. We started with simulated background communities of closely-related bacteria (i.e., members of the Enterobacteriaceae family), since the chance of false identification should be higher with more closely related organisms. We tested classification by Kraken 2 using various reference databases and confidence levels, as well as an additional confirmation step in which putative *Salmonella* reads were compared against "species"-specific reads from the *Salmonella* pan-genome [31]. To compare the detection sensitivity against other reference-based classification software, we also investigated these simulated libraries using the recently-released Metaphlan4 [21].

Choice of confidence level and database affects number of false positives

We first examined the impact of confidence level. Confidence scoring in Kraken 2 is a simple scheme in which the user defines a score threshold between 0 and 1 (default: 0). Each sequence is scored based on *kmer* mapping, and the label for that sequence is adjusted until the score meets or exceeds the confidence threshold. A more detailed explanation can be found in the software manual [32]. At confidence 0, the default setting, the majority of *Salmonella*-derived reads are correctly assigned, but there are many false positives (Fig. 1).

As confidence increases, the number of *Salmonella*-derived reads identified higher on the taxonomic tree increases; for example, *Salmonella*-derived reads could be classified

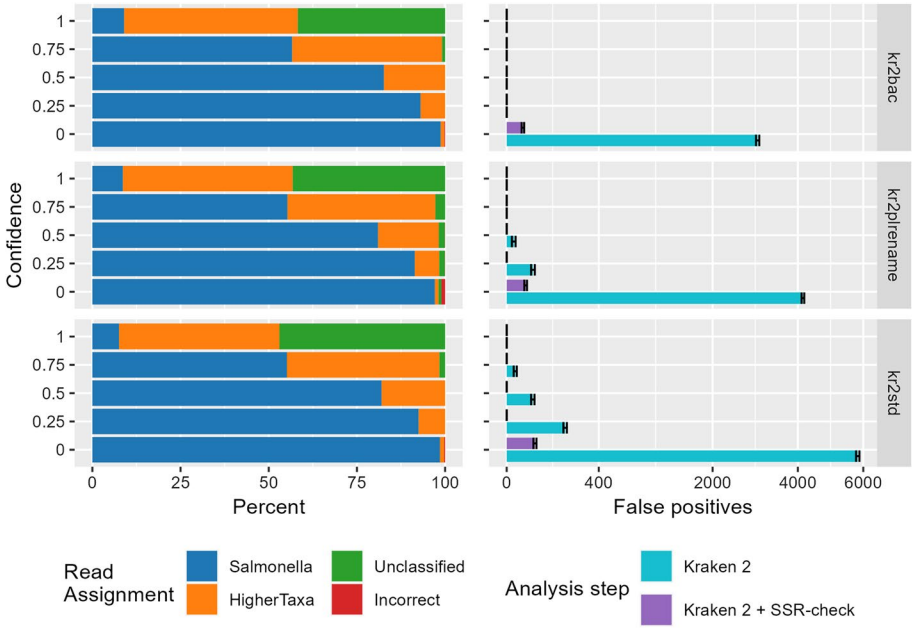

Fig. 1 Left panel: Stacked bars showing Kraken 2's classification of *Salmonella*-derived reads in the library with 0.001% *Salmonella*. In blue, *Salmonella*-derived reads identified explicitly as from the *Salmonella* genus; orange, those identified at a less specific taxonomic level; green, unclassified; and red misidentified as neither *Salmonella* nor an appropriate higher taxonomic group. Right panel: Number of non-*Salmonella*-derived reads classified as *Salmonella* (i.e. false positives) by Kraken 2, and remaining after checking Kraken 2 results against SSRs. Libraries contained 10 million reads each. Error bars are 1 std deviation. X-axis is square-root transformed to better display low values

as "enterobaceriaceae" or "gammaproteobacteria", or simply as "cellular organism". While these identities are not *incorrect*, they would not lead their libraries to be considered "positive for *Salmonella*". *Salmonella* reads which were falsely identified were almost all identified as other members of the Gammaproteobacteria, and mostly as closely related genera such as *Escherichia*, *Shigella*, and *Citrobacter*.

The prevalence of false positives at differing confidence levels can also be seen by their effect on precision in precision-recall curves (Fig. 2). This is most readily apparent in libaries with low counts of *Salmonella*-derived reads (bottom panel). Precision is very low at confidence 0 and near perfect at confidence 1, regardless of database used. However, database choice impacts precision and recall at intermediate confidence levels, with the kr2bac database showing near-perfect precision and high recall already at confidence 0.25 (Fig. 2, bottom panel).

Comparison to SSRs is quite effective at removing false positives

To remove false positives while retaining the best chance of detecting true positives, we added a comparison step analogous to that used in the SNIPE pipeline [26]. All reads identified by Kraken 2 as belonging to the *Salmonella* genus were then compared to 403 "species"-specific regions (SSR; though in this case they are genus- rather than species-specific) of 1000 bp length each from the *Salmonella* pan-genome. These SSRs were previously found by Laing et al. [31] by extracting 1000 bp-long regions shared by 211 closed *S. enterica* genomes, iteratively screening

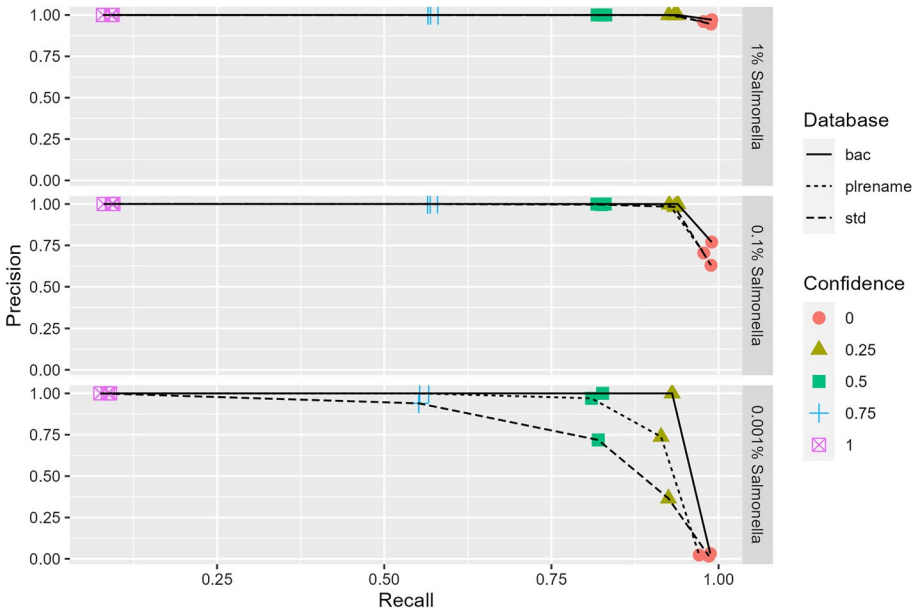

Fig. 2 Precision-recall plots for *Salmonella* detection via Kraken 2 classification in 10 million read libraries containing 100k (top panel), 10k (middle panel), and 100 (bottom panel) *Salmonella*-derived reads. Precision is a measure of specificity, with high precision indicating a low rate of false positives; recall is a measure of sensitivity, with high recall indicating a low rate of false negatives

these regions against the GenBank nr database, and discarding any region present in any genomic sequence except that of *S. enterica*.

This comparison substantially reduced the number of false positives remaining at the end of the analysis pipeline. For all three databases, however, false positives remained at confidence 0 (the Kraken 2 default) and were only completely absent at confidence ≥ 0.25 (Fig. 1, right panel).

Reads from novel organisms that are related to *Salmonella* are also filtered

We have previously collected genome sequence data for unusual isolates recovered from food and environmental sources, ten of which were mis-identified as *Salmonella* based on closest matches to published genomes by either MASH (1 *Citrobacter* spp.) [33], 16 S sequence analysis (6 *Enterobacter/Klebsiella* spp.) or detection of species-specific genes (3 *Citrobacter* spp.). These genomes have not been published and are therefore not incorporated into public databases. To test whether sequencing reads from these organisms pose a problem for the present workflow, a metagenome was created by simulating reads from 31 unpublished genomes to coverage levels of 35 to 82x (See Table S2, Additional file 3). While Kraken 2 analysis on its own classified 16,904 of the 40 million total reads as coming from *Salmonella*, none of these reads passed through the SSR-check step. Sequencing files are available at 10.5281/zenodo.8056523; given their usefulness as potential false positives, we suggest it is best to keep these sequences out of standard databases.

Limits of detection in a background of related species

A subsequent round of analysis investigated limits of detection using libraries with lower *Salmonella* content and the analysis parameters with the best precision and recall characteristics, i.e., the Kraken 2 bacteria database and confidence 0.25, with subsequent confirmation of putative *Salmonella* reads by comparison to SSRs. All libraries contained 10 million total reads, and included 5, 10, or 50 *Salmonella*-derived reads. At least one read was positively identified as *Salmonella* in 16/20 replicates of 50 *Salmonella* read libraries, 14/20 replicates of 10 *Salmonella* read libraries, and 12/20 replicates of 5 *Salmonella* read libraries (Table 1), giving a calculated LOD_{50} of 10.2 reads in a 10 million read library [34] (CI: 6.8–15.3). In comparison, Metaphlan4 was much less sensitive, requiring 1×10^4 *Salmonella*-derived reads in a 10 million read library (0.1 %) for reliable detection (Table 1), with a calculated LOD_{50} of 2106 reads in a 10 million read library (CI: 1247–3557).

Limits of detection in a real microbiome background

Previous rounds of analysis made use of fully simulated shotgun sequencing libraries containing reads from members of the Enterobacteriaceae family. To explore use of the analysis pipeline for detecting *Salmonella* in more realistic sets of sequences, libraries were created using published shotgun sequencing datasets from chicken gut microbiomes. *Salmonella* detection was attempted by (a) searching for two *Salmonella* marker genes and (b) using the strategy established above (using the Kraken 2 bacteria database and confidence 0.25, plus comparison to *Salmonella* SSRs). Marker genes *invA* and *stn* are commonly used for *Salmonella* detection in rapid tests such as quantitative PCR and loop-mediated isothermal amplification [35–37]. Read fragments of these genes could be reliably detected (100 % of replicates) in libraries with approx. 4×10^4 *Salmonella*-derived reads, with an LOD_{50} for one or more of the markers of 1754 (CI: 1067–2884) *Salmonella*-derived reads in a 40 million read library (Fig. 3, top panel). Using the established detection pipeline, *Salmonella* reads could be detected in all libraries with 40 *Salmonella*-derived reads, with an LOD_{50} of 5.5 (CI: 3.1–9.8).

Table 1 Number of *Salmonella*-detected at each spike-in level in 10-million read libraries

Salmonella reads in library	Replicates	Positive libraries: Kraken2+SSRs	Positive libraries: Metaphlan4
1×10^5	20	20	20
1×10^4	20	20	20
1×10^3	20	20	4
1×10^2	20	20	0
50	20	16	0
10	20	13	0
5	20	12	0
0	20	0	0

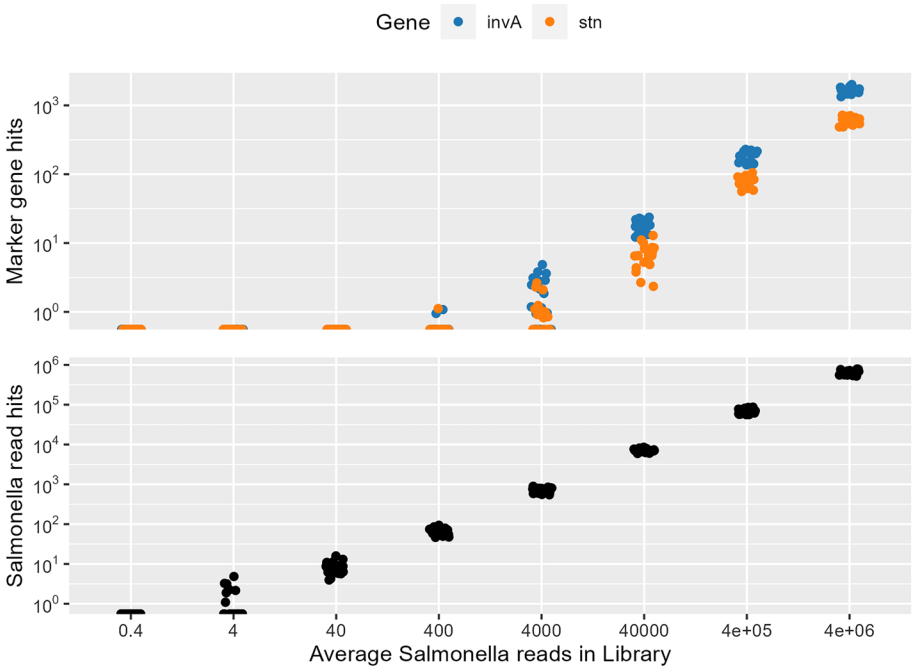

Fig. 3 Detection of *Salmonella* marker genes (top) or *Salmonella*-derived reads (bottom) using the established workflow in a chicken caecal microbiome background. Libraries contained 40 million total reads. Datapoints from individual replicates are shown. Y-axis is in log10 scale

Discussion

Here we investigate the conditions and parameter choices influencing sensitivity and specificity during metagenomic detection of *Salmonella*. We show the importance of appropriate software parameter choices: default parameters, notably a confidence setting of 0, lead to high false positive rates in simulated food microbiomes (Fig. 3. Moreover, choice of database has a pronounced impact on performance, with the kr2bac database showing improvements over the default kr2std database (Fig. 2). Optimal behaviour was obtained only with a carefully chosen combination of parameters and database. It is likely that the best combinations of parameters will vary depending on the pathogen of interest; our purpose here is not to optimize software settings for all possible targets, but rather to draw attention to the need to choose parameters carefully.

Use of species-specific regions (SSRs), as proposed in the SNIPE pipeline [26], is a promising approach for improving identifying and filtering false positives given by Kraken 2. In establishing their pipeline, Huang et al. used default parameters, which we found to be inadequate. At confidence 0 (the Kraken 2 default), false positives persist even after the SSR-comparison filtering step. Furthermore, their testing sets included a very limited number of closely related genomes as a confounding factor, whereas multiple members of the Enterobacteriaceae family could be expected to be present in sample types that are frequent targets for pathogen detection, including human clinical samples [38], food-animal microbiomes [39], or food products [40]. Thus, testing extensively in a dataset containing a large number and variety of related organisms was informative.

We were particularly interested in the sensitivity of pathogen detection. Low limits of detection make it possible to detect *Salmonella* even when it is a very small component

of the sample community. Additionally, extremely sensitive bioinformatic methods allow detection from shallower sequencing datasets, which would reduce costs. Our approach was able to correctly identify 100 % of *Salmonella*-positive sequencing libraries containing just 100 *Salmonella*-derived reads. Even with just five *Salmonella*-derived reads, more than half of library replicates were correctly identified as positive. By comparison, the recently-released Metaphlan4 software was very specific, but far less sensitive. However, one consideration in using such a sensitive detection strategy is the risk of contamination via carry-over between sequencing runs, a known issue with Illumina sequencers [41]. Samples contaminated in this way would legitimately contain reads identified as belonging to the pathogen of interest, and thus be considered positive [24, 42]. There is presently no way to overcome this issue in data analysis once sequencing has been performed; it can only be minimized during wet-lab procedures.

There are additional limitations to this analysis. Almost all members of the *Salmonella* genus are considered pathogenic [43], so identification at the genus level is sufficient for these organisms. Other genera contain both benign and pathogenic members, making species-level identification necessary. One growing concern with species-level identification is that, as reference databases grow, fewer reads can be classified at the species level when using k-mer-based taxonomic classifiers [44]. Still other species or subspecies are benign unless they carry certain virulence factors (for example, the majority of *E. coli* are harmless, but Shiga-toxin producing *E. coli* (STEC) cause gastrointestinal illness and even death [45]). In such cases, virulence genes or genetically linked markers must be detected for positive identification [15, 46]. We show that far higher pathogen numbers in a population are required for detection of marker genes (in this study, *invA* and *stn*) compared to general genomic reads.

We found that best results came from using the Kraken 2 bacteria database; however, we had prior knowledge that the pathogen-of-interest is bacterial. Diagnostic analyses where the cause of disease is unknown would require use of additional databases (ex. a virus database), and many additional SSRs for various species. The kraken 2-build function allows the production of custom databases, so it would be possible to create a combined bacteria-virus database, and to add in organisms of interest that are not yet included. Furthermore, our datasets included either no host reads (the simulated enterobac datasets) or a negligible number of reads matching to the *Gallus gallus* genome (the chicken caecal datasets). Depending on the sample type, some real metagenomic sequencing datasets can contain a high proportion of host reads, which can be a confounding factor in taxonomic identification of microbiome constituents, even after commonly-employed steps that aim to remove host reads. For samples from the human microbiome, using a standard database that includes human reads can greatly reduce false identification of sequencing reads [47]. For non-human hosts, incorporating host genomes into custom-built Kraken 2 databases could be advantageous and should be explored in future studies.

Conclusions

Shotgun sequencing is gaining popularity in many biological fields, including food safety. However, it is challenging to analyze the resulting datasets for the presence of pathogens with a high degree of both sensitivity and specificity. Robust analysis strategies are

essential, since false positives could lead to food recalls or production shut-downs and false negatives could lead to preventable illnesses. Many pipelines exist for metagenomics-based detection of foodborne pathogens [24], but these pipelines are often not tested on mock communities where the provenance of each read is known and false classification can be assessed. Here, we have investigated the impact of parameter and database choice on several popular approaches, using *Salmonella* as a model pathogen. We emphasize that careful consideration of software parameter and database choices is essential. With well-chosen parameters plus additional steps to confirm the taxonomic origin of reads, it is possible to detect pathogens with very high specificity and sensitivity.

Methods

Mock community

The mock community "enterobac" is composed of members of the Enterobaceriaceae family, to which the genus *Salmonella* belongs. Complete reference genomes for 62 species in the Enterobacteraceae family were selected using the NCBI genome browser [48] and downloaded from the NCBI RefSeq database (See Table S1, Additional file 2).

The art_illumina function of ART [49] was used to generate simulated shotgun sequencing reads for each genome with the following parameters: 25-fold coverage, paired reads of length 150 bp with insert size 300 bp, read length standard deviation of 10 bp, and an error profile from the Illumina HiSeq 2500. Reads from all genomes except *Salmonella enterica* subsp. enterica serovar Typhimurium str. LT2 were concatenated into "master" mock community files with a total of 26,976,269 paired-end reads.

Mock libraries

Libraries of 10 million paired reads were produced by randomly subsetting reads from the master file using BBMAPs' reformat function [50]. Read counts were chosen based on the desired number of *Salmonella* reads per library; for example, for the 10 % *Salmonella* library, 1×10^5 reads were selected from those produced from the *Salmonella* Typhimurium genome, and 9×10^5 reads were selected from the master mock community file. Twenty replicates were produced at each target level.

An additional mock community was also generated, comprised of unpublished genomic data from 31 strains erroneously identified as *Salmonella* by either MASH (1 *Citrobacter* spp.), [33] 16 S sequence analysis (6 *Enterobacter/Klebsiella* spp.) or detection of species-specific genes (3 *Citrobacter* spp.) (See Table S2, Additional file 3). Genomes had 35–82 fold coverage, with a total of 40 Million paired end reads. Libraries were produced similar to above. Illumina HiSeq short reads were synthesized from the draft genome assemblies and raw reads of the bacterial genomes using the FetaGenome2 (fabricate metagenome) tool developed in house [51]. Briefly, Art version 2.5.8 was used to simulate paired-end HiSeq reads of 150 bp in length with a 300 bp insert size. To simulate variability in coverage levels (e.g. higher coverage in plasmids vs chromosomal sequences), the FetaGenomePlasmidAware edition uses BWA to map reads to the original assembly to determine coverage depth of each contig in the given assembly, then uses the coverage report output to create more reads for higher-depth locations and fewer reads for low-depth locations of the genome. The simulated library was tested

with the current workflow, with Kraken 2 confidence of 0.25 and the kr2bac database, followed by confirmation by checking against *Salmonella* SSRs.

Kraken 2 reference databases

A pre-indexed version of the Kraken 2 [19] standard database ("kr2std"), which contains archaea, bacteria, viral, plasmid, human, and UniVec_Core sequences [52] was downloaded from https://benlangmead.github.io/aws-indexes/k2 on 01 Oct 2021 (database last updated 17 May 2021).

The Kraken 2 bacteria library and taxonomy were downloaded on 28 Oct 2021 according to the software manual instructions (see Supplementary material; Additional file 1). The unaltered Kraken 2 bacteria databases ("kr2bac") was built using these files. Database "kr_plrenamed_db" was built after altering the bacteria library file according to instructions from Doster et al. [25] (see Supplementary material, Additional file 1). Plasmids in the bacteria library fasta file were renamed using sed, and the database was then built as above.

Salmonella species specific regions (SSRs)

Laing et al. [31] investigated the *Salmonella* pan-genome and found 403 regions of 1000 bp each that were specific to the *Salmonella* genus. These regions were used to confirm the identity of reads classified as *Salmonella*-derived by Kraken 2. The position of on these regions on the *Salmonella* reference genome (*Salmonella enterica* subsp. enterica serovar Typhimurium str. LT2) was taken from the supplementary files [31] and the faidx function of samtools [53] was used to extract the sequences in fasta format. A blast-formatted database was then created using the sequences and BLAST CLI's makeblastdb command [54, 55].

Workflows

Custom Snakemake [56] workflows were written to carry out library setup and analyses. Each mock library was first subject to trimming with Trimmomatic [57] with parameters SLIDINGWINDOW:4:20 MINLEN:36. Singleton files passing quality check from the Trimmomatic output were concatenated with paired files to ensure minimal loss of sequences. Reads were then classified with Kraken 2 [19]. The first round of analysis was used to establish the best database and confidence level. For this, 10 million-read libraries with 1 %, 0.1 %, and 0.01 % *Salmonella* content were classified with each of the three Kraken 2 databases described above (kr2std, kr2bac, and kr2plrename) at five confidence levels: 0 (default), 0.25, 0.5, 0.75, and 1.

Output from this analysis was also used to establish the utility of comparison to SSRs [31] for removing false positives. Information about reads classified as members of the *Salmonella* genus ("putative *Salmonella* hits") was extracted from the Kraken 2 output and the origin of the read recorded using a custom Python script to determine the number of false positives (that is, reads originating from a non-*Salmonella* genome that were classified as a member of the *Salmonella* genus). Sequences from all Kraken 2 *Salmonella* hits were compared to the SSR database using the BLAST command line application's blastn function [54] with max_target_seqs=1 and max_hsps=1. The origin of these SSR *Salmonella* hits was again checked to determine remaining false positives.

The second round of analysis explored lower limits of detection in mock communities based on best practices from the above analyses. Libraries with 0.005 %, 0.001 %, and 0.0005 % *Salmonella* were classified with Kraken 2 against the bacteria database ("kr2bac") at 0.25 confidence. Kraken 2 *Salmonella* hits were extracted and compared to SSRs, and false positives were recorded, as above.

Mock libraries were also analyzed with Metaphlan4 [21] using the vJan21 database. All reads that passed the Trimmomatic step were combined into one file per library and analyzed with default parameters, using the output parameters "unclassified_estimation" and "-t rel_ab_w_read_stats". Individual library profiles were combined with the merge_metaphlan_tables.py script, and libraries with at least one read in the *Salmonella* genus were considered positive for *Salmonella*.

Limits of detection in a real metagenomic background

Limits of detection for *Salmonella*-derived reads were further explored using published chicken caecal shotgun libraries as the background microbiome. Sequencing files from Salaheen et al. [58] were retrieved from the European Nucleotide Archive (accession codes SRR5280289, SRR5280393, and SRR5280514). Briefly, the Salaheen et al. [58] study investigated the impact of antibiotic growth promotors on cecal microbiomes of Cobb-500 broiler chicks. Retrieved sequences were from control chickens which did not receive growth promotors. These reads were paired-end (2 x 151 bp) from an Illumina NextSeq 500. Sequencing files were concatenated to create master microbiome files containing 119,068,070 paired reads. The master files were classified with Kraken 2 [19] using the kr2bac library and confidence 0.2, and all reads matching to *Salmonella* were removed. This resulted in master files of 119,068,030 reads. Reads were also checked against a custom database derived from the chicken (Gallus gallus) reference genome to ensure that the number of host reads in the shotgun dataset was neglible.

The assembled genome of *Salmonella* Enteritidis strain CFIAFB20140150 (accession code SRR10859048) [59] was used to generate simulated paired-end HiSeq reads of 150 bp in length with a 300 bp insert sizeusing the art_illumina function of ART [49], as above. This strain was chosen based on its concurrent use in a laboratory spike-in study. Replicate libraries were produced by appending the appropriate number of *Salmonella* Enteritidis-derived reads to the master microbiome files, then subsetting libraries of 40 million reads using BBMAP's reformat function [50]. Libraries were produced at eight target levels, from 10 % (approx. 4 million S. Enteritidis-derived reads) to 0.000001 % (0.4 reads), and 20 replicates were produced per target level.

Libraries were analyzed using the above workflow, with the Kraken 2 bacteria database, confidence 0.25, and SSR checks. Additionally, DIAMOND-formatted databases of the *invA* and *stn* marker genes were created using animo acid sequences retrieved from NCBI (WP_000927219.1 and AAA21354.1, respectively) with the DIAMOND makedb function [60]. The presence of these genes in libraries was tested using DIAMOND's blastx function with a percent ID cutoff of 96.

Statistics

Plotting and statistical analyses were carried out in R v4.2.2 [61]. The full list of packages used is available in the supplementary material (see Supplementary material, Additional

file 1). LOD_{50} was calcluated via the log-log model by Wilrich and Wilrich [34] using a tool they provide online[1] Although this model was developed for calculating LODs in terms of bacterial CFU per gram of food matrix during spike-in experiments, we adapted the calculation for counts of pathogen-derived reads in sequencing libraries.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05952-x.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements
We are grateful for funding from the Ontario Ministry of Agriculture, Food, and Rural Affairs (OMAFRA). This research was enabled in part by support provided by SHARCNET (sharcnet.ca) and the Digital Research Alliance of Canada (alliancecan.ca). Thanks to Ashley Cooper at the CFIA for producing the mock library from unpublished genomes, and to Ryan Taylor at Carleton University for sharing his technical expertise.

Author Contributions
LMB, AW and CCconceived of the study. LMB carried out the study and drafted the manuscript. AW and CC contributed resources and revised the manuscript.

Funding
Funding for this project was provided by the Ontario Ministry of Agriculture, Food, and Rural Affairs (OMAFRA project number OAF-2020-101088).

Availability of data and materials
Genome sequences of unusual isolates are available at 10.5281/zenodo.8056523. Analysis code can be viewed at https://github.com/LMBradford/salmdetectpipeline.git.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 July 2023 Accepted: 7 October 2024
Published online: 03 December 2024

References

1. World Health Organization. WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference Grou. Geneva: World Health Organization; 2015. p. 2007–15.
2. Lee H, Yoon Y. Etiological agents implicated in foodborne illness world wide. Food Sci Anim Resour. 2021;41(1):1.
3. Banerjee G, Agarwal S, Marshall A, Jones DH, Sulaiman IM, Sur S, Banerjee P. Application of advanced genomic tools in food safety rapid diagnostics: challenges and opportunities. Curr Opin Food Sci. 2022;47:100886.
4. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Hilbert F, et al. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. EFSA J. 2019;17(12):05898.
5. Bell RL, Jarvis KG, Ottesen AR, McFarland MA, Brown EW. Recent and emerging innovations in salmonella detection: a food and environmental perspective. Microb Biotechnol. 2016;9(3):279–92.
6. Muhamad Rizal NS, Neoh H-M, Ramli R, ALK Periyasamy PR, Hanafiah A, Abdul Samat MN, Tan TL, Wong KK, Nathan S, Chieng S, et al. Advantages and limitations of 16s rrna next-generation sequencing for pathogen identification in the diagnostic microbiology laboratory: perspectives from a middle-income country. Diagnostics. 2020;10(10):816.

[1] https://www.wiwiss.fu-berlin.de/fachbereich/vwl/iso/ehemalige/wilrich/index.html.

7. Yap M, Ercolini D, Álvarez-Ordóñez A, O'Toole PW, O'Sullivan O, Cotter PD. Next-generation food research: use of meta-omic approaches for characterizing microbial communities along the food chain. Annu Rev Food Sci Technol. 2022;13:361–84.
8. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, et al. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol. 2019;79:96–115.
9. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: the next culture-independent game changer. Front Microbiol. 2017;8:1069.
10. Shah N, Tang H, Doak TG, Ye Y. Comparing bacterial communities inferred from 16s rrna gene sequencing and shotgun metagenomics. In: Biocomputing 2011, pp. 165–176. World Scientific, Singapore 2011.
11. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. Biochem Biophys Res Commun. 2016;469(4):967–77.
12. Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL, et al. Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. Appl Environ Microbiol. 2016;82(8):2433–43.
13. Duarte ASR, Röder T, Van Gompel L, Petersen TN, Hansen RB, Hansen IM, Bossers A, Aarestrup FM, Wagenaar JA, Hald T. Metagenomics-based approach to source-attribution of antimicrobial resistance determinants-identification of reservoir resistome signatures. Front Microbiol. 2021;11:601407.
14. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015;53(5):1685–92.
15. Carleton HA, Besser J, Williams-Newkirk AJ, Huang A, Trees E, Gerner-Smidt P. Metagenomic approaches for public health surveillance of foodborne infections: opportunities and challenges. Foodborne Pathog Dis. 2019;16(7):474–9.
16. Cocolin L, Mataragas M, Bourdichon F, Doulgeraki A, Pilet M-F, Jagadeesan B, Rantsiou K, Phister T. Next generation microbiological risk assessment meta-omics: the next need for integration. Int J Food Microbiol. 2018;287:10–7.
17. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, et al. Critical assessment of metagenome interpretation–a benchmark of metagenomics software. Nat Methods. 2017;14(11):1063–71.
18. Meyer F, Fritz A, Deng Z-L, Koslicki D, Lesker TR, Gurevich A, Robertson G, Alser M, Antipov D, Beghini F, et al. Critical assessment of metagenome interpretation: the second round of challenges. Nat Methods. 2022;19(4):429–40.
19. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20(1):1–13.
20. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R. Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 2016;12(6):1004957.
21. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. Nat Biotechnol. 2023;41:1633–44.
22. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with kaiju. Nat Commun. 2016;7(1):1–9.
23. Ounit R, Wanamaker S, Close TJ, Lonardi S. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):1–13.
24. Höper D, Grützke J, Brinkmann A, Mossong J, Matamoros S, Ellis RJ, Deneke C, Tausch SH, Cuesta I, Monzón S, et al. Proficiency testing of metagenomics-based detection of food-borne pathogens using a complex artificial sequencing dataset. Front Microbiol. 2020;11:575377.
25. Doster E, Rovira P, Noyes NR, Burgess BA, Yang X, Weinroth MD, Linke L, Magnuson R, Boucher C, Belk KE, et al. A cautionary report for pathogen identification using shotgun metagenomics; a comparison to aerobic culture and polymerase chain reaction for salmonella enterica identification. Front Microbiol. 2019;10:2499.
26. Huang L, Hong B, Yang W, Wang L, Yu R. Snipe: highly sensitive pathogen detection from metagenomic sequencing data. Brief Bioinform. 2021;22(5):064.
27. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):1–12.
28. Wright RJ, Comeau AM, Langille MG. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. bioRxiv 2022.
29. Thomas MK, Murray R, Flockhart L, Pintar K, Pollari F, Fazil A, Nesbitt A, Marshall B. Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. Foodborne Pathog Dis. 2013;10(7):639–48.
30. World Health Organization: Salmonella (non-typhoidal) [Fact Sheet]. https://www.who.int/news-room/fact-sheets/detail/salmonella-(non-typhoidal)
31. Laing CR, Whiteside MD, Gannon VP. Pan-genome analyses of the species salmonella enterica, and identification of genomic markers predictive for species, subspecies, and serovar. Front Microbiol. 2017;8:1345.
32. Lu J. Kraken 2 Manual. https://github.com/DerrickWood/kraken2/wiki/Manual
33. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using minhash. Genome Biol. 2016;17:1–14.
34. Wilrich C, Wilrich P-T. Estimation of the pod function and the lod of a qualitative microbiological measurement method. J AOAC Int. 2009;92(6):1763–72.
35. Rahn K, De Grandis S, Clarke R, McEwen S, Galan J, Ginocchio C, Curtiss Iii R, Gyles C. Amplification of an inva gene sequence of salmonella typhimurium by polymerase chain reaction as a specific method of detection of salmonella. Mol Cell Probes. 1992;6(4):271–9.
36. Moore M, Feist M. Real-time PCR method for *Salmonella* spp. targeting the *stn* gene. J Appl Microbiol. 2007;102(2):516–30.
37. Ou H, Wang Y, Gao J, Bai J, Zhang Q, Shi L, Wang X, Wang C. Rapid detection of salmonella based on loop-mediated isothermal amplification. Ann Palliat Med. 2021;10(6):6850–8.

38. King CH, Desai H, Sylvetsky AC, LoTempio J, Ayanyan S, Carrie J, Crandall KA, Fochtman BC, Gasparyan L, Gulzar N, et al. Baseline human gut microbiota profile in healthy people and standard reporting template. PLoS One. 2019;14(9):0206484.
39. Shang Y, Kumar S, Oakley B, Kim WK. Chicken gut microbiota: importance and detection technology. Front Vet Sci. 2018;5:254.
40. Baylis C, Uyttendaele M, Joosten H, Davies A, Heinz H. The enterobacteriaceae and their significance to the food industry, ilsi europe report series. Technical report, Washington, DC: International Life Sciences Institute; 2011.
41. Illumina: Reducing run-to-run carryover on the MiSeq using dilute sodium hypochlorite solution. San Diego, Illumina; 2013.
42. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. Analysis, optimization and verification of illumina-generated 16s rrna gene amplicon surveys. PLoS One. 2014;9(4):94249.
43. Eng S-K, Pusparajah P, Ab Mutalib N-S, Ser H-L, Chan K-G, Lee L-H. Salmonella: a review on pathogenesis, epidemiology and antibiotic resistance. Front Life Sci. 2015;8(3):284–93.
44. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. Refseq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol. 2018;19:1–10.
45. Majowicz SE, Scallan E, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Yeung DH, Kirk MD. Global incidence of human shiga toxin-producing *Escherichia coli* infections and deaths: a systematic review and knowledge synthesis. Foodborne Pathog Dis. 2014;11(6):447–55.
46. Riley LW. Distinguishing pathovars from nonpathovars: *Escherichia coli*. Microbiol Spectr. 2020;8(4):10. https://doi.org/10.1128/microbiolspec.ame-0014-2020
47. Gihawi A, Ge Y, Lu J, Puiu D, Xu A, Cooper CS, Brewer DS, Pertea M, Salzberg SL. Major data analysis errors invalidate cancer microbiome findings. MBio. 2023;14(5):01607–23.
48. NCBI Genome Browser. https://www.ncbi.nlm.nih.gov/datasets/genomes/?taxon=543
49. Huang W, Li L, Myers JR, Marth GT. Art: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.
50. Bushnell B. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley; 2014.
51. Low A. FetaGenome2. https://github.com/OLC-Bioinformatics/FetaGenome2
52. Langmead B. Kraken 2, KrakenUniq and Bracken indexes. https://benlangmead.github.io/aws-indexes/k2
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
54. National Center for Biotechnology Information (US): BLAST Command Line Applications User Manual. 2008.
55. Madden T. The blast sequence analysis tool. The NCBI handbook; 2003.
56. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. Sustainable data analysis with snakemake. F1000Research 2021;10:33.
57. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
58. Salaheen S, Kim S-W, Haley BJ, Van Kessel JAS, Biswas D. Alternative growth promoters modulate broiler gut microbiome and enhance body weight gain. Front Microbiol. 2017;8:2088.
59. Cooper AL, Low AJ, Koziol AG, Thomas MC, Leclair D, Tamber S, Wong A, Blais BW, Carrillo CD. Systematic evaluation of whole genome sequence-based predictions of salmonella serotype and antimicrobial resistance. Front Microbiol. 2020;11:549.
60. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using diamond. Nat Methods. 2021;18(4):366–8.
61. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 2021. R Foundation for Statistical Computing. https://www.R-project.org/

Publisher's Note