Method for evaluating genomic material purity using whole genome sequencing data.

Nathan D. Olson[1], **Justin Zook**[1], **Jayne Morrow**[1], **and Nancy Lin**[1]

[1]**Material Measurement Laboratory, National Institute of Standards and Technology**

ABSTRACT

Dummy abstract text.

Keywords: Biodetection, Test material, Reference material, Purity, Bioinformatics

INTRODUCTION

Rapid, sensitive and accurate assays for detecting bacterial pathogens in food, water, clinical samples, and suspicious biothreats are critical to public health and safety. Biodetection assays must be evaluated for assay sensitivity and specificity prior to deployment and then in the hands of the user to instill confidence in the actions made based on assay results (Ieven et al., 2013; Coates et al., 2011; EPA, 2004; ISO/TS, 2010; Guide, 1998; Feldsine et al., 2002). Test materials are used to validate assay performance. Test materials can be either purified cultures, genomic DNA or whole cells spiked into a matrix (EPA, 2004; ISO/TS, 2010; CLSI, 2010). Before being used to evaluate a biodetection assay, the material itself must be validated in terms of purity and identity to eliminate false positive results due to test material contaminants or false negatives due to the test material being the wrong strain (CLSI, 2010). There are a number of potential sources of microbial contaminants of test materials including the stock culture, preservation medium, and airborne and laboratory contaminants (Marron et al., 2013; Shrestha et al., 2013; Tanner et al., 1998).

Currently polymerase chain reaction (PCR) assays are the most commonly use method for evaluating test material purity. Other methods to detect contaminants using whole genome sequencing datasets have been developed, but they are not currently used to evaluate test material purity. A PCR assay was developed to analyze protist cultures. This assay uses endpoint PCR for prokaryotes and eukaryotes with template dilutions (Marron et al., 2013). The benefit to PCR-based approaches is that they can be cost effective and fast if an applicable protocol exists. While PCR assays can detect low levels of contaminants, this approach does not easily scale to multiple contaminants and test materials. More importantly, PCR assays can only target specific contaminants, which biases the purity assessment to known potential contaminants. The bioinformatics tools developed to identify genomic contaminants in metagenomic datasets, which include sequencing data from all organisms in a sample, can also be used to evaluate test material purity. For example DeconSeq (Schmieder and Edwards, 2011a) and a similar method QC-Chain (Zhou et al., 2013) were developed to identify contaminants based on analysis of 16S ribosomal ribonucleic acid (rRNA) gene sequences or comparison of a subset of reads to a reference database using Basic Local Alignment Search Tool (BLAST). Metagonomic-based methods are ideally able to identify contaminants without any prior knowledge or assumptions regarding the identity of the organism(s). However, methods based on 16S rRNA gene identification have limited resolution, as 16S rRNA sequences can only provide genus level taxonomic resolution at best. The benefit to using metagenomic tools developed for 16S rRNA is that prior knowledge of the identity of the contaminant is not required; however, this method is unable to identify contaminants to the species level or higher.

Another approach to evaluating test material purity is through shotgun whole genome sequencing, i.e., sequence all DNA in a purportedly single organism sample. There are a number of metagenomic read classification algorithms developed to determine the taxonomic composition of a sequence dataset of unknown composition. These algorithms tend to use one of three primary strategies for taxonomic assignment. The first method consist of aligning reads to a reference database that contains assemblies of microbial genomes (Buchfink et al., 2015; Francis et al., 2013). This approach, while exaustive, is computationally expensive. The second type of method focuses on marker genes, genes common to different phylogenetic groups, which reduces the computational cost (Segata et al., 2012; Liu et al., 2011). The disadvantage of using only marker genes is that information required to discriminate closely related genomes may not be present in the marker genes. The third method uses a *k*-mer based approach, where taxonomic composition is determined based the abundance of DNA sequences of length *k* in the sequence dataset and a reference database (Ounit et al., 2015; Menzel et al., 2016; Wood and Salzberg, 2014).

In this work, we present the results of a proof of concept study to measure the purity of single organism test materials using whole genome sequencing data combined with a metagenomic read classification algorithm. We choose to use *Pathoscope*, a method that aligns sequences to a database of genome assemblies. It was developed to detect pathogens and identify strains using whole genome sequencing data (Francis et al., 2013). *Pathoscope* benefits from the large sample size obtained using all sequence data for higher sensitivity (compared to marker gene based methods) and leverages algorithmic advances for whole genome sequence mapping. We will first present the specificity of the method using simulated data for single organisms. Then, we evaluate sensitivity of the method using simulated contaminanted test material datasets.

METHODS

Simulated whole gnome sequence data was used to evaluat the suitability of using whole genome sequence data and metagenomic taxonomic classification methods for validating test material purity. Simulated data from single genomes was used to assess method specificity. To assess method sensitivity, contaminant datasets comprised of pairwise combinations of single genomes spiked with a defined proportion of contaminant reads, reads simulated from a different genome.

To best approximate real sequencing data reads with simulated using empirically determined error model and insert size distributions. The whole genome sequencing data was simulated using the ART sequencing read simulator as the algorithm using an empirical sequencing error model (Huang et al., 2012). Reads were simulated with ART simulator using the Illumina MiSeq error model for 2 × 230 base pair paired end reads with an insert size of 690 ± 10 base pairs(average ± standard deviation) and 20 X mean coverage for each strain. The insert size parameters were defined based on the observed average and standard deviation insert size of the NIST RM8375-MG002 MiSeq sequencing data (**?**).

The taxonomic composition of simulated datasest was assessed using the Pathoscope metagenomic taxonomic classifier (Francis et al., 2013). This method was selected as it combines the use of a large reference database reducing potential biases due to contaminant seqeunces not present in the database and efficient whole genome read mapping algorithms. This method uses an expectation maximization algorithm where the sequence data are first mapped to a database comprised of all sequence data in the Genbank nt database. Then, through an iterative process, it re-assigns ambiguously mapped reads based on the proportion of reads mapped unambigously to individual taxa in the database. The PathoScope 2.0 taxonomic read classification pipeline has three steps; (1) PathoQC - read quality filtering and trimming using the PRINSEQ algorithm (Schmieder and Edwards, 2011b), (2) PathoMap - mapping reads to a reference database using the bowtie2 algorithm (Langmead and Salzberg, 2012), (3) PathoID - expectation-maximization classification algorithm. The annotated Genbank nt database provided by the PathoScope developers was used as the reference database (`ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz`).

Specificity

Method specificity was first assessed to characterize the baseline accuracy of the read classifier. Method specificity was defined as the proportion of reads in a single organism simulated dataset incorrectly assigned to a taxonomy different from the test material taxonomy. Sequence data was simulated for 406 strains, from 9 genera (Table 1). We will refer to the genome used to generate the reads as the target genome. The genomes included in the simulation study were limited to the number of closed genomes in the Genbank database (`http://www.ncbi.nlm.nih.gov/genbank/`, accessed 10/18/2013) belonging to the genera of interest (Table 1). Due to the large number of closed genomes from the genera *Bacillus*, *Escherichia*, and *Salmonella*, these genera were limited to the species *Bacillus cereus*, *Escherichia coli*, and *Salmonella enterica*. The taxononomic heirarchy for the target genome and simulated read assignment match levels were determined using the R package (Scott Chamberlain and Eduard Szocs, 2013; Chamberlain et al., 2016).

Sensitivity

Method sensitivity was assessed using simulated contaminated datasets to evaluate at how well the method is able to detect genomic contaminants at a range of contaminat concentrations. Representative genomes for 8 of the 9 genus were used to generate the simulated contaminant datasets (Table 2). An *Escherichia coli* strain was selected as a representative of both and *Shigella* as the genus *Shigella* phylogenetically resides within the species *Eschericha coli* (Lan and Reeves, 2002). For each pairwise combination of representative genomes the simulated contaminant dataset was comprised of a randomly selected subset of reads from the target and contaminant simulated single genome sequence dataset. The simulated datasets were subsampled at defined proportions with p representing the proportion of reads from the contaminant single genome dataset subsampled and $1 - p$ the proportion of reads from the target genome simulated dataset. A 10 fold range of contaminant proportions were simulated with p ranging from 0.1 to 10^{-8}, resulting in 512 simulated contaminant datasets. This approach simulates the proportions of cells in a test material and not the amount of DNA, assuming unbiased DNA extraction. To generate the simulated contaminant datasets single organism simulated datasets were first generated for the 8 representative genomes using the same methods as used in the first part of the study. The resulting simulated sequencing data was first processed using PathoQC and PathoMap steps in the PathoScope pipeline. The output from the PathoMap step (sam file, sequence alignment file `https://samtools.github.io/hts-specs/SAMv1.pdf`) for the target and contaminant datasets were subsampled and combined and the resulting sam file was processed by PathoID, the third step in the PathoScope pipeline. Subsampling the sam files instead of the simulated sequence files greatly reduces the computational cost of the analysis as the simulated reads were only processed by the first two steps in PathoScope pipeline rather then for every simulated contaminant dataset.

Reproducibility

The seed number for the random number generator was randomly assigned and recorded for each dataset.

To facilitate reproducibility and transparency, a Docker (`www.docker.com`) container is available with installed pipeline dependencies (`www.registry.hub.docker.com/u/natedolson/docker-pathoscope/`). The script used to run the simulations are available at `https://github.com/nate-d-olson/genomic_purity`. Pathoscope results were processed using the statistical programing language R (R Core Team, 2016), and intermediate analysis and data summaries were organized using ProjectTemplate (White, 2014) and archived in a github repository (`https://github.com/nate-d-olson/genomic_purity_analysis`).

RESULTS AND DISCUSSION

Specificity

Simulated sequence data from individual isolates was used to assess the genomic purity assessment method specificity. Here we use specificity as a measure of the ability of the method to correctly assign reads to the taxonomy of the genome the sequencing reads were simulated from, the target genome. True negatives (TNs) are reads assigned to the target genome's species, genus, family, ect., depending on the match stringency, and false positives (FPs) are reads incorrectly assigned to a different species, genus, family, ect., and specificity = TN/(FP+TN). Method specificity was evaluated by characterizing the read assignment results based on the level of agreement between the genome and assigned taxonomy (Fig. 1). Overall high proportion of matches at species and genus level. Some genus have low specificity at the species and genus levels. For *Shigella* most likely due to matches with *Escherichia* (Fig. 2). The cumulative match proportions do not always reach 1.00, for example *Staphylococcus* genomes. This might be due to exclusion of unclassified and unknown matches (NCBI taxid 12908 and 0 respectively) from match level analysis.

Most of the genera had genus level or higher match proportions excluding a few outliers (Fig. 3). *Escherichia*, *Shigella*, and *Staphylococcus* are notable exceptions. As discussed previously the taxonomic ambiguities for *Shigella* and *Escherichia* are responsible for the overall lower genus level match proportions. Another example of low genus level matches is the *Bacillus* genome with genus match proportion

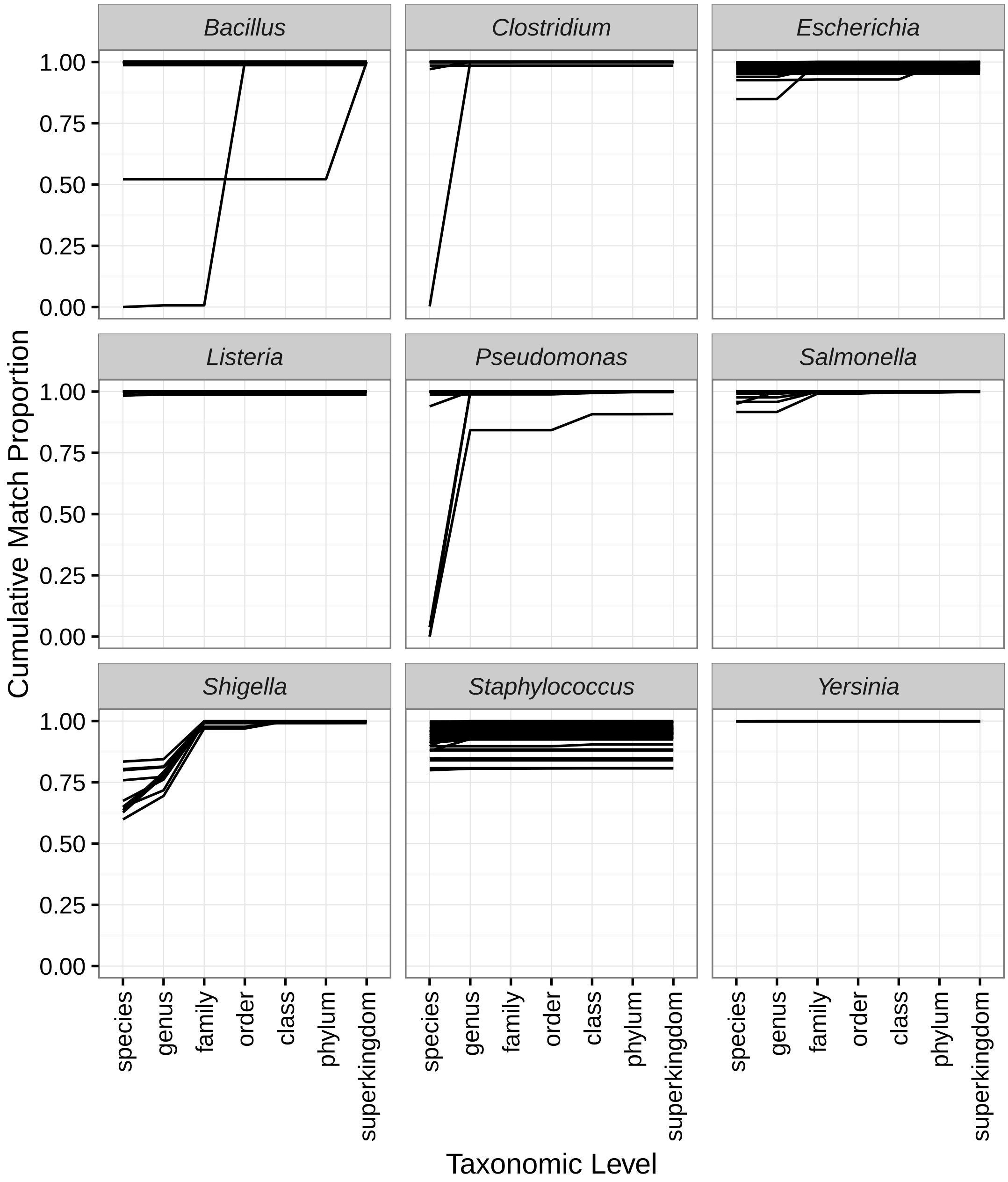

Figure 1. Cumulative taxonomic match results for genomic purity assessments of simulated sequence data from single genomes. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes are grouped by genus.

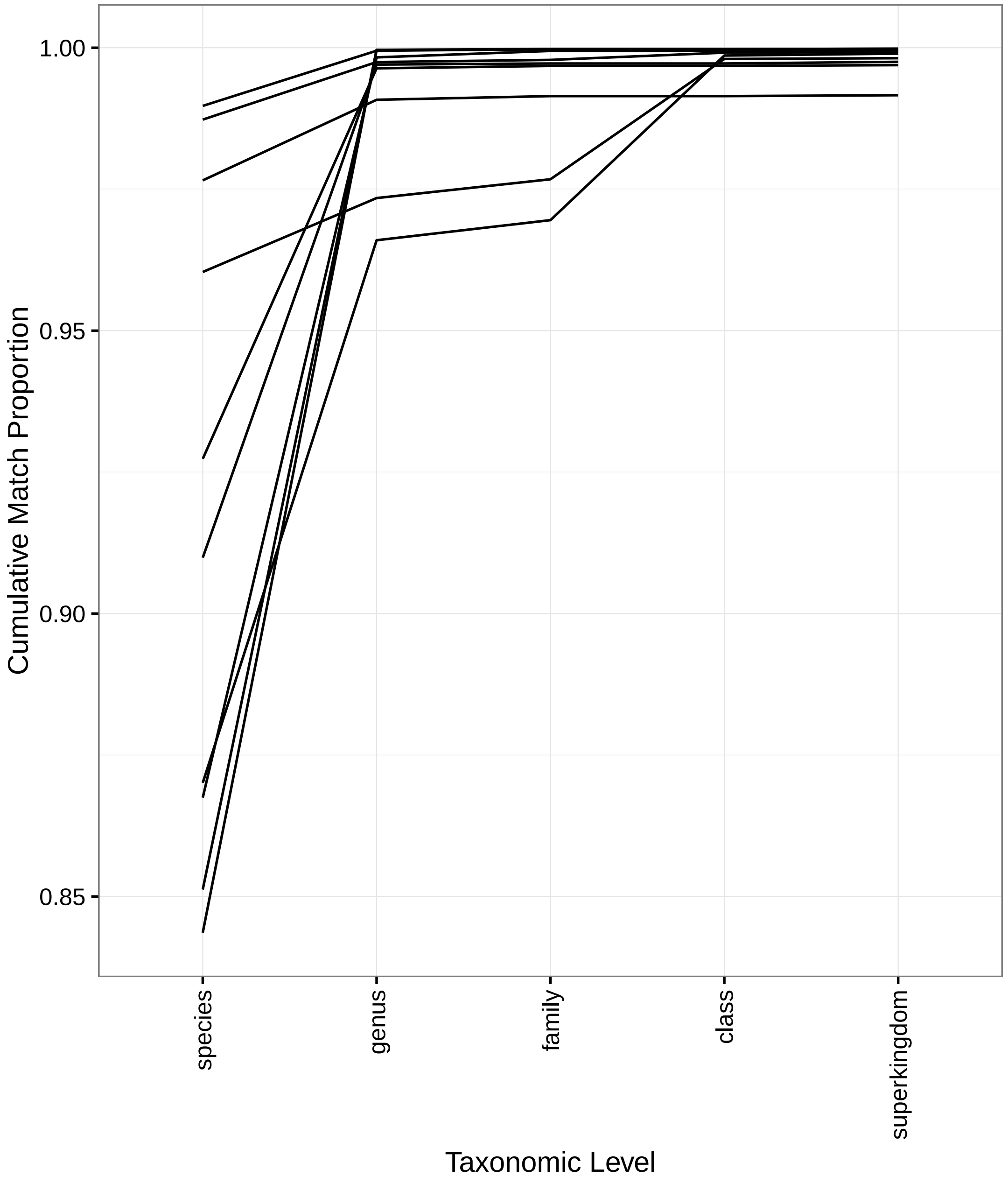

Figure 2. Cumulative taxonomic match results for genomic purity assessment for *Shigella* considering matches to *E. coli* as species level matches. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes are grouped by genus.

Genus	N	Genome Size (Mb)
Bacillus	76	5.05 (3.07-7.59)
Escherichia	62	5.11 (3.98-5.86)
Pseudomonas	57	6.18 (4.17-7.01)
Staphylococcus	49	2.82 (2.69-3.08)
Salmonella	44	4.88 (4.46-5.27)
Listeria	39	2.97 (2.78-3.11)
Clostridium	32	4.02 (2.55-6.67)
Yersinia	19	4.73 (4.62-4.94)
Francisella	18	1.89 (1.85-2.05)
Shigella	10	4.74 (4.48-5.22)

Table 1. Breakdown of the number of genomes by genus used to generate single genome simultated datasets. N indicates the number of genomes, and Genome Size is presented as the median and range (minimum to maximum) genome size

close to zero, *Bacillus infantis* string NRRL B 14911. While the *B. infantis* strain was originally classified as *Bacillus* the species is phylogenetically distinct from other members of the genus (Ko et al., 2006). It is important to consider the strain and genome being characterized, as taxonomic ambiguities (e.g. *Shigella* and *Escherichia*) can lead to lower than expected specificity and the identification of false positive contaminants.

Sensitivity

To evaluate genomic purity assessment methods we generated simulated contaminant datasets as pairwise combinations of representative genomes from 8 of the genera used in the specificity section of the study (Table 2). Due to the overall high proportion of reads matched to the correct genome in the method specificity study, the simulated contaminant datasets were evaluated at the genus level for sensitivity. For all of the genomes selected for the sensitivity study, the proportion of simulated reads that matched at species level or higher was 0.98 (Table 2).

Representative Strain	Species	C Mb	C Acc	P Mb	P Acc
Bacillus anthracis str. Ames	1.00	5.23	AE016879.1		
Clostridium botulinum A str. Hall	1.00	3.76	CP000727.1		
Escherichia coli O157:H7 str. EC4115	0.98	5.57	CP001164.1	0.13	CP001163.1, CP001165.1
Francisella tularensis subsp. tularensis SCHU S4	1.00	1.89	AJ749949.2		
Pseudomonas aeruginosa PAO1	1.00	6.26	AE004091.2		
Salmonella enterica subsp. enterica serovar Typhimurium str. D23580	1.00	4.88	FN424405.1		
Staphylococcus aureus subsp. aureus ED133	0.98	2.83	CP001996.1		
Yersinia pestis CO92	1.00	4.65	AL590842.1	0.18	AL109969.1, AL117189.1, AL117211.1

Table 2. Representative strains used in simulated contaminant datasets. Species indicates the proportion of simulated reads assigned to the correct taxa at the species level or higher. DNA size (Mb) and Genbank accession numbers (Acc) are indicated for chromosomes (C) and plasmids (P). Escherichia coli O157:H7 str. EC4115 and Yersinia pestis CO92 have two and three plasmids respectively.

To evaluate sensitivity, we plot the proportion of reads assigned to the contaminant genus or species versus the proportion of reads simulated from the contaminating genome. While the proportion of contaminant reads in the simulated datasets was not equal to the defined contaminant proportion, the proportion of reads assigned to the contaminant genus was comparable to the expected proportion (Fig. 4). This was especially true for datasets containing mixtures of *B. anthracis*, *Y. pestis*, *E. coli*, and *S. enteria* as they had similar sized genomes (Table 2). Three contaminants were detected when spiked in at contaminant proportions of 10^{-8}, *B. anthracis* in *E. coli* as well *S. enteria* and *E. coli* in *Y. pestis*. Interestingly the proportion of assigned reads did not decrease with decreasing contaminant proportions after 10^{-4}.

The lowest detectable proportion of simulated contaminant level varied by both contaminant and target genome. All organisms had comparable minimum contamination levels for which reads were assigned to the contaminant genome. Two notable exceptions are *Escherichia* and *Yersinia*, where *Bacillus*, and *Salmonella* and *Escherichia* were detected at the lowest contaminant levels respectively. As the re-

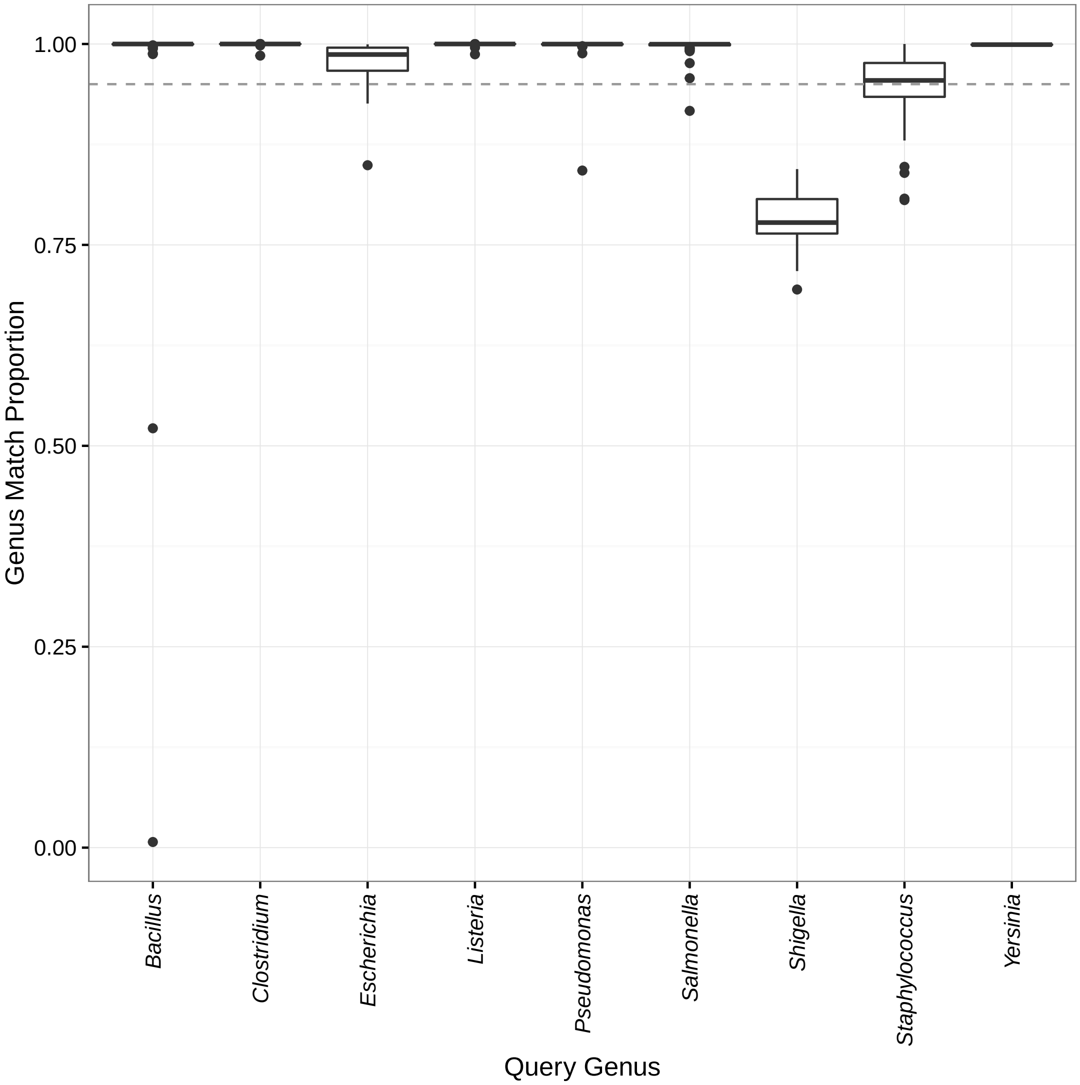

Figure 3. Distribution of the proportion of reads assigned to the source genome at or above the genus level. Horizontal grey line highlights a match proportion of 0.95. Boxplots hinges represent the 25th and 75th percentiles, line through box represent is the median, whiskers are the 95% confidence interval, and the black dots are outliers.

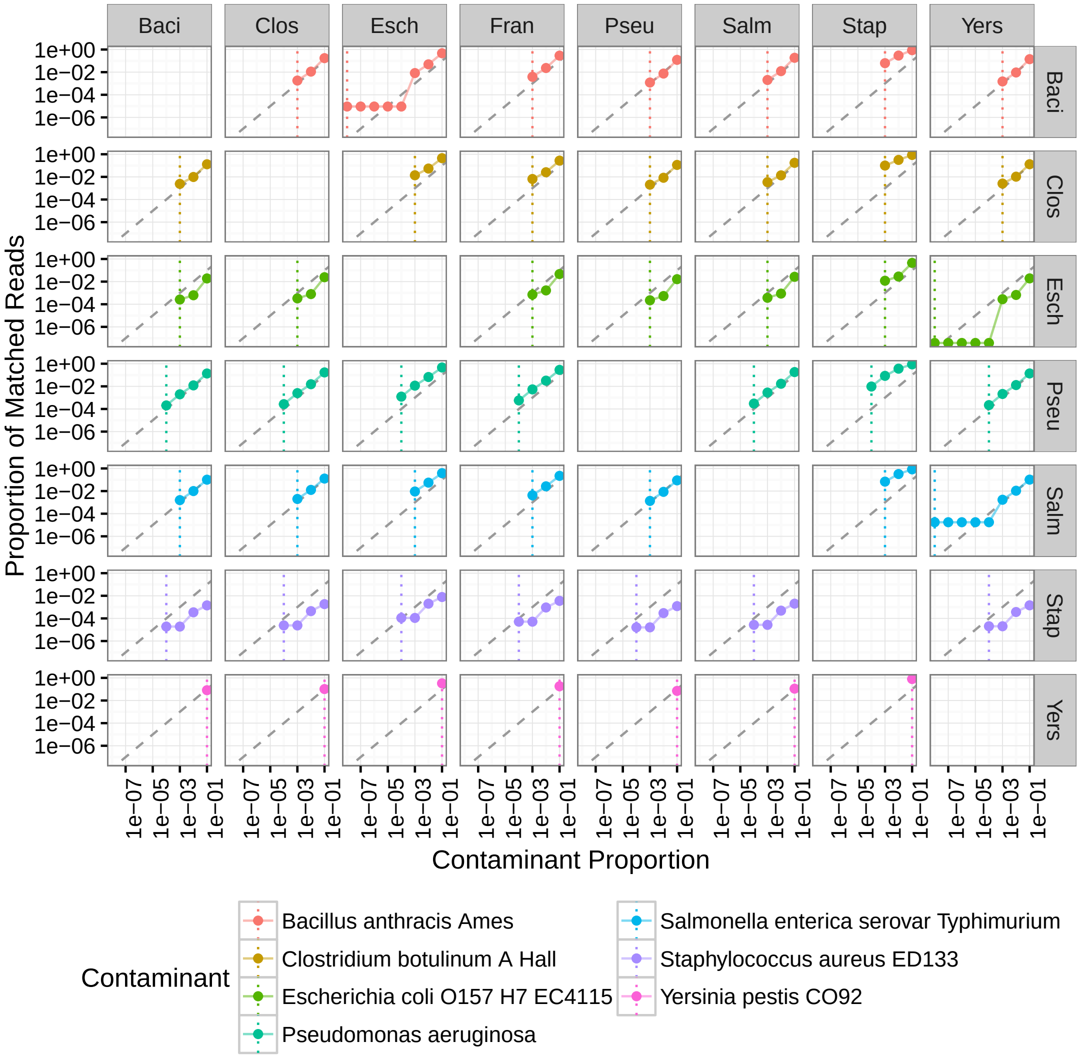

Figure 4. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

sults are from simulated data and based on proportions of simulated reads, these values do not indicate a limit of detection for the method.

CONCLUSIONS

- Proof of concept study additional work required to validate use in assessing the purity of a test material.
- Use of other taxonomic classification methods are likely to have different sensitivity and specificity results.
- Need to evaluate the suitability of the reference database for used the genome and contaminant of interest.
- Work to further expand the taxonomic database to include genomes from uncultured organism using either metagenome datasets for single cell datasets along with efforts to address issues related to taxonomic ambiguities will help to improve the method applicability.

ACKNOWLEDGMENTS

The authors would like to thanks Dr. Steven Lund for his assistance in developing the study. The Department of Homeland Security (DHS) Science and Technology Directorate supported this work under the Interagency Agreement HSHQPM-12-X-00078 with the National Institute of Standards and Technology (NIST). Opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of DHS, NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendations or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

REFERENCES

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60.

Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., Foster, Z., and O'Donnell, J. (2016). *taxize: Taxonomic information from around the web*. R package version 0.7.4.

CLSI (2010). Characterization and Qualification of Commutable Reference Materials for Laboratory Medicine ; Approved Guideline. Technical Report 22.

Coates, S. G., Brunelle, S. L., and Davenport, M. G. (2011). Development of standard method performance requirements for biological threat agent detection methods. *Journal of AOAC International*, 94(4):1328–37.

EPA (2004). Quality Assurance/Quality Control Guidance for Laboratories Performing PCR Analyses on Environmental Samples October 2004. *October*, (October).

Feldsine, P., Abeyta, C., and Andrews, W. (2002). AOAC international methods committee guidelines for validation of qualitative and quantiative food microbiological official methods of analysis. Technical Report May.

Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. a., and Johnson, W. E. (2013). Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome research*.

Guide, E. (1998). The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics. Technical report.

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

Ieven, M., Finch, R., and van Belkum, a. (2013). European quality clearance of new microbiological diagnostics. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(1):29–38.

ISO/TS (2010). Microbiology of food and animal feeding stuffs - specific requirements and quidance for proficiency testing by interlaboratory comparison. Technical report.

Ko, K. S., Oh, W. S., Lee, M. Y., Lee, J. H., Lee, H., Peck, K. R., Lee, N. Y., and Song, J.-H. (2006). Bacillus infantis sp. nov. and bacillus idriensis sp. nov., isolated from a patient with neonatal sepsis. *International journal of systematic and evolutionary microbiology*, 56(11):2541–2544.

Lan, R. and Reeves, P. R. (2002). Escherichia coli in disguise: molecular origins of shigella. *Microbes and infection*, 4(11):1125–1132.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12(1):1.

Marron, A. O., Akam, M., and Walker, G. (2013). A Duplex PCR-Based Assay for Measuring the Amount of Bacterial Contamination in a Nucleic Acid Extract from a Culture of Free-Living Protists. *PloS one*, 8(4):e61732.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7.

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):1.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schmieder, R. and Edwards, R. (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3):e17288.

Schmieder, R. and Edwards, R. (2011b). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

Scott Chamberlain and Eduard Szocs (2013). taxize - taxonomic search and retrieval in r. *F1000Research*.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814.

Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When Is a Microbial Culture Pure ? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing.

Tanner, M. A., Goebel, B. M., Dojka, M. A., and Pace, N. R. (1998). Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. *Appl. Envir. Microbiol.*, 64(8):3110–3113.

White, J. M. (2014). *ProjectTemplate: Automates the creation of new statistical analysis projects.* R package version 0.6.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):1.

Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE*, 8(4):e60234.