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

- basic introduction - more detailed background - general problem - main result summary - explanation of what main result reveals - general context - broader perspective

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

INTRODUCTION

Contamination of microbiological materials such as cell cultures and genomic DNA with biological organisms is a common problem in research and diagnostic laboratories (REF). Microbiological materials whether whole cell or genomic DNA free of organismal contaminants are needed for research use and validation of detection assays. These materials include but are not limited to environmental strains isolated for in lab use, strains obtained from an other laboratory or culture collection, and reference materials. Regardless of the material source, it is good practice for the material provider and user to validate the material is suitable for their application. Material validation often requires ensuring that the material is free of organismal contaminants that would interfere with there application. Application contaminant requirements may require the material be free of contaminants with varying levels of taxonomic similarity to the material or present below a defined threshold.

Detection of organismal contaminants, assessing a materials genomic purity, is performed using three primary methods; culturing, polymerase chain reaction (PCR), and whole genome sequencing. For whole cell material purity is most commonly evaluated using traditional culture based methods (http://www.microbiol.org/resources/monographswhite-papers/usp-microbiological-best-laboratory-practices/??). Culture based genomic purity assessment is limited in that it is only able to detect contaminants that are culturable and phenotypically differentiable from the material strain. Additionally, culture methods are not suitable for detection of organismal contaminants in genomic material. PCR is another commonly used technique to detect organismal contaminants. While, PCR is able to detect contaminants in both culture and genomic material the method is limited in the kinds of contaminants it is able to detect and the throughput of the method. The type of contaminants and throughput of the method is limited because the each contaminant detection assay must be designed to detect a specific contaminant (Heck et al., 2016)(Marron et al., 2013). The third type of organismal contamination detection method is whole genome sequencing. Whole genome sequencing similar to PCR can be used for both culture and genomic material but is also less restricted in the type of detectable contaminants. Two examples of whole genome sequencing based detection method are DeconSeq (Schmieder and Edwards, 2011a) and a similar method QC-Chain (Zhou et al., 2013). These two method 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). While DeconSeq and QC-Chain are able to detect prokaryotic contaminants within a culture or genomic material with a single assay, they cannot detect non-prokaryotic contaminants or contaminants with the sample 16S rRNA gene sequence as the material strain. As whole genome sequencing data include genomic information from all organisms within a sample including eukaryotic and viral contaminants.

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 provide a baseline assessment of the method using simulated data for single organisms to characterize the contaminant detection false positive rate. Then, we evaluate the methods ability to detect contaminants from genus other then the material strain using simulated contaminated test material datasets.

METHODS

Simulated whole genome sequence data was used to evaluate 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 characterize the rate at which the method correctly classifies reads as the test material. To characterize the ability of the method to detect contaminants, simulated 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 were simulated using an empirically determined error model and insert size distributions. The whole genome sequencing data was simulated using the ART sequencing read simulator (Huang et al., 2012). Reads were simulated with ART simulator using the Illumina MiSeq error model for 2×230 base pair (bp) paired end reads with an insert size of 690 ± 10 bp (average $\pm$ standard deviation) and 20 X mean coverage. 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 sequences 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`).

Single Genome - Baseline Assessment

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).

Simulated Contaminants

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 contaminant 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. *Make Sure to Revise for Clarity - Maybe include a figure/diagram.* 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 baseline assessment. The resulting simulated sequencing data was first processed using the 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 as described above 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.

Bioinformatic Pipeline

To facilitate repeatability and transparency, a Docker (`www.docker.com`) container is available with installed pipeline dependencies (`www.registry.hub.docker.com/u/natedolson/docker-pathoscop`
The script used to run the simulations are available at `https://github.com/nate-d-olson/genomic_purity`. Additionally, seeds number for the random number generator was randomly assigned and recorded for each dataset so that the same simulated datasets could be regenerated. 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`) along with the source file for this manuscript.

RESULTS

Single Genome - Baselines Assessment

Simulated sequence data from individual isolates was used to assess the genomic purity assessment method specificity. 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). 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. 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 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.

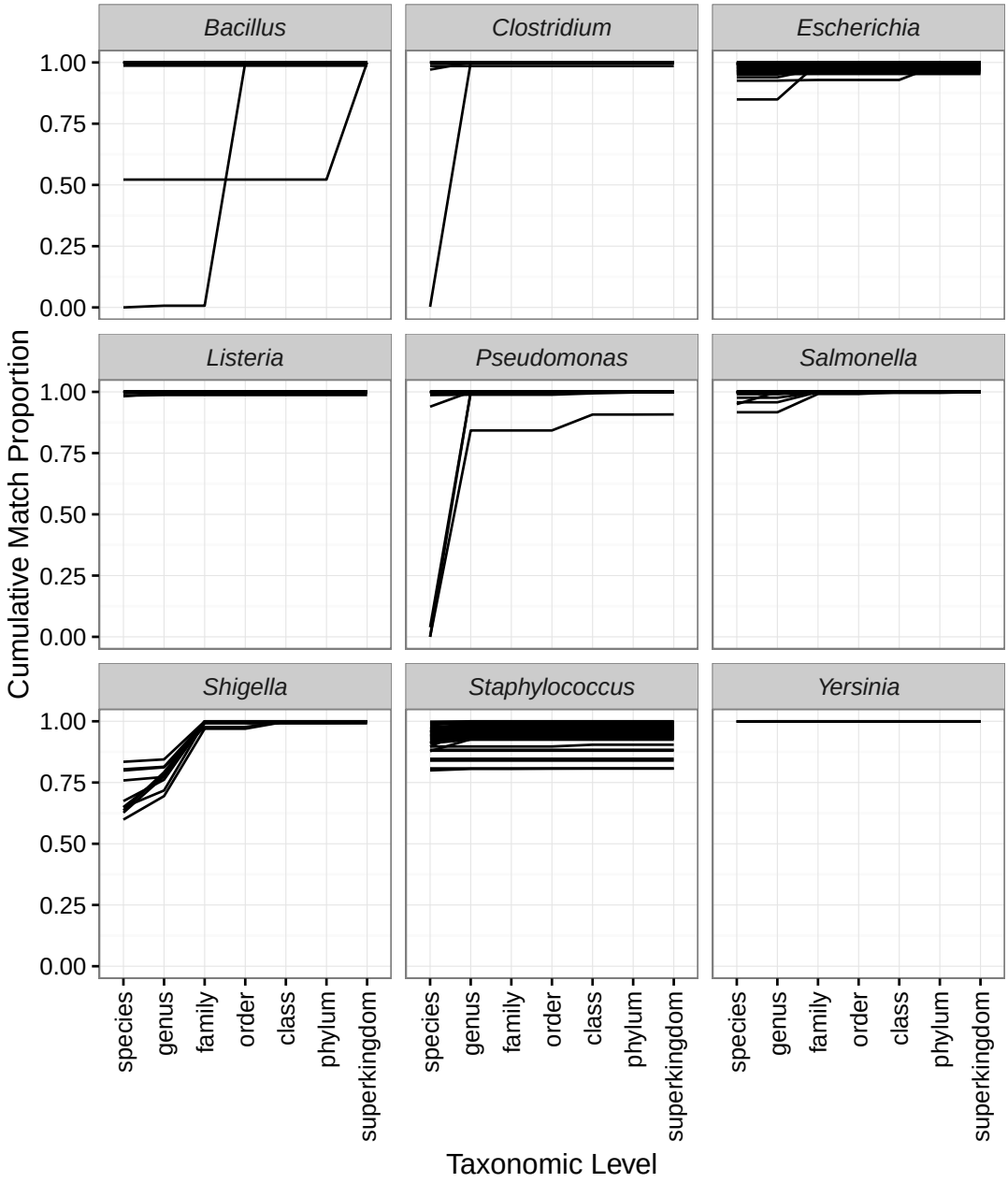

Figure 1. Cumulative taxonomic match results for genomic purity assessments of simulated sequence data from single genomes. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level for an individual genome. Genomes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

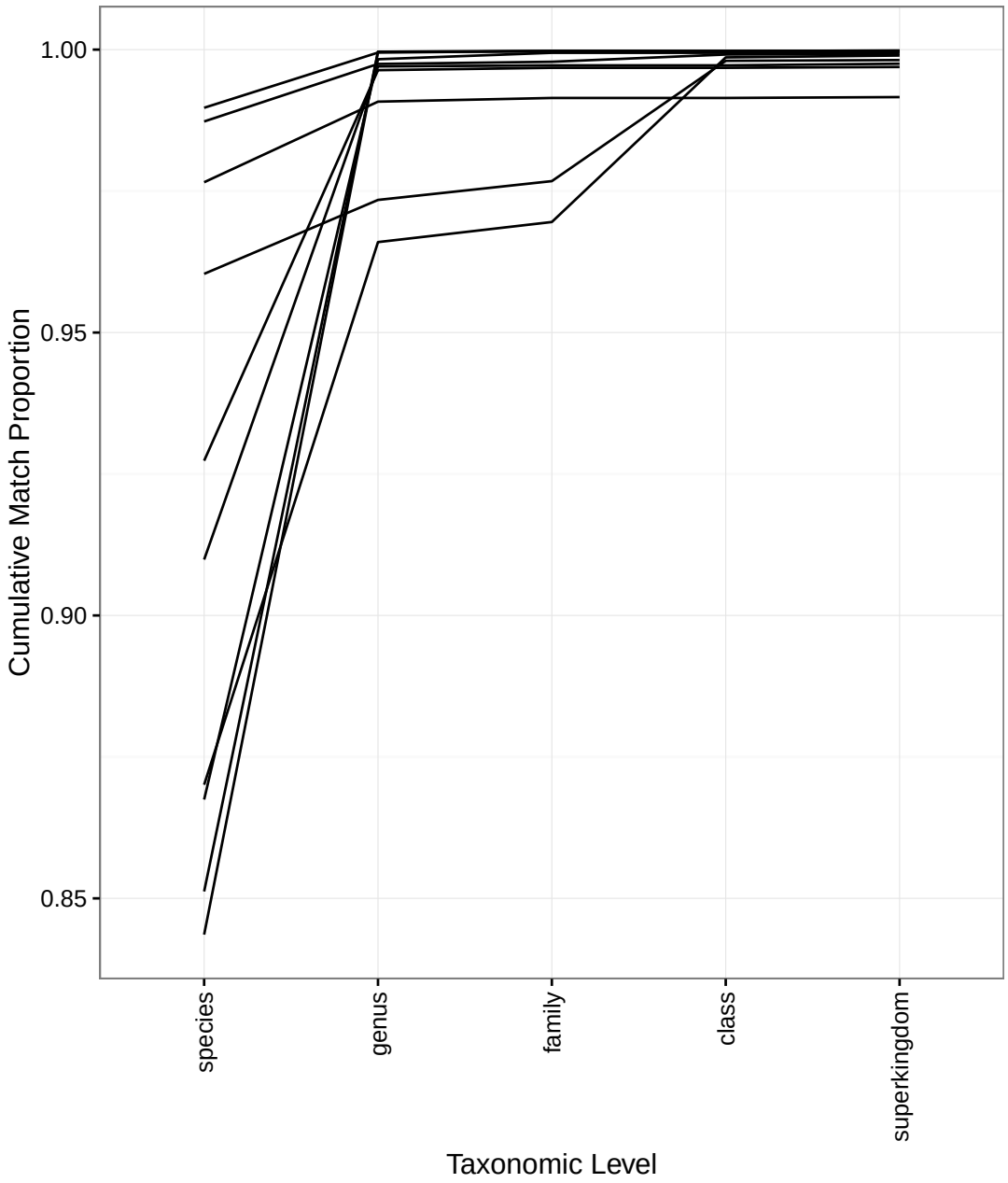

Figure 2. Cumulative taxonomic match results for genomic purity assessment for *Shigella* considering matches to *E. coli* as species level matches. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes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

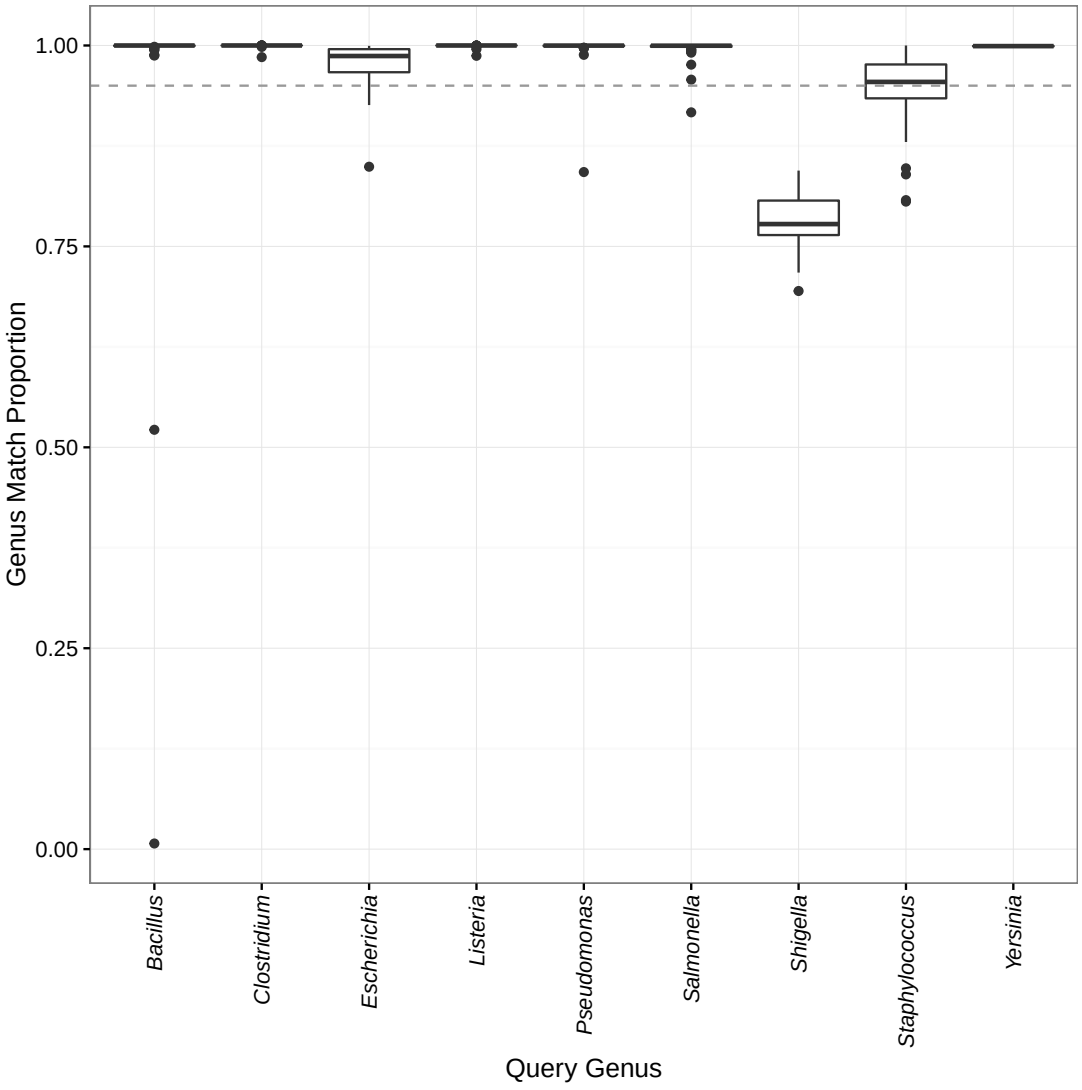

Figure 3. Distribution of the proportion of reads assigned to the source genome at or above the genus level. Horizontal grey line highlights a match proportion of 0.95. Boxplots hinges represent the 25th and 75th percentiles, line through box represent is the median, whiskers are the 95% confidence interval,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

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 simulated datasets. N indicates the number of genomes, and Genome Size is presented as the median and range (minimum to maximum) genome size

Simulated Contaminants - Detection Assessment

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 baseline assessment 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 all of the genomes selected for the detection assessment 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). extitEscherichia coli O157:H7 str. EC4115 and extitYersinia pestis CO92 have two and three plasmids respectively.

To evaluate the method's ability to detect contaminants, 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 results are from simulated data and based on proportions of simulated reads, these values do not indicate a limit of detection for the method.

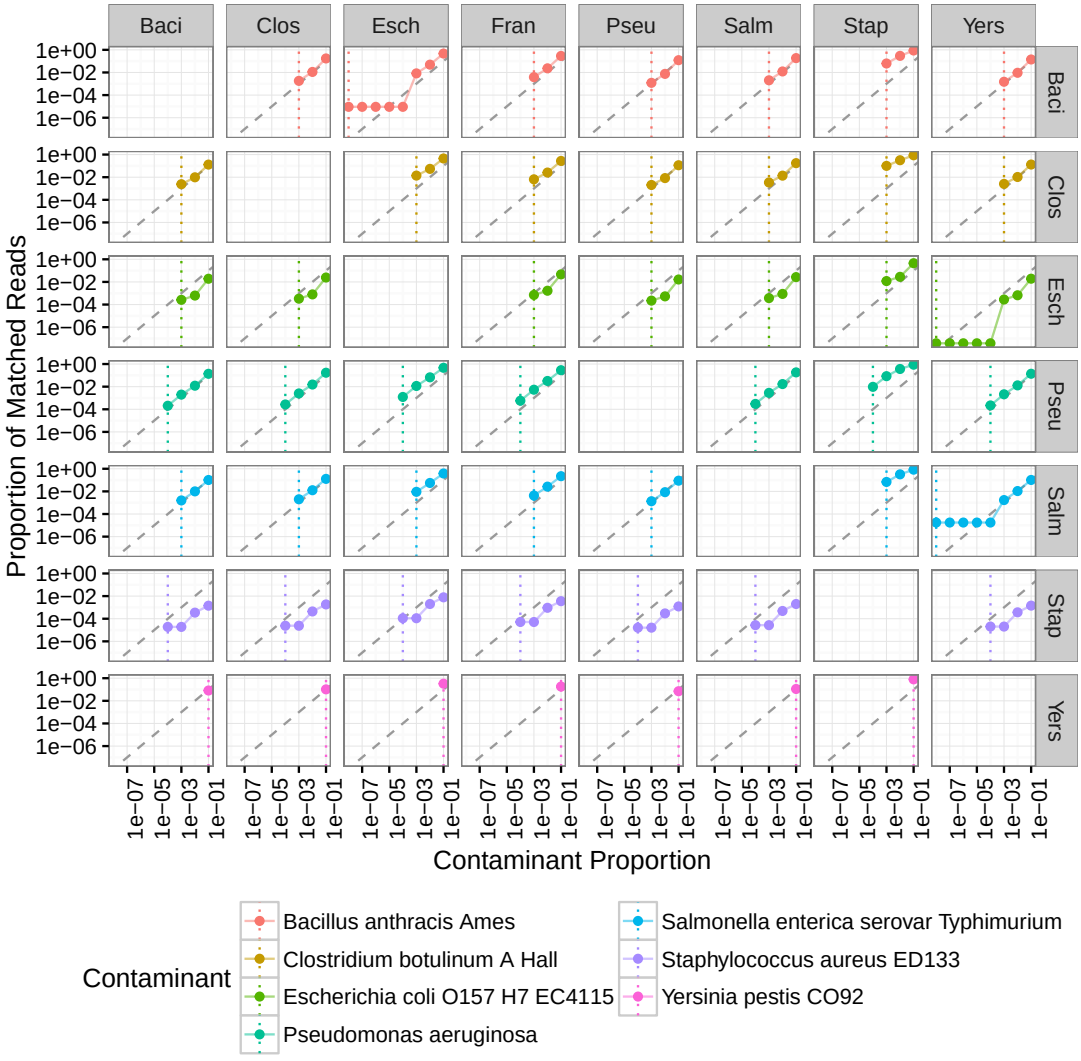

Figure 4. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

DISCUSSION

* Methods presented here represent an initial step in the development of whole genome sequencing based methods for detecting genomic contaminants microbial materials such as cell cultures or DNA extracts. * General conclusions regarding method performance * General statements regarding method limitations * Characterization of limit of detection, ability to differentiate between closely related strains. * How would this method be applied to different applications * Culture collections * how the results from the study support (or not) the use of the method * able to identify contaminants unrelated above the species level * limitation in ability to identify species or strain level contaminants * Isolate characterization * how the results from the study support (or not) the use of the method * validate non-mixed culture (to species level) * what additional work is needed to validate the method for the specific application * to validate single isolate additional analysis required * Reference materials * how the results from the study support (or not) the use of the method * able to identify contaminants * what additional work is needed to validate the method for the specific application * unable to detect species level contaminants * for reference materials and quantitative detection assays * characterization of limits of detection * articulate where the current methods for purity assessment have fallen short and where this work would advance capabilities * Current methods such as traditional micro culture techniques and PCR * limitations regarding a priori * culturable * WGS - less a priori assumptions * theoretically higher limits of detection (only need more sequence data) * Using common whole genome sequencing methods * data can be used for other applications (REF pepr)

CONCLUSIONS

Reference materials and strains are commonly used in basic and applied research settings. To ensure the validity of the conclusions drawn from the results of experiments using these materials the material genomic purity should be evaluated. We have demonstrated that whole genome sequencing paired with taxonomic read classification methods are able to detect genomic contaminants at levels down to **NEED TO FILL IN** when evaluating contaminants as the genus level. While the methods used in this study produced high specificity at the genus level, other classification algorithms may result in higher specificity. Additionally, long read sequencing methods such as Pac Bio and Oxford Nanopore, have the potential to increase method specificity. The method sensitivity was dependent on the contaminant and not the target material. This is due to the method used to generate the simulated contaminant datasets as the number of reads used in the contaminated dataset is dependent on the genome size. Increasing sequencing depth is likely to result in increased sensitivity. We have presented a proof of concept study for a novel method for evaluating test material purity. With the decreasing cost of whole genome sequencing this method provides a viable alternative to other commonly used methods for evaluating test material purity. When using whole genome sequencing in combination with metagenomic taxonomic read classifiers users should make sure to validate and optimize the methods for their specific use case. Validation and optimization would include selection of the appropriate database, evaluation of method sensitivity and specificity in a manner similar to what was presented here, as well as evaluation of different taxonomic classification algorithms.

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

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.

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*.

Heck, K., Machineski, G. S., Alvarenga, D. O., Vaz, M. G. M. V., de Mello Varani, A., and Fiore, M. F. (2016). Evaluating methods for purifying cyanobacterial cultures by qpcr and high-throughput illumina sequencing. *Journal of Microbiological Methods*.

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.

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.

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.

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*.

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

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.