Pightling *et al. Genome Biology* (2019) 20:286
https://doi.org/10.1186/s13059-019-1914-x

Genome Biology

SHORT REPORT Open Access

Within-species contamination of bacterial whole-genome sequence data has a greater influence on clustering analyses than between-species contamination

Arthur W. Pightling*, James B. Pettengill, Yu Wang, Hugh Rand and Errol Strain

Abstract

Although it is assumed that contamination in bacterial whole-genome sequencing causes errors, the influences of contamination on clustering analyses, such as single-nucleotide polymorphism discovery, phylogenetics, and multi-locus sequencing typing, have not been quantified. By developing and analyzing 720 *Listeria monocytogenes*, *Salmonella enterica*, and *Escherichia coli* short-read datasets, we demonstrate that within-species contamination causes errors that confound clustering analyses, while between-species contamination generally does not. Contaminant reads mapping to references or becoming incorporated into chimeric sequences during assembly are the sources of those errors. Contamination sufficient to influence clustering analyses is present in public sequence databases.

Keywords: *Listeria monocytogenes*, *Salmonella enterica*, *Escherichia coli*, Whole-genome sequencing, Contamination, Phylogenetics, Single-nucleotide polymorphism, Multi-locus sequence typing, Comparative genomics, Clustering analyses, SNP, MLST

Main text

Whole-genome sequence (WGS) analysis is valuable for studying bacteria in many disciplines, including genetics, evolutionary biology, ecology, clinical microbiology, and microbial forensics [1–5]. Researchers cluster genomes with phylogenetic analyses and by counting nucleotide or allele differences. Contamination of eukaryotic data can cause misleading results [6, 7]. For prokaryotes, it is assumed that contamination causes error [8], and tools are available to detect it [9–13], but evidence supporting this assumption is lacking. To measure the influences of contamination on clustering analyses, we generated 720 sets of simulated *Listeria monocytogenes*, *Salmonella enterica*, and *Escherichia coli* Illumina MiSeq reads. These datasets include from 10 to 50% of within-species (at 0.05, 0.5, and 5% genomic distances) and between-species contamination. We also identified 24 sets of closely related bacteria (clusters) within which the contamination datasets can be analyzed. With these tools, we found that within-species contamination caused substantial errors in single-nucleotide polymorphism (SNP) and multi-locus sequence typing (MLST) pipelines, while between-species contamination resulted in fewer errors. Read mapping and assembly behavior explains this observation—reads from the same species are mapped to references or incorporated into the same contiguous sequences (contigs) as subject reads, while reads from different species usually are not.

We measured SNP and allele distances between subjects and closely related isolates ("nearest neighbors") with the CFSAN SNP Pipeline and core-genome MLST (cgMLST) workflows [14–16] (Additional file 1: Table S1). We also performed phylogenetic analyses to provide bootstrap supports for the monophyly of subjects and their nearest neighbors. Importantly, only the subject data are simulated; all other data are real (Additional file 1: Figure S1). This approach provides as realistic a dataset as possible that produces results that apply to real-world situations.

* Correspondence: Arthur.Pightling@fda.hhs.gov; apightling@gmail.com
Center for Food Safety and Applied Nutrition, US Food and Drug Administration, College Park, MD, USA

We observed increased SNP counts for all three species at 40 and 50% levels of contamination with 0.5 and 5% distant genomes (median 5–154) relative to controls (median 1–3; Fig. 1a–c, Additional file 1: Tables S2 and S3). For *S. enterica* and *E. coli*, there were smaller but significant increases at 50% contamination with 0.05% distant genomes (median 12–14) and for one of the two between-species contaminants (median 7–13). Bootstrap support at 40 and 50% levels of within-species contamination decreased for *L. monocytogenes* and *E. coli* (median 0.63–0.88 and 0.00–0.92, respectively) compared to controls (median 0.91–0.92 and 0.97), although not all decreases were significant (Fig. 1d–f). For *S. enterica*, we saw small decreases with 50% contamination by 0.05 (median 0.86) and 0.5% (median 0.96) distant genomes relative to controls (median 1.00 for each). For *L. monocytogenes* and *S. enterica*, between-species contamination caused no decreases in bootstrap support (median 0.92–0.93 and 1.00, respectively), and support only slightly decreased for *E. coli* (median 0.92–0.99). With the MLST workflows, each type of contamination influenced allele counts. Still, the 0.5 and 5% distant genomes had the greatest influence (median 3–294 and 14–418) when compared to controls (median 2–5; Fig. 2a–c, Additional file 1: Tables S2 and S3). The numbers of missing and partial alleles were also greatest for the 0.5 and 5% contaminants (median 1–463) relative to controls (median 0–6; Fig. 2d–f). Errors at lower levels for the MLST are likely due to the absence of filtering steps commonly found in SNP pipelines.

To gain insight into these results, we examined the percent of reads mapped to references. Median values were highest for 0.05 and 0.5% within-species contamination (median 96–100%) and lowest for between-species (median 50–91%), while 5% within-species contamination

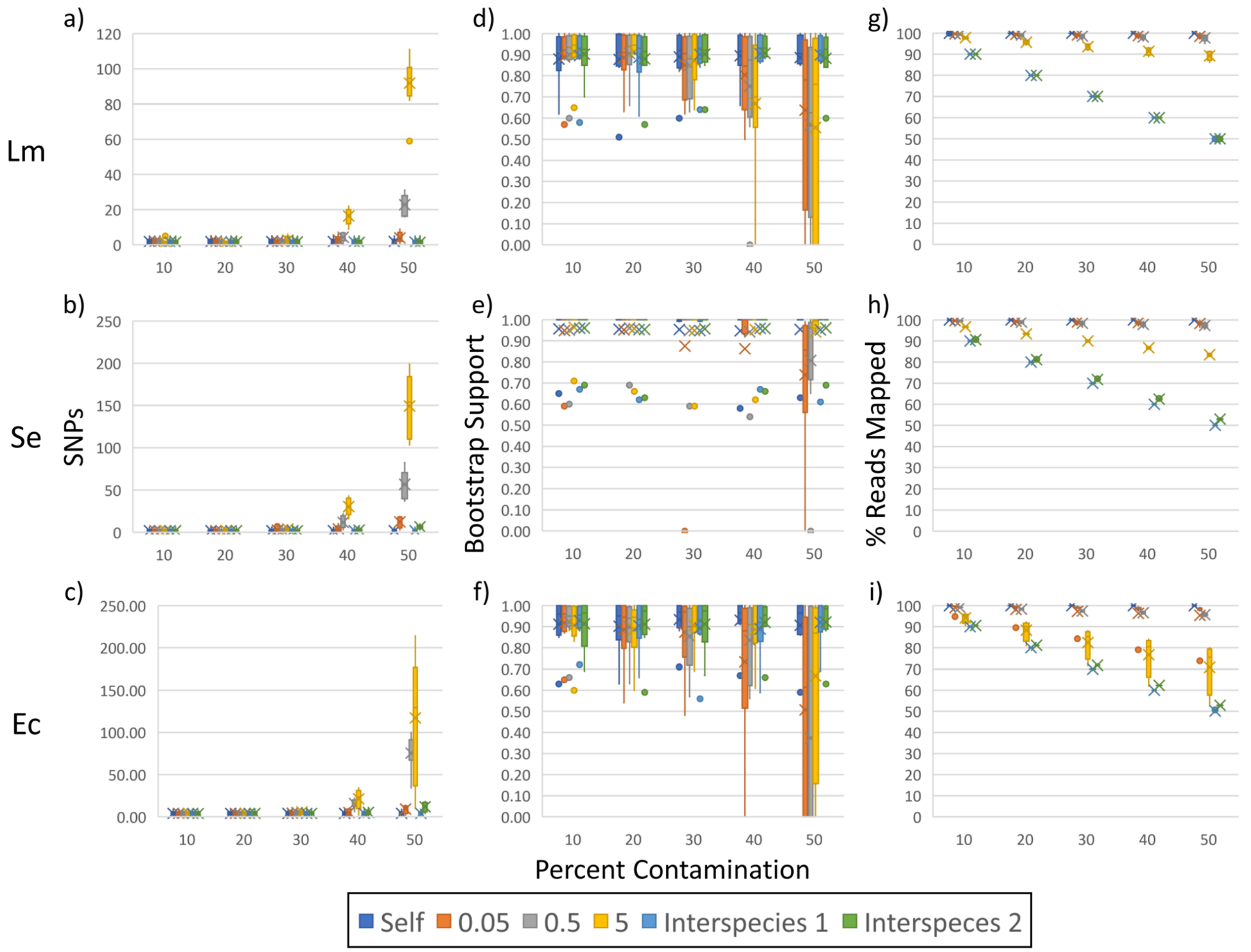

Fig. 1 Results of SNP and phylogenetic analyses for contaminated datasets. We contaminated simulated *Listeria monocytogenes* (Lm), *Salmonella enterica* (Se), and *Escherichia coli* (Ec) MiSeq data with reads from themselves as controls (Self); genomes from the same species at 0.05, 0.5, and 5% genetic distances; and genomes from different species (e.g., we contaminated Lm with Se and Ec, and we contaminated Se with Lm and Ec) at 10–50% levels. For each contamination type at each level, results for 8 datasets are shown. Panels **a**-**c** show SNP distances, **d**-**f** bootstrap supports, and **g**-**i** percent reads mapped

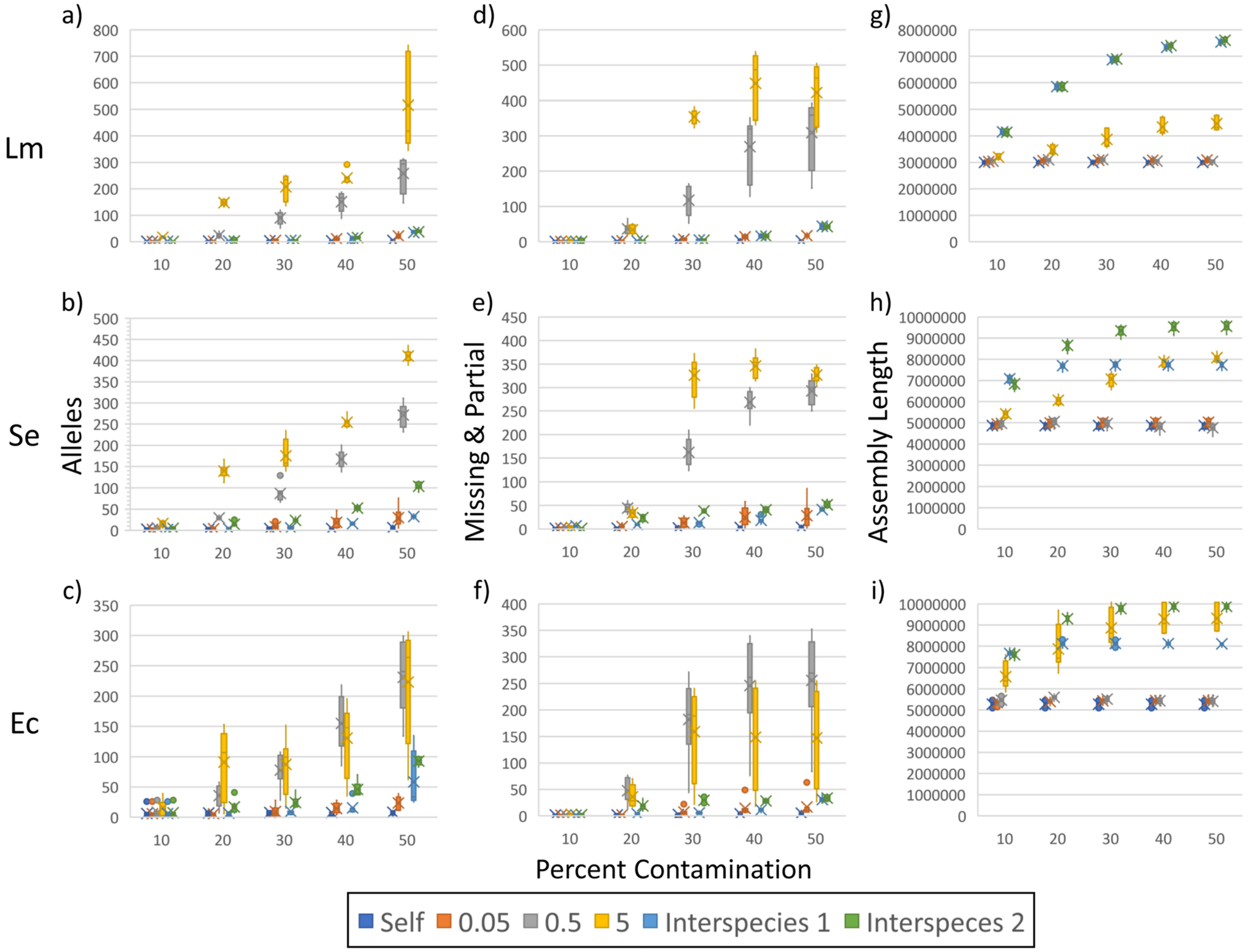

Fig. 2 Results of MLST analyses and assembly lengths for contaminated datasets. We contaminated simulated *Listeria monocytogenes* (Lm), *Salmonella enterica* (Se), and *Escherichia coli* (Ec) MiSeq data with reads from themselves as controls (Self); genomes from the same species at 0.05, 0.5, and 5% genetic distances; and genomes from different species (e.g., we contaminated Lm with Se and Ec, and we contaminated Se with Lm and Ec) at 10–50% levels. For each contamination type at each level, results for 8 datasets are shown. Panels **a**-**c** show allele counts, **d**-**f** numbers of missing and partial alleles, and **g**-**i** assembly lengths

yielded intermediate results (median 76–98%; Fig. 1g–i, Additional file 1: Tables S2 and S3). For between-species contamination, there is an inverse relationship between contamination levels and the percent of reads mapped to references. For example, at 10% contamination, approximately 90% of reads mapped. It appears that the more distant mapped contaminant reads are, the higher the SNP counts. Contaminant reads that are similar enough to the reference to be mapped but distant enough from the subject to introduce variation will generate errors. In turn, these errors may reduce bootstrap support. A similar relationship exists between allele distances and assembly lengths. Median assembly lengths for 0.05 and 0.5% within-species data are similar to controls (median 3.0–5.6 and 3.0–5.3 megabases [Mb], respectively), while between-species contaminants yielded larger assemblies (median 4.1–9.9 Mb) and the 5% within-species contamination dataset yielded intermediate assemblies (median 3.1–9.1 Mb; Fig. 2g–i).

To measure contamination in public sequence databases, we used ConFindr [13] to analyze 10,000 randomly selected fastq datasets for each of *L. monocytogenes*, *S. enterica*, and *E. coli* (Additional file 2: Table S4). We detected contamination in 8.92, 6.38, and 5.47% of the data, respectively (Additional file 1: Table S5). We detected between-species contamination (1.23, 0.29, and 0.15%) less often than within-species contamination (7.69, 6.09, and 5.33%), consistent with Low et al. [13]. We also analyzed the simulated data with ConFindr and used that information to estimate levels of contamination in the databases that may confound SNP and MLST workflows (Additional file 1: Figure S2 and Table S5). Approximately 1.48 (*L. monocytogenes*), 2.22 (*S. enterica*), and 0.87% (*E. coli*) of the data are contaminated at levels that are likely to influence SNP analyses. Roughly 2.26 (*L. monocytogenes*), 5.06 (*S. enterica*), and 1.26% (*E. coli*) of the data are contaminated at levels that may influence MLST analyses.

In summary, we show that within-species contamination (especially by 0.5 and 5% distant genomes) causes more errors in SNP counts, allele counts, and phylogenetic analyses of bacterial genomes [17] than between-species contamination. While other workflows may not yield the exact numbers measured here, the observation that contaminant reads are mapped to references and included in contigs of the same species, resulting in errors, is likely to hold. This study also shows that contamination that may cause errors in clustering analyses is present in public sequence databases. Therefore, it is important that studies include steps to detect within-species contamination.

Methods

We searched the National Center for Biotechnology Information's (NCBI's) database for closed *Listeria monocytogenes*, *Salmonella enterica*, and *Escherichia coli* genomes (e.g., "*Listeria monocytogenes*"[Organism] AND ("complete genome"[filter] AND all[filter] NOT anomalous[filter])) and downloaded all assemblies. We identified those that are 0–9 SNPs distant to other genomes ("nearest neighbors") using the "min_dist_same" and "min_dist_opp" measurements in the NCBI metadata files [18–20]. We used the NCBI's Isolates Browser [21] to identify closed genomes with closely related isolates that are part of NCBI SNP trees with at least 5 taxa [22]. We assembled 16,839 *L. monocytogenes*, 127,357 *S. enterica*, and 33,821 *Escherichia coli* Illumina datasets with SPAdes v3.12.0 (spades.py --careful -1 forward.fastq -2 reverse.fastq) [23]. We removed contigs that were less than 500 nucleotides. We aligned closed and draft assemblies with NUCmer v3.1 (nucmer --prefix=ref_qry closed.fna draft.fna) and estimated SNP distances with show-snps (show-snps -Clr ref_qry.delta > ref_qry.snps) [24]. We selected closed genomes for further analyses that are approximately 0.05, 0.5, and 5% from draft genomes of the same species (based upon closed assembly length estimates calculated with QUASTv4.5 [25]). For most subjects, within-species contamination represents (i) closely related genomes of the same serotype and clonal complex, with 0–2 locus differences (average 0.22; as measured with the program mlst; 0.05%) [26–28]; (ii) distantly related genomes of the same serotype but different clonal complex and 2–6 locus variants (average 4.1; 0.5%); and (iii) genomes of a different serotype and clonal complex with 7 locus variants (average 7; 5%; Additional file 1: Table S1). When unavailable, we predicted serotypes for *S. enterica* with SeqSero [29] and *E. coli* with SerotypeFinder [30]. We generated simulated reads using closed subject assemblies, within-species draft contaminant assemblies, and between-species draft contaminant assemblies, with ART_Illumina v2.5.8 (art_illumina -ss MSv1 -i assembly.fasta -p -l 230 -f 20 -m 295 -s 10 -o paired_data) [31]—all assemblies were generated from real sequencing data. Contamination fastq files were made by randomly selecting subject and contaminant reads at indicated levels (in this case 10–50% contamination) and combining them into paired read files with 20-fold depth of coverage (github.com/apightling/contamination; e.g., select_reads.pl subject_1.fq subject_2.fq 10 contaminant_1.fq contaminant_2.fq output_prefix).

We identified SNP clusters that contain subject genome sequences with the NCBI's Isolates Browser. If SNP clusters had more than 20 taxa, counting the subjects and their nearest neighbors, we randomly selected subsets for further analyses. We also ensured that the subjects and nearest neighbors formed monophyletic groups in phylogenetic trees. We generated SNP matrices with the CFSAN SNP Pipeline v1.0, using the subject assembly as a reference to minimize errors [32]. Alignments of SNPs that were detected by mapping reads to the reference were phylogenetically analyzed with GARLI v2.01.1067 [33] (100 replicates, K80 and HKY). We reported supports for monophyly of subjects and nearest neighbors; if the they were no longer monophyletic, we recorded a support of 0.

We assembled simulated data with SPAdes v3.12.0 and measured assembly statistics with QUAST v4.5. We analyzed *Listeria monocytogenes* assemblies with the LmCGST core-genome multi-locus sequence typing (cgMLST) tool and *Salmonella enterica* assemblies with an *S. enterica* cgMLST tool described in Pettengill et al. [15]. We analyzed *E. coli* assemblies with a cgMLST developed using the same approach. Partial alleles are those loci whose lengths are less than 60% of the predicted lengths, and missing alleles are those loci that are less than 60% of predicted lengths and less than 80% identical to the reference.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1914-x.

Additional file 1: Figure S1. Phylogenetic tree of 9 *Listeria monocytogenes* genomes with study subject and nearest neighbor labeled. **Figure S2.** Results of ConFindr analysis of contamination datasets generated for this study. **Table S1.** Contextual information for genome sequences used for this study. **Table S2.** Results of SNP pipeline and core-genome multi locus sequence typing analyses. **Table S3.** *P*-values for results of clustering analyses. **Table S5.** Percent of contamination detected in data from NCBI. **Table S6.** NCBI accession numbers for data generated during this study.

Additional file 2: Table S4. ConFindr results from analysis of 10,000 *Listeria monocytogenes, Salmonella enterica,* and *Escherichia coli* fastq datasets. (XLS 7913 kb)

Additional file 3. Review history.

Acknowledgements

We thank Maria Balkey for depositing the sequence data to NCBI.

Review history

The review history is available as Additional file 3.

Peer review information

Kevin Pang was the primary handling editor for this manuscript and handled its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

AWP, JBP, YW, HR, and ES conceived and designed the experiment. AWP and JBP performed the analyses. AWP, JBP, YW, HR, and ES wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by the US Food and Drug Administration's Center for Food Safety and Applied Nutrition.

Availability of data and materials

The datasets generated and analyzed during the current study are available in the NCBI repository with BioProject number PRJNA561589 [34]. Accession numbers are listed in Additional file 1: Table S6. The datasets are also available at figshare with DOI https://doi.org/10.6084/m9.figshare.c.4282706.v1 [35]. The scripts made for this study are available in GitHub (https://github.com/apightling/contamination) [36]. An archival version is available at Zenodo with DOI https://doi.org/10.5281/zenodo.3552954 [37].

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: 22 April 2019 Accepted: 6 December 2019
Published online: 18 December 2019

References

1. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601–12.
2. Lekunberri I, Gasol JM, Acinas SG, Gomez-Consarnau L, Crespo BG, Casamayor EO, Massana R, Pedros-Alio C, Pinhassi J. The phylogenetic and ecological context of cultured and whole genome-sequenced planktonic bacteria from the coastal NW Mediterranean Sea. Syst Appl Microbiol. 2014;37(3):216–28.
3. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol. 2014;52(7):2365–70.
4. van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin Microbiol Rev. 2001;14(3):547–60.
5. Wilson MR, Allard MW, Brown EW. The forensic analysis of foodborne bacterial pathogens in the age of whole-genome sequencing. Cladistics. 2013;29:449–4610.
6. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8(1):61–5.
7. Dittami SM, Corre E. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp Saccharina japonica using Taxoblast. PeerJ. 2017;5:e4073.
8. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci. 2015;10:18.
9. Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. Int J Syst Evol Microbiol. 2017;67(6):2053–7.
10. Olson ND, Zook JM, Morrow JB, Lin NJ. Challenging a bioinformatic tool's ability to detect microbial contaminants using in silico whole genome sequencing data. PeerJ. 2017;5:e3729.
11. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55.
12. Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova N, Woyke T, Kyrpides N, et al. ProDeGe: a computational protocol for fully automated decontamination of genomes. ISME J. 2016;10(1):269–72.
13. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ. 2019;7:e6995.
14. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Science. 2015;1:e20.
15. Pettengill JB, Pightling AW, Baugher JD, Rand H, Strain E. Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of *Salmonella* samples. PLoS One. 2016;11(11):e0166162.
16. Pightling AW, Petronella N, Pagotto F. The *Listeria monocytogenes* Core-Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. BMC Microbiol. 2015;15:224.
17. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Front Microbiol. 2018;9:1482.
18. Index of /pathogen/Results/Listeria/. ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Listeria/. Accessed July 2018.
19. Index of /pathogen/Results/Salmonella/. ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Salmonella/. Accessed July 2018.
20. Index of /pathogen/Results/Escherichia_coli_Shigella/. ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Escherichia_coli_Shigella/. Accessed July 2018.
21. Isolates Browser [ncbi.nlm.nih.gov/pathogens/isolates#/search/?newtree=on]. Accessed July 2018.
22. Methods description for ftp://ncbi.nlm.nih.gov/pathogen/ [ftp.ncbi.nlm.nih.gov/pathogen/Methods.txt]. Accessed July 2018.
23. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
24. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
26. PubMLST [https://pubmlst.org/]. Accessed Sept 2019.
27. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595.
28. mlst [https://github.com/tseemann/mlst]. Accessed Sept 2019.
29. 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.
30. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J Clin Microbiol. 2015;53(8):2410–26.
31. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.
32. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One. 2014;9(8):e104579.
33. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin; 2006.
34. Pightling A, Pettengill J, Wang Y, Rand H, Strain E: *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli* raw sequence reads. NCBI SRA. 2019. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA561589.

35. Pightling A, Pettengill J, Wang Y, Rand H, Strain E. Contamination Dataset 191125. Figshare. 2019; https://doi.org/10.6084/m9.figshare.c.4282706.v1.
36. Pightling A, Pettengill J, Wang Y, Rand H, Strain E. apightling/contamination. Github. 2019; https://github.com/apightling/contamination.
37. Pightling A, Pettengill J, Wang Y, Rand H, Strain E. apightling/contamination. Zenodo. 2019; https://doi.org/10.5281/zenodo.3552954.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.