# ConFindr: Rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data

**Andrew J Low** Equal first author, [1] , **Adam G Koziol** [1] , **Paul A Manninger** [1] , **Burton W Blais** [1] , **Catherine D Carrillo** Corresp. Equal first author, [1]

[1] Ottawa Laboratory (Carling), Canadian Food Inspection Agency, Ottawa, Ontario, Canada

Corresponding Author: Catherine D Carrillo
Email address: catherine.carrillo@canada.ca

Whole-genome sequencing (WGS) of bacterial pathogens is currently widely used to support public-health investigations. The ability to assess WGS data quality is critical to underpin the reliability of downstream analyses. Sequence contamination is a quality issue that could potentially impact WGS-based findings; however, existing tools do not readily identify contamination from closely-related organisms. To address this gap, we have developed a computational pipeline, ConFindr, for detection of intraspecies contamination. ConFindr determines the presence of contaminating sequences based on the identification of multiple alleles of core, single-copy, ribosomal-protein genes in raw sequencing reads. The performance of this tool was assessed using simulated and lab-generated Illumina short-read WGS data with varying levels of contamination (0-20% of reads) and varying genetic distance between the designated target and contaminant strains. Intraspecies and cross-species contamination was reliably detected in datasets containing 5% or more reads from a second, unrelated strain. ConFindr detected intraspecies contamination with higher sensitivity than existing tools, while also being able to automatically detect cross-species contamination with similar sensitivity. The implementation of ConFindr in quality-control pipelines will help to improve the reliability of WGS databases as well as the accuracy of downstream analyses. ConFindr is written in Python, and is freely available under the MIT License at github.com/OLC-Bioinformatics/ConFindr.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27499v1 | CC BY 4.0 Open Access | rec: 22 Jan 2019, publ: 22 Jan 2019

# ConFindr: Rapid Detection of Intraspecies and Cross-Species Contamination in Bacterial Whole-Genome Sequence Data

**Andrew J. Low**[1], **Adam G. Koziol**[1], **Paul A. Manninger**[1], **Burton W. Blais**[1], **and Catherine D. Carrillo**[1]

[1]**Ottawa Laboratory (Carling), Canadian Food Inspection Agency, Ottawa, Ontario, Canada**

Corresponding author:
Catherine D. Carrillo[1]

Email address: catherine.carrillo@canada.ca

## ABSTRACT

Whole-genome sequencing (WGS) of bacterial pathogens is currently widely used to support public-health investigations. The ability to assess WGS data quality is critical to underpin the reliability of downstream analyses. Sequence contamination is a quality issue that could potentially impact WGS-based findings; however, existing tools do not readily identify contamination from closely-related organisms. To address this gap, we have developed a computational pipeline, ConFindr, for detection of intraspecies contamination. ConFindr determines the presence of contaminating sequences based on the identification of multiple alleles of core, single-copy, ribosomal-protein genes in raw sequencing reads. The performance of this tool was assessed using simulated and lab-generated Illumina short-read WGS data with varying levels of contamination (0-20% of reads) and varying genetic distance between the designated target and contaminant strains. Intraspecies and cross-species contamination was reliably detected in datasets containing 5% or more reads from a second, unrelated strain. ConFindr detected intraspecies contamination with higher sensitivity than existing tools, while also being able to automatically detect cross-species contamination with similar sensitivity. The implementation of ConFindr in quality-control pipelines will help to improve the reliability of WGS databases as well as the accuracy of downstream analyses. ConFindr is written in Python, and is freely available under the MIT License at github.com/OLC-Bioinformatics/ConFindr.

## INTRODUCTION

Public-health microbiology laboratories increasingly apply bacterial whole-genome sequence (WGS) analyses for pathogen identification, high-resolution typing and risk profiling (Ronholm et al., 2016; Allard et al., 2016; Taboada et al., 2017). Reductions in cost for generating WGS data have led to the widespread use of this technology for tracking foodborne pathogens internationally, and public databases currently include sequences for hundreds of thousands of isolates. A significant effort has been undertaken to produce guidelines and minimum standards for sequence data quality, particularly when such data is used to support regulatory activities (Lambert et al., 2017; W.A. Rossen et al., 2017).

Quality assessment tools are typically integrated into bioinformatics workflows to ensure the reliability of WGS data (Koren et al., 2014; Page et al., 2016). For example, FastQC is used to assess the per-base quality of raw reads to identify problems with the sequencing libraries or runs (Andrews, 2010). Tools such as QUAST can be used to evaluate the quality of *de novo* assemblies, identify misassemblies, determine error rates, and more (Gurevich et al., 2013). These tools can be extremely valuable for identifying inferior datasets; however, assessing contamination is outside of their current scope.

The presence of contamination in WGS data is recognized as an important sequence quality issue (Merchant et al., 2014; Ballenghien et al., 2017; Robertson et al., 2018; Cornet et al., 2018). Introduction of contaminants can occur at many stages in the generation of bacterial sequence data. For example, cultures recovered from samples may not be adequately purified, or cross-contamination could occur during

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27499v1 | CC BY 4.0 Open Access | rec: 22 Jan 2019, publ: 22 Jan 2019

preparation of genomic DNA or sequencing-libraries (Merchant et al., 2014). Carryover contamination results from the presence of residual fragments from previous sequencing runs (Souvorov et al., 2018). While integration of controls can help to identify pervasive contamination issues, they are not effective for the identification of sporadic contamination events.

Cross-species contamination in short-read WGS data can be readily identified by taxonomic classification of sequence reads using reference databases (Wood and Salzberg, 2014; Merchant et al., 2014; Ounit et al., 2015; Mallet et al., 2017). Contamination can also be inferred following *de novo* assembly of short sequencing reads into a contiguous bacterial chromosome. Contiguity can be impacted by presence of contaminating sequencing reads, but also by factors such as the assembler used, length of the sequencing reads, presence of repeat regions, GC content and coverage (Lin et al., 2011; Jünemann et al., 2014; Souvorov et al., 2018). Contamination may be indicated by a highly fragmented assembly, or a genome size that is larger than expected (Robertson et al., 2018). However, establishment of appropriate cutoffs requires determination of acceptable ranges within a species, and atypical strains may fall outside of these limits.

Intraspecies contamination is far more difficult to detect because read-classification approaches cannot be used. In some studies, the quality of metagenomic or single-cell sequencing data is assessed by evaluating core genes to determine the completeness and degree of contamination of assemblies (Hess et al., 2011; Parks et al., 2015). One of the tools developed for this purpose is CheckM, which determines the presence of contamination based on the identification of multiple copies of lineage-specific, ubiquitous, single-copy genes (Parks et al., 2015). To our knowledge, there are no tools designed and evaluated specifically for the detection of intraspecies contamination in bacterial-isolate sequence data.

We have developed a bioinformatics tool, ConFindr, which can accurately and rapidly identify intra- and cross-species contamination based on the analysis of raw sequencing reads. We evaluated the performance of this tool for detecting contamination in Illumina short-read WGS data derived from priority foodborne pathogens *Listeria monocytogenes*, *Salmonella enterica* and Shiga-toxin producing *Escherichia coli* (STEC).

# METHODS

## ConFindr Workflow and Implementation

ConFindr determines the presence of contaminating sequencing reads based on the analysis of the set of 53 genes encoding the bacterial ribosomal-protein subunits that are used in the ribosomal multilocus sequence typing scheme (rMLST) (Jolley et al., 2012). The rMLST genes are typically present as single copies and are conserved across the entire bacterial domain, with some exceptions where multiple alleles for a gene exist or no gene exists. ConFindr works on the principle that a genome containing more than one allele for any rMLST gene is contaminated, taking into consideration the known exceptions.

In its first step, ConFindr uses a screening functionality provided in Mash (Ondov et al., 2016) to determine which genera are present in a sample. This screen is done against a custom database derived from the NCBI RefSeq genomes (https://www.ncbi.nlm.nih.gov/refseq/) with one genome representing each species (O'Leary et al., 2015). If more than one genus is detected, ConFindr reports cross-species contamination for the sample and does not proceed further. If only one genus is present, ConFindr creates a genus-specific rMLST database by extracting all rMLST sequences associated with the target genus, excluding genes known to have multiple alleles, and proceeds to attempt to find contamination by searching for multiple alleles of one or more of the benchmark rMLST genes.

To search for multiple alleles, ConFindr begins by using BBDuk (Bushnell, 2014) to extract reads that are likely part of the rMLST gene set. These baited reads are stringently trimmed, again using BBDuk, and then aligned to the rMLST genes using BBMap (Bushnell, 2014). The resulting BAM file is then parsed in order to find 'Contaminating Single Nucleotide Variants' (cSNVs) – that is, sites in the pileup where more than one base is present. Since all rMLST genes are known to be present as single copies, the occurrence of multi-base sites in the pileup indicates multiple alleles, and therefore contamination. To be called as a cSNV, at least 2 bases of the minor variant with a Phred score of 20 or greater must be present at that site, and at least 5 percent of bases must support the minor variant (though these parameters can be changed by the user). ConFindr determines that a sample is contaminated if multiple genera are found in the Mash screen step or if three or more cSNVs are found.

## *in silico* Dataset Creation

To create *in silico* datasets, we selected complete assemblies from RefSeq for *E. coli*, *S. enterica*, and *L. monocytogenes* (accessions NC_002695.1, NC_003198.1, and NC_003210.1, respectively) and generated variants of these genomes with 100, 500, 1000, and 2000 SNVs using a custom script available at https://github.com/lowandrew/MutantCreator). Simulated reads were then created from both variant and base genomes using ART v2.5.8 (Huang et al., 2012) and mixed together in proportions of 0, 1, 5, 10, and 20 percent contamination using scripts found at https://github.com/lowandrew/FastQMixer to a total coverage depth of approximately 60X. Five replicates were created for each mixed read set.

## Test Datasets

We generated a test dataset of 48 samples comprised of intra- and cross-species mixes of *E. coli*, *S. enterica*, and *L. monocytogenes* isolates with varying levels of relatedness (Table 1). Average nucleotide identity (ANI) was calculated using OrthoANI version 1.4.0 (Lee et al., 2016). This dataset was made both *in silico* by mixing together reads from previous runs of these isolates using the reformat.sh program of the BBMap package (Bushnell, 2014) to a coverage depth of 80X, as well as by sequencing lab-generated mixes of genomic DNA (gDNA).

To generate WGS data, bacterial isolates were cultured in Brain Heart Infusion (BHI) broth (Oxoid Ltd., Basingstoke, Hampshire, England) for 4 to 6 h at 36 ° C, and gDNA was extracted using the Maxwell 16 Cell SEV DNA Purification Kit (Promega, Madison, WI). DNA was quantified using the Quant-it High-Sensitivity DNA Assay Kit (Life Technologies Inc., Burlington, ON). Sequencing libraries were constructed from 1 ng of gDNA using the Nextera XT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA) and the Nextera XT Index Kit (Illumina, Inc.) according to manufacturers' instructions. Genomic sequencing was performed on the Illumina MiSeq Platform (Illumina, Inc.) using a 600-cycle MiSeq Reagent kit v3 (Illumina, Inc.).

## Nucleotide Sequence Accession Numbers

Raw data have been deposited at DDBJ/EMBL/GenBank under BioProject PRJNA507762. The accession numbers and strain descriptions are listed in the Supplemental Table S1.

## Genome Assemblies and Quality Metrics

All of the read sets created were also put through the process of *de novo* assembly. Briefly, reads were quality trimmed using bbduk.sh and error corrected with tadpole.sh (both of the BBMap package (Bushnell, 2014)) and then assembled using SKESA v2.3.0 (Souvorov et al., 2018). Exact commands used to carry this out can be found in Supplemental Table S2. Genome quality statistics were assessed using QUAST v4.6.3 (Gurevich et al., 2013).

## Calculation of Number of SNVs Between rMLST Types

To calculate the number of SNVs between rMLST types within *E. coli*, *S. enterica*, and *L. monocytogenes*, we retrieved all rMLST allele sequences and the list of profiles (accessed at https://pubmlst.org/rmlst/, November 1, 2018). We then extracted sequences for each allele within each rMLST type (1641 types for *L. monocytogenes*, 3062 types for *E. coli*, and 7255 for *S. enterica*). The number of SNVs between every sequence type pair within each species was calculated by aligning each gene in the first type against each gene in the second type using the pairwise2 module in biopython (Cock et al., 2009).

## Dataset Testing

To detect contamination in the datasets generated, ConFindr v0.4.4 was run on default settings on all samples generated. Kraken v1.0 (Wood and Salzberg, 2014) was run on fastq files that had been trimmed to a quality of 15 with bbduk.sh against the standard Kraken database. Exact commands used to carry this out can be found in Supplemental Table S2. Strains/species were determined to be present if at least 0.5 percent of classified reads could be assigned to them. CheckM v1.0.11 (Parks et al., 2015) was run with the lineage_wf workflow for each assembly, and samples were called as contaminated if their contamination level was 2.5 percent or greater. We note that this is a fairly conservative cut-off and CheckM may be able to detect contamination with more sensitivity, but in our experience CheckM results with low levels of contamination require further manual analysis. As we focus on automated detection of contamination, we feel that this is an appropriate cut-off.

**Table 1.** Summary of test dataset used for ConFindr Evaluation

| Target Strain | Contaminant Strain | rMLST SNVs | ANI | Contamination Levels |
|---|---|---|---|---|
| *E. coli* O121:H19 (OLF17053-3) | *E. coli* O121:H19 (OLC2152) | 0 | 99.98 | 0, 20 |
| | *E. coli* O121:H19 (OLC2152) | 11 | 98.86 | 0, 5, 10, 20 |
| | *E. coli* O15:H14 (OLF17030) | 11 | 98.78 | 0, 5, 10, 20 |
| | *E. coli* O8:H28 (OLF17043) | 11 | 98.83 | 0, 5, 10, 20 |
| | *Enterobacter cancerogenus* (OLC1687) | N/A | 78.86 | 0, 5 |
| *S. Heidelberg* (OLC2542) | *S. Heidelberg* (OLC2000) | 0 | 99.99 | 0, 20 |
| | *S. Bredeney* (OLC2229) | 32 | 98.36 | 0, 5, 10 20 |
| | *S. Typhimurium* (OLF13104-7) | 24 | 99.08 | 0, 5, 10 20 |
| | *S. Dublin* (OLF18064-1) | 33 | 98.82 | 0, 5, 10 20 |
| | *Citrobacter freundii* (OLC1136) | N/A | 81.83 | 0, 5, 10 20 |
| *L. monocytogenes* (OLF10129) | *L. monocytogenes* (OLF11041-1) | 0 | 99.99 | 0, 20 |
| | *L. monocytogenes* (OLF13043-2) | 16 | 99.54 | 0, 5, 10, 20 |
| | *L. monocytogenes* (OLF15140) | 10 | 99.45 | 0, 5, 10, 20 |
| | *L. monocytogenes* (OLF09168) | 133 | 94.85 | 0, 5, 10, 20 |
| | *Listeria innocua* (OLC0004) | 420 | 88.26 | 0, 5, 10 |
| | *Enterococcus faecalis* (OLC0147) | N/A | 66.36 | 0, 5 |

# RESULTS

## Identification of contaminating SNVs within rMLST genes using ConFindr

As ConFindr is based on finding contaminating SNVs (cSNVs) within the rMLST genes in raw reads, we first looked at the reliability of detection of cSNVs in simulated data with different levels of contamination. Synthetic mutants with 100 to 2000 SNVs relative to reference genomes were generated *in silico*, and the number of SNVs occurring within rMLST genes in these mutants was calculated. The number of cSNVs found by ConFindr in the *in silico* datasets at 60 times coverage was compared with the predicted number of SNVs within rMLST genes in the two isolates making up the contaminated sample (Figure 1). As the relative contamination increased, ConFindr's estimate of the number of cSNVs in the sample approached the expected number of SNVs. Contamination at 5% was reliably detected when the contaminant had at least 16 SNVs within the rMLST genes, at 10% with at least 7 SNVs and at 20% with at least 3 SNVs within target genes.

## Diversity among rMLST sequences types

To illustrate the genetic diversity within the rMLST scheme, we calculated the numbers of SNVs between ribosomal sequence types (rSTs) for *L. monocytogenes*, *S. enterica*, and *E. coli* (Figure 2). Over 99 percent of all pairs in all three species had three or more SNVs, which is the cutoff chosen in ConFindr as the minimum number of cSNVs that need to be found before a sample will be considered contaminated. Over 80 percent of the sequence types have 16 or more SNVs relative to others. Therefore, ConFindr should almost always be able to detect contamination between two isolates with different rMLST types.

## ConFindr detects contamination with more sensitivity that existing tools

We compared ConFindr to existing tools capable of detecting contamination, CheckM and Kraken using both *in silico* and lab-generated datasets. Mixes were binned based on the Average Nucleotide Identity (ANI) of the two samples being mixed - those with >99 percent ANI, representing very closely related mixes, those with between 98 and 99 percent ANI, representing same-species mixes between strains not as closely related, and those with ANI of less than 98 percent, representing distantly related same-species mixes and cross-species mixes (Table 1).

ConFindr was more sensitive than either CheckM or Kraken for intraspecies contamination detection (Figure 3, panels A, B, D and E), and comparable to both for cross-species contamination detection (Figure 3, Panels C and F). ConFindr detected contamination successfully in all cases except for one simulated mix of closely related strains at 5 percent contamination (Supplemental Table S1), while both CheckM and Kraken required either more distance between species or a higher level of contamination for its determination.

## Assembly Metrics are Insufficient for Contamination Detection

We assembled the contaminated datasets used in this study to assess metrics such as number of contigs, N50 and total length for contaminated datasets relative to uncontaminated datasets. At 5% contamination, there was an increase in the number of contigs and the total length of the assembly, and a decrease in N50 (Table 2). The relative increase or decrease in these metrics varied depending on the strain used as the contaminant. For example, contamination of *L. monocytogenes* strain OLF10129 with OLF15140 appeared to have a smaller impact than contamination with the more distantly related isolate OLF09168. Statistics on all assemblies at all contamination levels are available in Supplemental Table S1.

## Contamination in SRA Data

To evaluate the prevalence of contamination in public databases, we randomly selected 500 isolates sequenced on Illumina instruments from the Sequence Read Archive (SRA) for *E. coli*, *S. enterica*, and *L. monocytogenes*. A full list of accessions can be found in Supplemental Table S4. Of the 1500 samples examined, 78 (5.27%) were determined to be contaminated by ConFindr (Table 3, Supplemental Table S4). For all species, intraspecies contamination appeared to be more prevalent than cross-species contamination.

## Confirmation of contamination in *E. coli* WGS data

In a recent WGS analysis performed in our laboratories, hybrid assemblies of *E. coli* samples led to an incorrect serotype determination (Table 4, sample 1, isolate 3). Three strains of *E. coli* from two samples

**Table 2.** Assembly metrics for intraspecies contamination dataset

| Strain 1 | Strain 2 (Contaminant) | Percent Contaminant | N50 | # Contigs | Total Length |
|---|---|---|---|---|---|
| *L. monocytogenes* (OLF10129) | NA | 0 | 338684 | 16 | 2966006 |
| | OLF13043-2 | 5 | 291474 | 47 | 3007464 |
| | | 10 | 112978 | 83 | 3073200 |
| | OLF15140 | 5 | 302686 | 26 | 2971529 |
| | | 10 | 134503 | 78 | 3008481 |
| | OLF09168 | 5 | 320971 | 57 | 2992774 |
| | | 10 | 271165 | 453 | 3333771 |
| *S.* Heidelberg (OLC2542) | NA | 0 | 693768 | 29 | 4856249 |
| | *S.* Bredeney (OLC2229) | 5 | 381278 | 34 | 4859111 |
| | | 10 | 228154 | 111 | 4936326 |
| | *S.* Typhimurium (OLF13104-7) | 5 | 235407 | 66 | 4910684 |
| | | 10 | 162612 | 162 | 5080654 |
| | *S.* Dublin (OLF18064-1) | 5 | 387491 | 56 | 4881296 |
| | | 10 | 272678 | 132 | 5022778 |
| *E. coli* O121:H19 (OLF17053-3) | NA | 0 | 134602 | 196 | 5150254 |
| | O174:H19 (OLF17021-7) | 5 | 122561 | 234 | 5183143 |
| | | 10 | 50560 | 433 | 5337747 |
| | O8:H28 (OLF17043) | 5 | 119304 | 197 | 5154985 |
| | | 10 | 31707 | 477 | 5329336 |
| | O15:H14 (OLF17030) | 5 | 121490 | 230 | 5179176 |
| | | 10 | 32121 | 509 | 5385110 |

**Table 3.** Application of ConFindr to the assessment of contamination in published genomes of *L. monocytogenes*, *S. enterica*, and *E. coli*

| Species | Intraspecies Contamination | Cross-species Contamination |
|---|---|---|
| *L. monocytogenes* | 27/500 (5.4%) | 5/500 (1%) |
| *S. enterica* | 17/500 (3.4%) | 1/500 (0.2%) |
| *E. coli* | 26/500 (5.2%) | 2/500 (0.4%) |

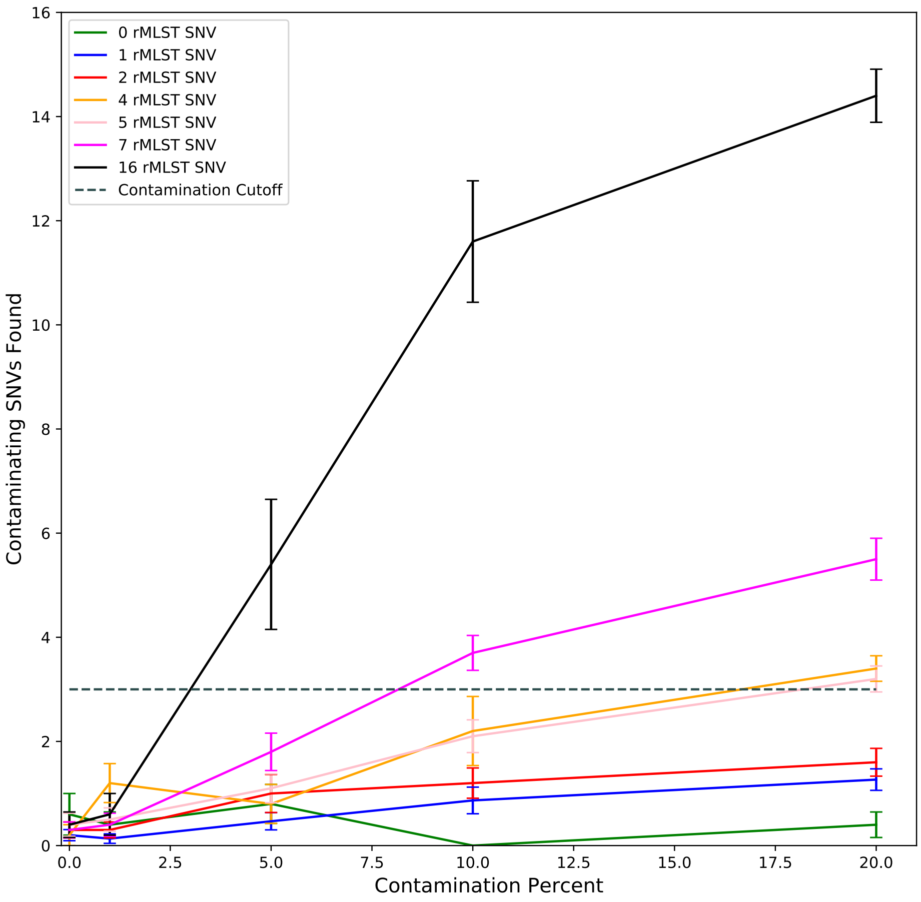

**Figure 1. Detection of contaminating SNVs within rMLST genes by ConFindr.** Sequencing reads were generated *in silico* from complete assemblies for *E. coli*, *S. enterica*, and *L. monocytogenes* and synthetic mutants containing 100, 500, 1000 and 2000 randomly-distributed SNVs. Reads were mixed to generate datasets with 0, 1, 5, 10 and 20% contamination. Datasets were binned according to the number of SNVs (0 to 16) occurring within rMLST genes in the contaminant relative to the parent strain (Supplementary Table S3. The number of contaminating SNVs identified in each dataset was plotted relative to percent contamination of the sample. Error bars indicate standard error for a minimum of 5 replicates.

were sequenced in duplicate or triplicate. Presumptive contamination was identified due to higher number of contigs or larger genome size relative to duplicates from the same sample (Table 4, bold). In one sample, the serotype of an isolate was incorrectly determined (O159:H2). Contamination was confirmed by analysis of samples with ConFindr (Table 4, bold).

## Runtime Considerations and Installation

ConFindr can be installed with a single command via bioconda (Grüning et al., 2018), and completes analysis on a sample in under one minute when using 4 threads and less than 4 GB of RAM. These features make it practical for ConFindr to be installed and run as a standard quality control step in bioinformatics pipelines.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27499v1 | CC BY 4.0 Open Access | rec: 22 Jan 2019, publ: 22 Jan 2019

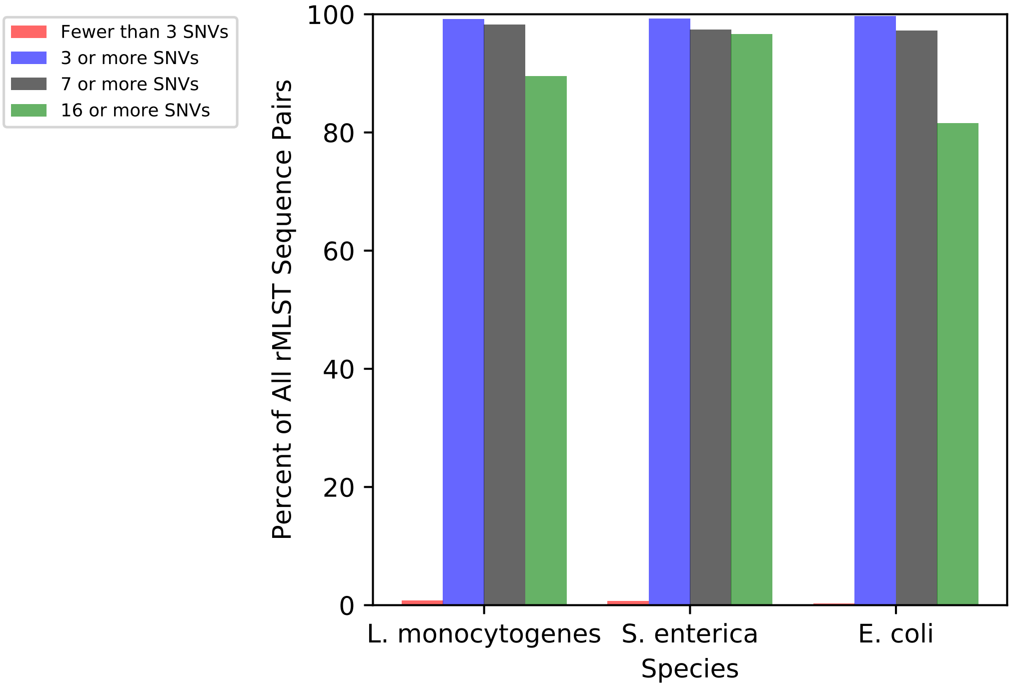

**Figure 2.** **SNV distance between all pairs of rMLST sequence types for *L. monocytogenes*, *S. enterica*, and *E. coli***

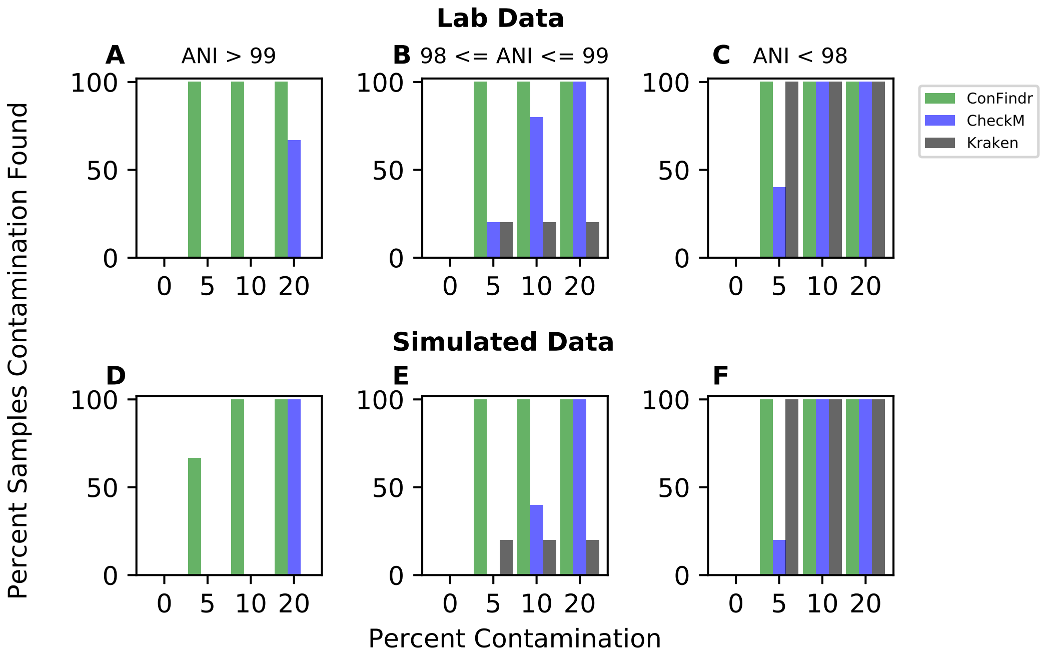

**Figure 3.** **Performance of ConFindr compared with CheckM and Kraken for closely related, related, and distantly related mixes (ANI >99, between 98 and 99, and <98, respectively).** A, B, and C show results using laboratory data, while D, E, and F show results with simulated data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27499v1 | CC BY 4.0 Open Access | rec: 22 Jan 2019, publ: 22 Jan 2019

**Table 4.** Intraspecies contamination in *E. coli*

| Sample | Isolate | Serotype | MLST | rMLST | Contigs | N50 | Genome Size | ConFindr |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | O159:H19 | 1611 | 52368 | 76 | 187212 | 5087345 | Clean |
| | 2 | O159:H19 | 1611 | 52368 | 73 | 187212 | 5085921 | Clean |
| | 3 | **O159:H2** | 1611 | 52368 | **193** | 186411 | **5500652** | **Contaminated** |
| | 4 | O83:H31 | 372 | 1854 | 31 | 436072 | 1967515 | Clean |
| | 5 | O83:H31 | 372 | 1854 | **112** | 433519 | **5077348** | **Contaminated** |
| 2 | 1 | O8:H28 | 4496 | 33427 | 55 | 218837 | 4858743 | Clean |
| | 2 | O8:H28 | 4496 | 33427 | 55 | 218837 | 4857891 | Clean |
| | 3 | O8:H28 | 4496 | 33427 | **151** | 218837 | **4985955** | **Contaminated** |

# DISCUSSION

In creating ConFindr, we wanted a tool that would be broadly applicable to the bacterial domain, while also providing enough resolution to detect contamination between closely related isolates. We selected the 53 ribosomal protein genes used in the rMLST scheme as they are present in all bacteria and provide enough diversity for high-resolution characterization (Jolley et al., 2012). While there are duplicate copies of some of the genes within the scheme (e.g., *L. monocytogenes* has two alleles of BACT000014), the scheme is actively curated, and these exceptions are known and handled by ConFindr. The advantage of this core-gene approach is that the tool can be integrated into pipelines aimed at the analysis of multiple bacterial species. While single-copy core genes have been used to evaluate the quality and completeness of metagenomic assemblies (Hess et al., 2011; Parks et al., 2015), this approach has not been commonly applied to bacterial isolate WGS data. In the current study, we found ConFindr performed equally well for three species, including both Gram-positive and Gram-negative bacteria, and we would expect similar performance for other species covered by the rMLST database. Due to its reliance on the rMLST scheme, ConFindr is intended for use in bacteria and is not for the detection of contamination in archaeal or eukaryotic samples. Nonetheless, a similar approach of using broadly-conserved core single-copy genes would likely be effective for addressing contamination within other domains.

The sensitivity of ConFindr for detection of intraspecies contamination is dependent on sequence coverage, as well as the number of SNVs occurring within the conserved ribosomal proteins genes used in the analysis. ConFindr is unable to detect contamination if the contaminating isolate has fewer than 3 SNVs within the rMLST genes used in the tool. This cut-off was chosen as ConFindr will occasionally detect one or two false positive cSNVs in rMLST genes (Figure 1), but we have yet to see an example with 3 or more false positive cSNVs. In practice, this means that ConFindr may sometimes miss contamination between two strains that have only one or two SNVs within the rMLST genes. However, our analysis of the rMLST database demonstrates that greater than 99% of the rMLST profiles for *L. monocytogenes*, *S. enterica* and *E. coli* differed by more than 3 SNVs relative to all other profiles in the database indicating that this tool would generally be effective for detection of contamination with unrelated strains (Figure 2).

The combined length of the rMLST genes is approximately 20 kilobases, representing only 0.7% of the genome in *L. monocytogenes* and 0.4% of the genome in *E. coli*. Examining this small fraction of the genome limits the sensitivity of ConFindr. This limitation could be overcome by using core-genome multi-locus sequence typing (cgMLST) schemes for species where they are available; however, doing this would increase the size of the databases used by ConFindr, increase the runtime, and would require additional manual curation of the cgMLST schemes used to ensure reliability in an automated system. Moreover, we found ConFindr to be more sensitive than CheckM for intraspecies contamination, despite the use of a smaller number of core genes relative to CheckM which uses a larger number of lineage-specific core genes. This is likely because ConFindr works at the read level while CheckM works on assemblies (Parks et al., 2015). If contamination is at a low level (e.g., one SNV in a gene, at a low contamination percentage), variant positions would likely get lost in the assembly process, limiting the sensitivity of assembly-based approaches. Furthermore, different assemblers or read preprocessing steps may change the results found by assembly-based tools for contamination detection.

In the present study, cross-species contamination was easily identified based on *de novo* assembly metrics (Supplemental Table S1); however, assemblies with low-level intraspecies contamination had assembly metrics similar to uncontaminated assemblies (Table 2). This is consistent with observations in our laboratory (Table 4). Typical assembly metrics vary among species and strains, making it difficult to develop robust standards for these metrics. For example, *S. enterica* genomes tend to have higher N50 values and assemble into fewer contigs than *E. coli* (e.g., Table 2). Ultimately, this variability makes it difficult to develop standard cutoffs that can be integrated into automated tools.

We applied ConFindr to the evaluation of 1500 samples in the public SRA repository and identified intraspecies contamination in 5.13% of the samples (Table 3). Notably, intraspecies contamination was more prevalent than cross-species contamination. A recent assessment of 67758 publically-available *Salmonella* sequences determined that 1.87% of samples had cross-species contamination based on a read classification approach (Robertson et al., 2018). Prevalence of cross-species sequence contamination in public repositories is a known issue that has been described in a number of studies (Merchant et al., 2014; Mukherjee et al., 2015; Lee et al., 2017; Cornet et al., 2018). Very few studies have looked at intraspecies contamination in public repositories, and we could not identify any studies evaluating prevalence of intraspecies contamination in foodborne pathogens. While the effects of intraspecies contamination are poorly understood, the relatively high proportion of samples determined to be contaminated by ConFindr highlights the need to further investigate the impacts of intraspecies contamination on WGS-based analyses.

WGS pipelines for public-health microbiology often include analyses of SNVs among a group of isolates to assess evolutionary relatedness and/or detection of genetic targets (e.g. serotype markers, virulence determinants) (Lambert et al., 2015; Ronholm et al., 2016; Allard et al., 2016; Chen et al., 2017). The impact of using contaminated data in these analyses is not well understood as validation schemes for bioinformatics pipelines do not often assess the effect of contamination on results of analyses to determine acceptable limits. We found one report on the development and validation of the SNVPhyl pipeline that incorporated an assessment of the impact of contamination (Petkau et al., 2017). In this evaluation, the number of SNVs detected decreased as contamination with a closely related strain increased, and detection of clusters of epidemiologically-related isolates was impacted with greater than 10% contamination (Petkau et al., 2017). While few studies of the impact of contamination on phylogenomic analyses exist, most SNV detection pipelines use cut-offs for coverage and relative nucleotide abundance at a given position to ensure validity of a SNV (Davis et al., 2015; Petkau et al., 2017). The presence of intraspecies contamination in the analysis of a sample would ultimately result in the exclusion of valid SNVs and could have impacts on the resulting phylogenetic tree topology. Contamination may have more important effects on detection of genetic markers in WGS data. For example, in a recent analysis in our laboratory, contamination impacted the accuracy of the determination of an *E. coli* serotype (Table 4). Similarly, intraspecies contamination could result in detection of virulence and antibiotic resistance genes, as well as pathogenicity islands or other horizontally acquired genes that are part of the contaminant strain and not the target strain.

# CONCLUSION

We have developed a novel bioinformatics pipeline (ConFindr) for detection of contaminating reads in raw short-read bacterial WGS data and have demonstrated its applicability for quality assessment of data derived from the priority foodborne pathogens *L. monocytogenes*, *S. enterica* and STEC. To our knowledge, this is the first automated tool developed specifically for this purpose. ConFindr outperforms existing bioinformatics tools for detection of intraspecies contamination in bacterial WGS data and can reliably detect cross-species contamination. It may be possible to adapt the approach used by ConFindr for long read data, but this is reserved for future work due to the drastic differences in error profile that may impact the identification of contaminating SNVs. ConFindr should be universally applicable to bacterial genomes and can be easily implemented in quality-control pipelines for WGS analysis. Further studies are needed to better understand the effects of contamination on WGS analyses, as well as establish what acceptable levels of contamination are when analyzing WGS data. The integration of tools such as ConFindr in quality-analysis pipelines will improve the reliability of public WGS databases, and the accuracy of downstream analyses.

ConFindr source code and documentation are freely available under the MIT Licence at https://github.com/OLC-Bioinformatics/ConFindr.

## ACKNOWLEDGMENTS

We gratefully acknowledge Julie Shay, Mohamed Elmufti and Austin Markell for helpful comments on the draft manuscript. We would also like to thank Katie Eloranta and Jennifer Liu for providing sequence data used in Table 4 of the manuscript, and users who have tested ConFindr and suggested improvements and modifications.

## REFERENCES

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., and Timme, R. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *Journal of Clinical Microbiology*, 54(8):1975–1983.

Andrews, S. (2010). Fastqc. a quality control tool for high throughput sequence data.

Ballenghien, M., Faivre, N., and Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, 15(1):25.

Bushnell, B. (2014). BBmap: A fast, accurate, splice-aware aligner.

Chen, Y., Luo, Y., Carleton, H., Timme, R., Melka, D., Muruvanda, T., Wang, C., Kastanis, G., Katz, L. S., Turner, L., Fritzinger, A., Moore, T., Stones, R., Blankenship, J., Salter, M., Parish, M., Hammack, T. S., Evans, P. S., Tarr, C. L., Allard, M. W., Strain, E. A., and Brown, E. W. (2017). Whole genome and core genome multilocus sequence typing and single nucleotide polymorphism analyses of *Listeria monocytogenes* isolates associated with an outbreak linked to cheese, united states, 2013. *Applied and Environmental Microbiology*, 83(15).

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Cornet, L., Meunier, L., Van Vlierberghe, M., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., and Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE*, 13(7):1–26.

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., and Strain, E. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Computer Science*, 1:e20.

Grüning, B., Dale, R., Sjödin, A., Chapman, B., Rowe, J., Tomkins-Tinch, C., Valieris, R., Köster, J., and Team, B. (2018). Bioconda: A sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15:475–476.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Hess, M., Sczyrba, A., Egan, R., Kim, T., Chokhawala, H., Schroth, G., Luo, S., Clark, D., Chen, F., and Zhang, T. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331:463.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., and Maiden, M. C. J. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, 158(4):1005–1015.

Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., and Harmsen, D. (2014). Gabenchtob: A genome assembly benchmark tuned on bacteria and benchtop sequencers. *PLOS ONE*, 9(9):1–12.

Koren, S., Treangen, T. J., Hill, C. M., Pop, M., and Phillippy, A. M. (2014). Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*, 15(1):126.

Lambert, D., Carrillo, C. D., Koziol, A. G., Manninger, P., and Blais, B. W. (2015). Genesippr: A rapid whole-genome approach for the identification and characterization of foodborne pathogens such as priority shiga toxigenic *Escherichia coli*. *PLOS ONE*, 10(4):1–19.

Lambert, D., Pightling, A., Griffiths, E., Van Domselaar, G., Evans, P., Berthelet, S., Craig, D., Chandry, P. S., Stones, R., Brinkman, F., Angers-Loustau, A., Kreysa, J., Tong, W., and Blais, B. (2017). Baseline practices for the application of genomic data supporting regulatory food safety. *Journal of AOAC International*, 100(3):721–731.

Lee, I., Chalita, M., Ha, S.-M., Na, S.-I., Yoon, S.-H., and Chun, J. (2017). ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16s rna gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, 67(6):2053–2057.

Lee, I., Ouk Kim, Y., Park, S.-C., and Chun, J. (2016). Orthoani: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2):1100–1103.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., and Deng, H. (2011). Comparative studies of *de novo* assembly tools for next-generation sequencing technologies. *Bioinformatics*, 27(15):2031–2037.

Mallet, L., Bitard-Feildel, T., Cerutti, F., and Chiapello, H. (2017). Phyloligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. *Bioinformatics*, 33(20):3283–3285.

Merchant, S., E Wood, D., and Salzberg, S. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, 2:e675.

Mukherjee, S., Huntemann, M., Ivanova, N., C Kyrpides, N., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina Phix control. *Standards in genomic sciences*, 10:18.

O'Leary, N., Wright, M., Brister, J., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., and Pruitt, K. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44.

Ondov, B. D., Treangen, T., Melsted, P., Mallonee, A., Bergman, N., Koren, S., and Phillippy, A. (2016). Mash: Fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17.

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236.

Page, A. J., De Silva, N., Hunt, M., Quail, M. A., Parkhill, J., Harris, S. R., Otto, T. D., and Keane, J. A. (2016). Robust high-throughput prokaryote de novo assembly and improvement pipeline for illumina data. *Microbial Genomics*, 2(8):–.

Parks, D., Imelfort, M., T Skennerton, C., Philip, H., and Tyson, G. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25.

Petkau, A., Mabon, P., Sieffert, C., C Knox, N., Cabral, J., Iskander, M., Iskander, M., Weedmark, K., Zaheer, R., Katz, L., Nadon, C., Reimer, A., Taboada, E., G Beiko, R., Hsiao, W., Brinkman, F., Graham, M., and Van Domselaar, G. (2017). SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microbial genomics*, 3:e000116.

Robertson, J., Yoshida, C., Kruczkiewicz, P., Nadon, C., Nichani, A., Taboada, E., and Nash, J. (2018). Comprehensive assessment of the quality of *Salmonella* whole genome sequence data available in public sequence databases using the Salmonella *in silico* Typing Resource (SISTR). *Microbial Genomics*, 4.

Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clinical Microbiology Reviews*, 29(4):837–857.

Souvorov, A., Agarwala, R., and Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology*, 19(1):153.

Taboada, E. N., Graham, M. R., Carriço, J. A., and Van Domselaar, G. (2017). Food safety in the age of next generation sequencing, bioinformatics, and open data access. *Frontiers in Microbiology*, 8:909.

W.A. Rossen, J., W. Friedrich, A., and Moran-Gilad, J. (2017). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clinical Microbiology and Infection*, 24.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46.