

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

Nathan D. Olson¹, Justin Zook¹, Jayne Morrow¹, and Nancy Lin¹

¹Material 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. Another PCR based assay was used to evaluate the purity of cyanobacterial cultures (Heck et al., 2016). 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). Metagenomic-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 exhaustive, 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 contaminated test material datasets.

METHODS

Simulated whole gnom 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 \pm 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 dataset 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 unambiguously 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 taxonomic hierarchy 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 *Escherichia 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 than 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-pathoscope/). 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 programming 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 AND DISCUSSION

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

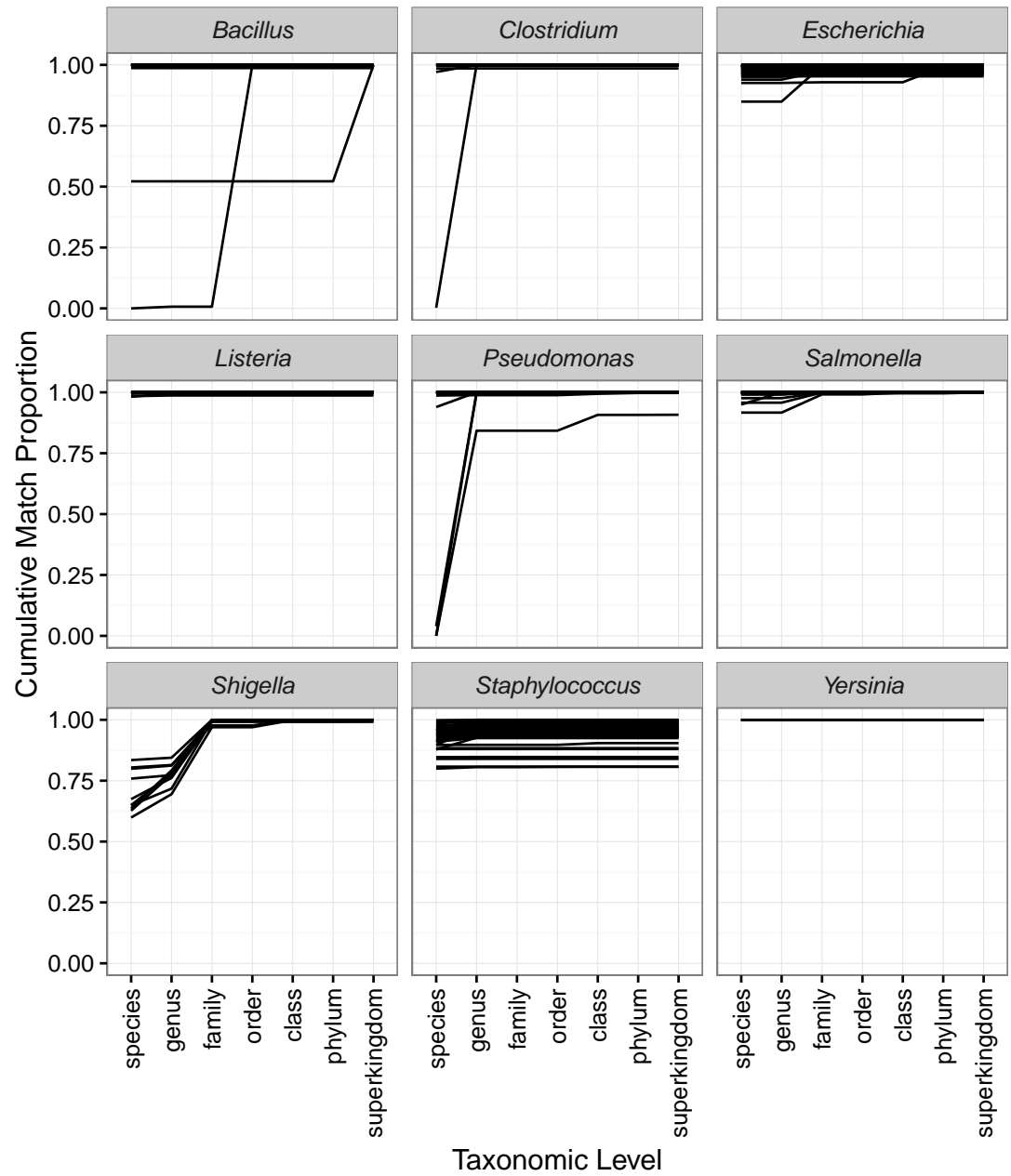


Figure 1. Cumulative taxonomic match results for genomic purity assessments of simulated sequence data from single genomes. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes are grouped by genus.

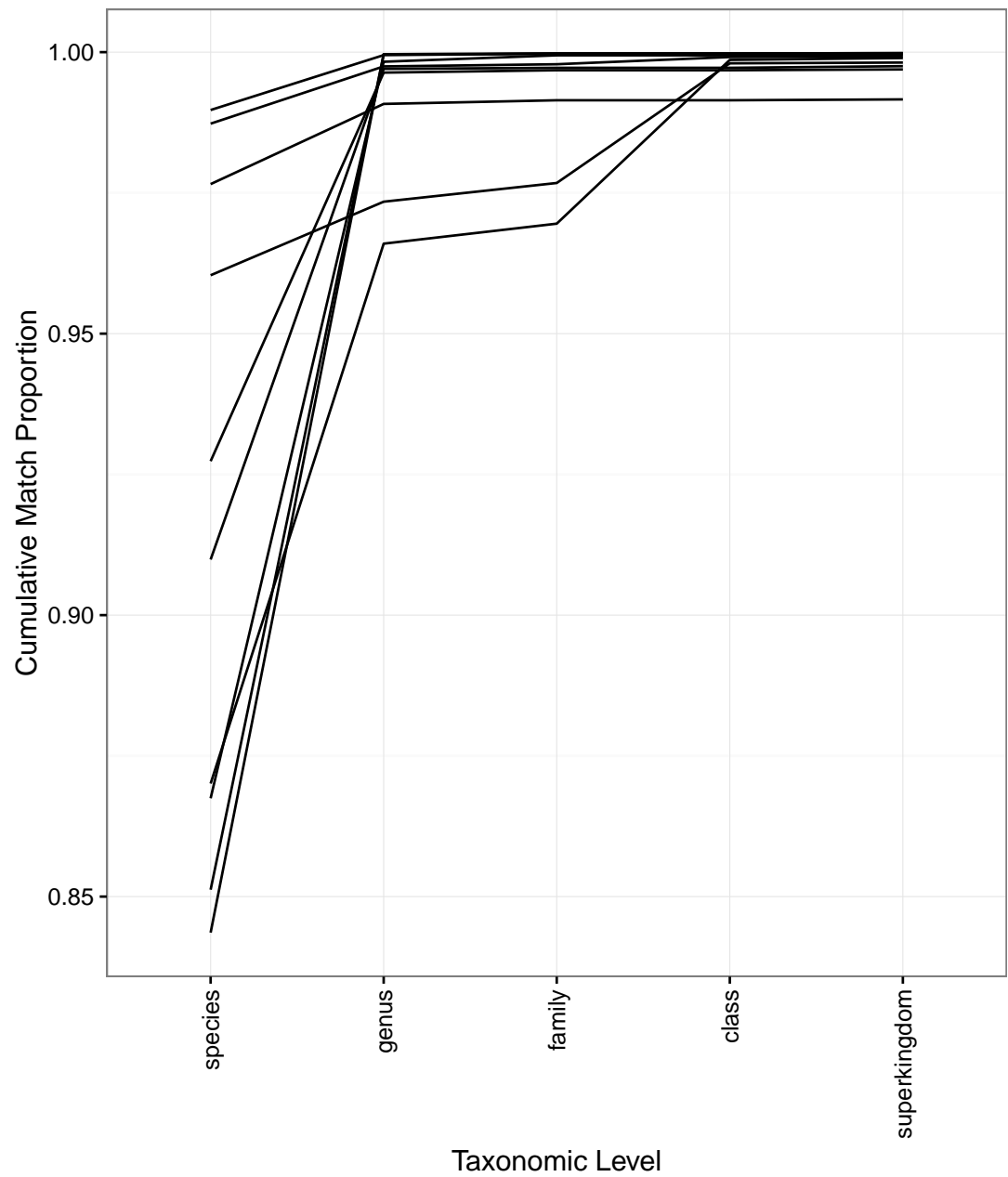


Figure 2. Cumulative taxonomic match results for genomic purity assessment for *Shigella* considering matches to *E. coli* as species level matches. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes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

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

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.

Simulated Contaminants

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
<i>Bacillus anthracis</i> str. Ames	1.00	5.23	AE016879.1		
<i>Clostridium botulinum</i> A str. Hall	1.00	3.76	CP000727.1		
<i>Escherichia coli</i> O157:H7 str. EC4115	0.98	5.57	CP001164.1	0.13	CP001163.1, CP001165.1
<i>Francisella tularensis</i> subsp. tularensis SCHU S4	1.00	1.89	AJ749949.2		
<i>Pseudomonas aeruginosa</i> PAO1	1.00	6.26	AE004091.2		
<i>Salmonella enterica</i> subsp. enterica serovar Typhimurium str. D23580	1.00	4.88	FN424405.1		
<i>Staphylococcus aureus</i> subsp. aureus ED133	0.98	2.83	CP001996.1		
<i>Yersinia pestis</i> 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 as-

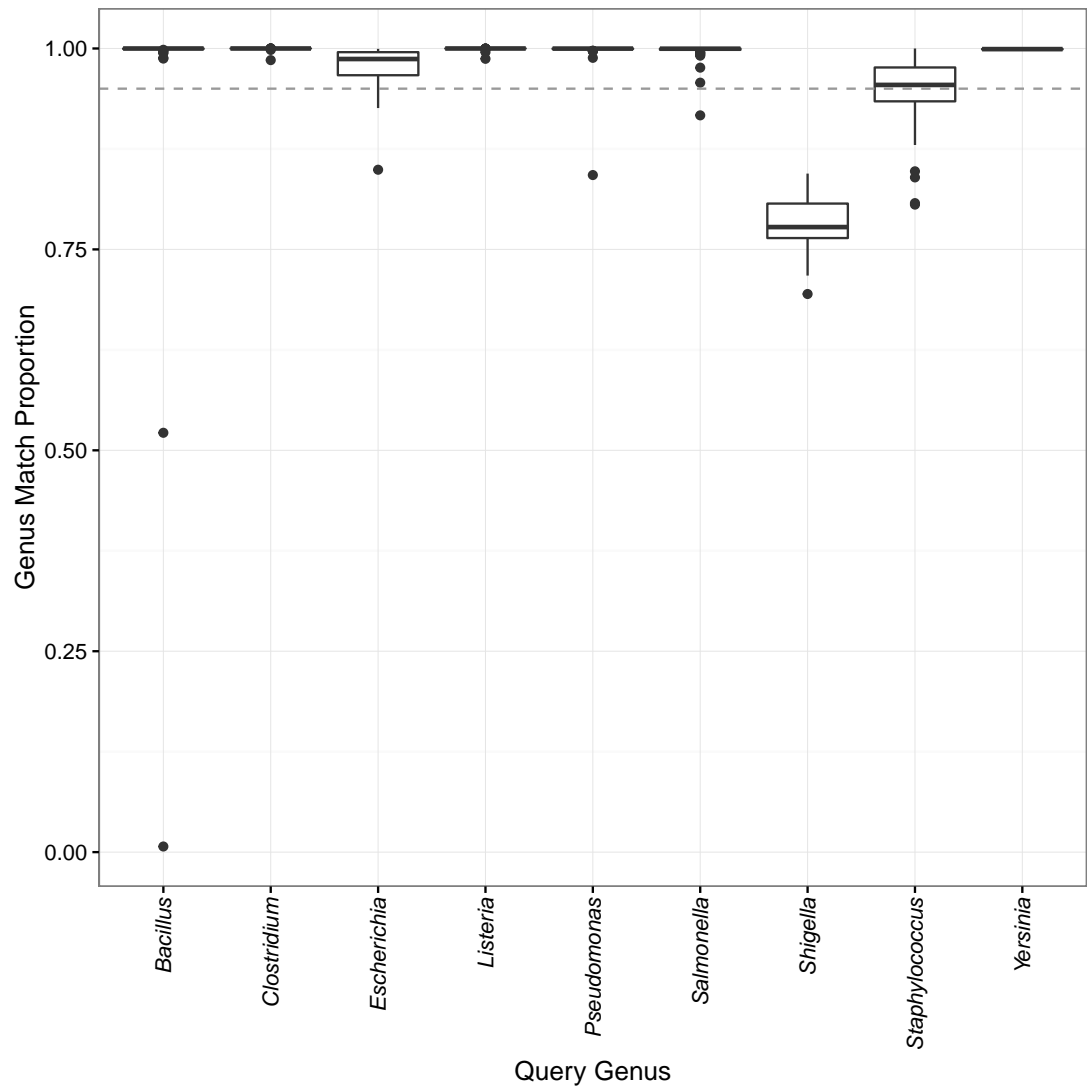


Figure 3. Distribution of the proportion of reads assigned to the source genome at or above the genus level. Horizontal grey line highlights a match proportion of 0.95. Boxplots hinges represent the 25th and 75th percentiles, line through box represent is the median, whiskers are the 95% confidence interval,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

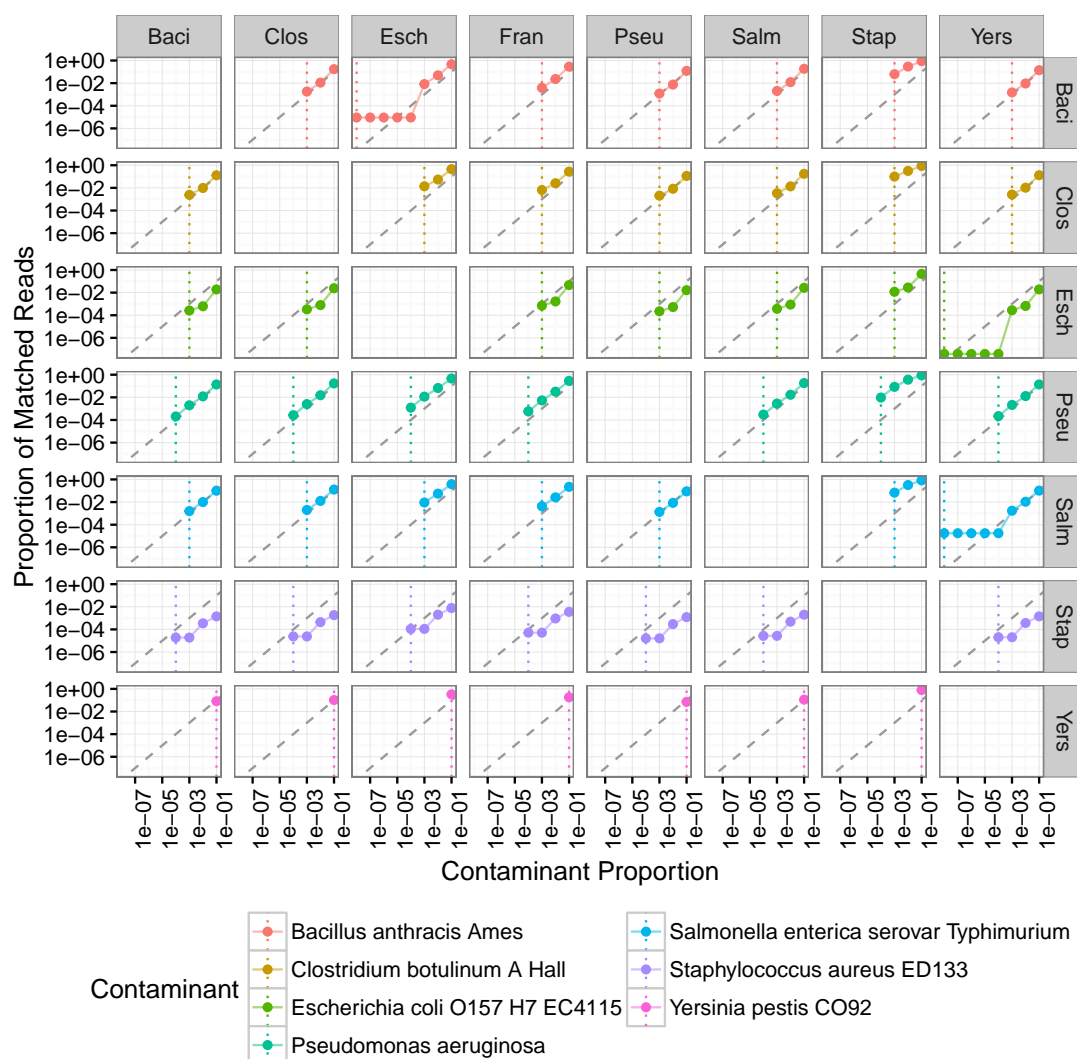


Figure 4. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

signed 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.

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

189 depth is likely to result in increased sensitivity. We have presented a proof of concept study for a novel
190 method for evaluating test material purity. With the decreasing cost of whole genome sequencing this
191 method provides a viable alternative to other commonly used methods for evaluating test material purity.
192 When using whole genome sequencing in combination with metagenomic taxonomic read classifiers
193 users should make sure to validate and optimize the methods for their specific use case. Validation and
194 optimization would include selection of the appropriate database, evaluation of method sensitivity and
195 specificity in a manner similar to what was presented here, as well as evaluation of different taxonomic
196 classification algorithms.

197 ACKNOWLEDGMENTS

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

207 REFERENCES

- 208 Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond.
 209 *Nature methods*, 12(1):59–60.
- 210 Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., Foster, Z., and
 211 O'Donnell, J. (2016). *taxize: Taxonomic information from around the web*. R package version 0.7.4.
- 212 CLSI (2010). Characterization and Qualification of Commutable Reference Materials for Laboratory
 213 Medicine ; Approved Guideline. Technical Report 22.
- 214 Coates, S. G., Brunelle, S. L., and Davenport, M. G. (2011). Development of standard method perfor-
 215 mance requirements for biological threat agent detection methods. *Journal of AOAC International*,
 216 94(4):1328–37.
- 217 EPA (2004). Quality Assurance/Quality Control Guidance for Laboratories Performing PCR Analyses
 218 on Environmental Samples October 2004. *October*, (October).
- 219 Feldsine, P., Abeyta, C., and Andrews, W. (2002). AOAC international methods committee guidelines for
 220 validation of qualitative and quantitative food microbiological official methods of analysis. Technical
 221 Report May.
- 222 Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q.,
 223 Schaalje, G. B., Clement, M. J., Crandall, K. a., and Johnson, W. E. (2013). Pathoscope: Species
 224 identification and strain attribution with unassembled sequencing data. *Genome research*.
- 225 Guide, E. (1998). The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method
 226 Validation and Related Topics. Technical report.
- 227 Heck, K., Machineski, G. S., Alvarenga, D. O., Vaz, M. G. M. V., de Mello Varani, A., and Fiore,
 228 M. F. (2016). Evaluating methods for purifying cyanobacterial cultures by qpcr and high-throughput
 229 illumina sequencing. *Journal of Microbiological Methods*.
- 230 Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read
 231 simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4.
- 232 Ieven, M., Finch, R., and van Belkum, a. (2013). European quality clearance of new microbiological
 233 diagnostics. *Clinical microbiology and infection : the official publication of the European Society of*
 234 *Clinical Microbiology and Infectious Diseases*, 19(1):29–38.
- 235 ISO/TS (2010). Microbiology of food and animal feeding stuffs - specific requirements and guidance
 236 for proficiency testing by interlaboratory comparison. Technical report.
- 237 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).
 238 *Bacillus infantis* sp. nov. and *bacillus idriensis* sp. nov., isolated from a patient with neonatal sepsis.
 239 *International journal of systematic and evolutionary microbiology*, 56(11):2541–2544.
- 240 Lan, R. and Reeves, P. R. (2002). *Escherichia coli* in disguise: molecular origins of shigella. *Microbes*
 241 *and infection*, 4(11):1125–1132.
- 242 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*,
 243 9(4):357–9.
- 244 Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of
 245 taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12(1):1.
- 246 Marron, A. O., Akam, M., and Walker, G. (2013). A Duplex PCR-Based Assay for Measuring the
 247 Amount of Bacterial Contamination in a Nucleic Acid Extract from a Culture of Free-Living Protists.
 248 *PloS one*, 8(4):e61732.

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

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

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

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

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

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

261 Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metage-
 262 nomic microbial community profiling using unique clade-specific marker genes. *Nature methods*,
 263 9(8):811–814.

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

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

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

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

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