

Using metagenomic methods to detect organismal contaminants in microbial materials.

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

¹Material Measurement Laboratory, National Institute of Standards and Technology

ABSTRACT

TODO

Keywords: Biodetection, Microbial Material, Reference Material, Purity, Bioinformatics

INTRODUCTION

Shotgun metagenomic sequencing is used to characterize environmental samples and detect pathogens in complex samples. The same method can also be used to detect contaminants in microbial materials, cell cultures and genomic DNA from clinical or environmental isolates. Microbial materials free of contaminants are needed for biodetection assay validation (Ieven et al., 2013; Coates et al., 2011) and basic research using model systems (Shrestha et al., 2013). Current methods for detecting contaminants in microbial materials use traditional methods such as culture, microscopy and polymerase chain reaction (PCR) (REF). Culture and microscopy based methods are not appropriate for genomic DNA materials and assumes the contaminants are phenotypically distinct from the material isolate it is contaminating. While, PCR based methods can be used to detect contaminants in genomic DNA, the method is limited as contaminant detection assays are contaminant specific and not amenable to high-throughput applications (Heck et al., 2016; Marron et al., 2013). In contrast to these methods, shotgun metagenomic methods can be used to detect contaminants in both cell cultures and genomic DNA materials and only require that the contaminant is genotypically differentiable from the material strain.

Shotgun metagenomics consists of two main steps, whole genome sequencing of genomic DNA, and analyzing the resulting sequencing data, most commonly using a taxonomic assignment algorithm (Thomas et al., 2012). For genomic DNA material, the material itself can be sequenced, whereas the genomic DNA must be extracted from cell cultures prior to sequencing. After sequencing a taxonomic assignment algorithm is used to characterize the sequencing data. There are a number of different types of classification algorithms with varying classification accuracy and computational performance (Bazinet and Cummings, 2012; Menzel and Krogh, 2016). All methods require a reference database for classification. In order to detect a contaminant within a microbial material, the contaminant or an organism more closely related to the contaminant than the material must be present in the database. As taxonomic classification algorithms are constantly improving, reference databases are expanding, and the cost of sequencing drops, shotgun metagenomic sequencing provides an alternative to current methods for detecting contaminants in microbial materials.

In this work, we present the results of a proof of concept study evaluating the suitability of whole genome sequencing data combined with a taxonomic assignment algorithm for detecting organismal contaminants in microbial materials. We first provide a baseline assessment of the method using simulated sequencing data from single organisms to characterize the types of false positive contaminants the method may report. Then, evaluate the method's ability to detect organismal contaminants in microbial material strains using sequencing data simulated to replicate microbial materials with different organismal contaminants at a range of concentrations.

METHODS

Simulated whole genome sequence data was used to evaluate the suitability of using whole genome sequence data and metagenomic taxonomic classification methods for detecting organismal contaminants in microbial materials. 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 method's ability to detect contaminants, simulated contaminant datasets comprised of pairwise combinations of single genomes spiked with a defined proportions of contaminant reads, reads simulated from a different genome.

To best approximate real sequencing data reads were simulated using an empirical error model and insert size distribution. 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 (Olson et al., 2016).

The taxonomic composition of simulated dataset was assessed using the Pathoscope sequence taxonomic classifier (Francis et al., 2013). Pathoscope 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, 2011), (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

Simulated sequencing data from individual genomes was used to characterize the false positive contaminants reported by Pathoscope. 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*, genomes from 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, Taxize (Scott Chamberlain and Eduard Szocs, 2013; Chamberlain et al., 2016).

Simulated Contaminants

Simulated contaminated datasets were used to evaluate how contaminant detection varied by material and contaminant strain over 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. *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 contaminated 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 material and contaminant datasets were subsampled as described above then combined. 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 once 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, seed numbers for the random number generator were 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

Single Genome - Baselines Assessment

We first assessed baseline performance of the proposed method for detecting organismal contaminants in microbial materials. Our analysis included taxonomic classification results for simulated sequencing data from 388 genomes, representing 9 different genera (Table 1). For 105 out of 388 genomes, Pathoscope estimated that 99% of the material was the same species as the genome the sequencing data was simulated from (Fig. 1). The estimated proportion of the material identified as the correct species varied by genus, with none of the *Shigella* genomes and five of the 49 *Staphylococcus* genomes having proportions greater than 0.9. *Shigella* and *Staphylococcus* along with *Escherichia* represent 87 of the 105 genomes with a less than 0.99 estimated match proportions at the species level. Excluding *Shigella*, *Escherichia*, and *Staphylococcus* the median estimated proportion matching at the species level or higher is 0.9995. The low species level match proportions were due to false positive contaminants as the input sequencing data were simulated from individual genome sequences.

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

We characterized the false positive contaminants responsible for the genera *Shigella*, *Escherichia*, and *Staphylococcus*, as well as genomes of other genera with species match proportions less than 0.9. The false positive contaminants were grouped into three types, taxonomic ambiguities, phage, and method artifacts. Taxonomic ambiguities were defined as contaminants with highly similar genome sequences

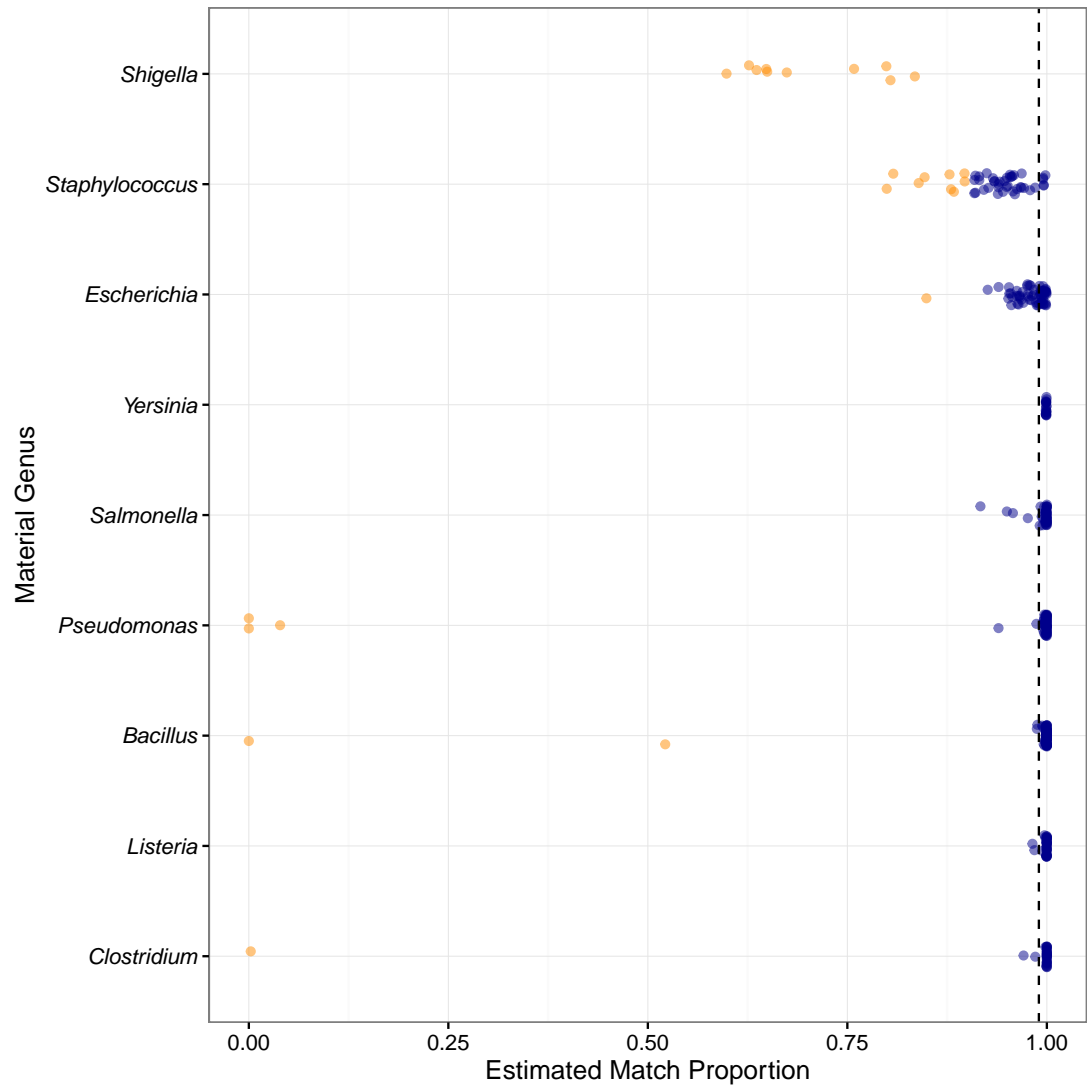


Figure 1. Species level estimated match proportion varies by material genus. The proportion of the material, simulated sequence data from individual genomes, was estimated by Pathoscope. The estimated match proportion is the total proportion of the material with taxonomic assignments to the genome species, subspecies, strain, or isolate levels. The vertical dashed line indicates the 0.99 match proportion. Orange points are genomes with species level match proportions less than 0.90 and blue points greater than 0.90

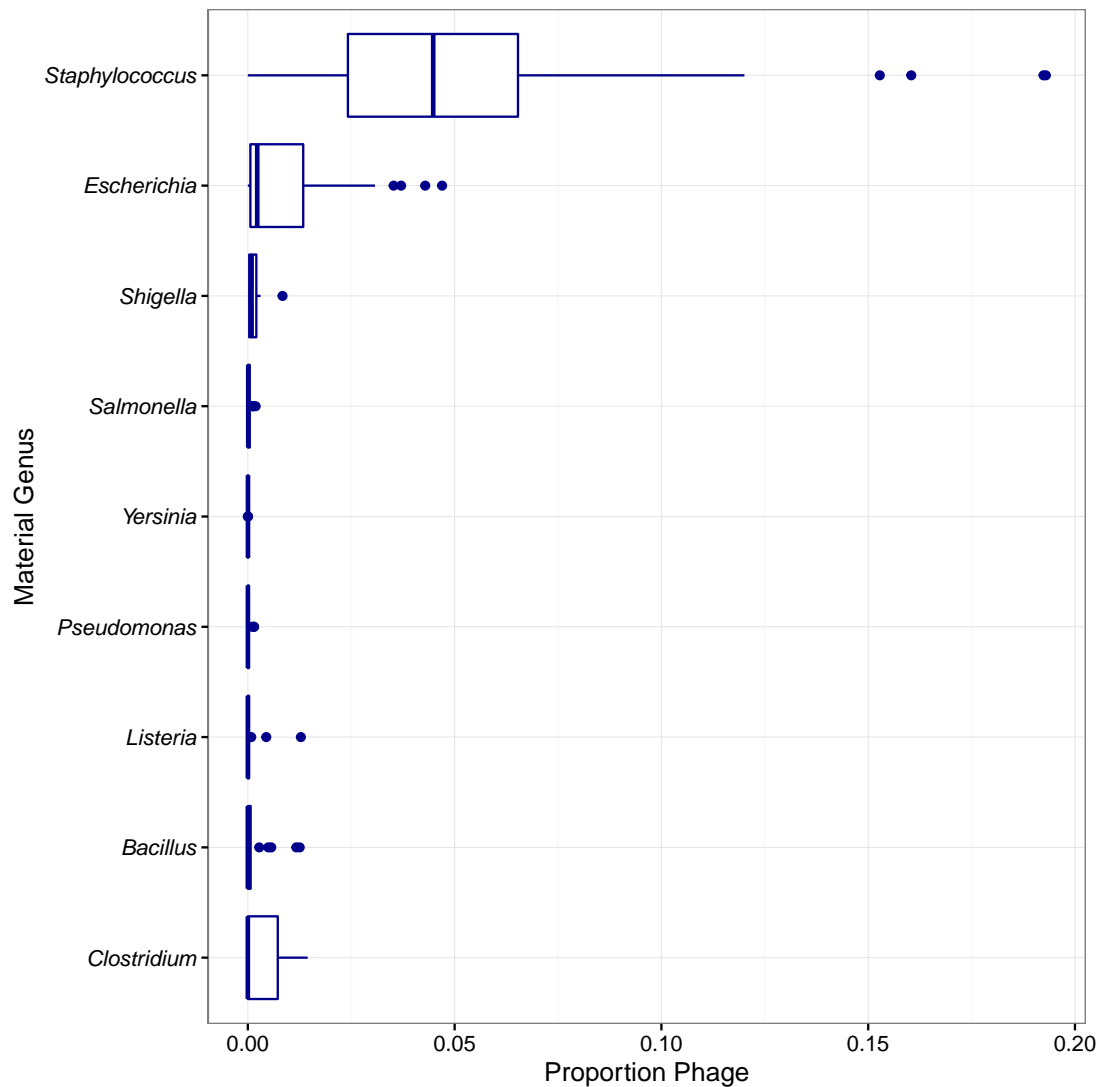


Figure 2. Estimated proportion of phage in the simulated single genome datasets by genera.

130 to the material genome sequence but taxonomically classified as different species. For example the low
 131 match percentage for *Clostridium autoethanogenum* strain DSM10061 was due to *Clostridium ljung-*
 132 *dahlia* strain DSM13528 assigned the top proportion (VALUE) instead of *C. autoethanogenum*. Sim-
 133 ilarly, *Escherichia coli* strain UMNK88 low match, due to two bacteria in the same family as *E. coli*,
 134 Enterobacteriaceae, *Providencia stuartii* and *Salmonella enterica* subsp. *enterica* serovar Heidelberg
 135 with estimated proportions of VALUE and VALUE respectively. Taxonomic ambiguities can be due to a
 136 species being incorrectly assigned to the wrong taxonomic group for example the *Bacillus* genome with
 137 species match proportion close to zero, *Bacillus infantis* string NRRL B 14911. While the *B. infantis*
 138 strain was originally classified as *Bacillus*, the species is phylogenetically distinct from other members of
 139 the genus (Ko et al., 2006). Taxonomic ambiguities are at least partially responsible for the low species
 140 level match proportions for *Shigella* and *Escherichia*, as *Shigella* is not phylogenetically distinct from *E.*
 141 *coli*(REF). When including matches to *E. coli* as species level matches, the median match proportions
 142 increase from 0.66 to 0.92. Though considerably higher, this match proportion is still low relative to the
 143 other genera **INCLUDE VALUE DIFFERENCE**.

144 Phage, the second type of false positive contaminant, were reported by Pathoscope as present in
 145 varying proportions for genomes from all 9 genera investigated (Fig. 2). Most notably, low proportions
 146 of species level matches for *E. coli* and *Staphylococcus* is partly due to relatively higher proportions of

147 matches to phage, compared to the other genera investigated (Fig. 2). All of the false positive phage
 148 contaminants were specific to the taxonomy of the genome the sequence data was simulated from. The
 149 phage contaminants may represent errors in the database, where sequence data from the host organisms
 150 genome is misassembled into the phage genome, or genomic sequence is shared between the phage and
 151 the host, such as CRISPR, and lysogenic phage (REF).

152 **Method Artifacts** The third type of false positive contaminant, are method artifacts. Method arti-
 153 fact false positive contaminants are defined as false positives due to misclassified or unclassified
 154 sequences in the database, genome assemblies in the database including sequence data from organis-
 155 mal or reagent contaminants, and errors associated with the classification algorithm. *Bacillus subtilis*
 156 BEST7613 genome had low estimated species level match proportion due to *Synechocystis* sp. PCC
 157 6803 substr. PCC-P being estimated as comprising 47% of the material (Kanesaki et al., 2012). *Syne-*
 158 *chocystis* is in a different phylum compared to *Bacillus*, cyanobacteria versus firmicutes. The high match
 159 proportion is not due to taxonomic ambiguities but potentially an error in the database. Low species level
 160 match proportions can also be due to the database containing unclassified sequence data for organisms
 161 highly similar to the material genome. For example the low match proportion for *Pseudomonas* strain
 162 FGI182 was due to matches to unclassified bacteria, bacterium 142412, and unclassified *Pseudomonas*
 163 species, *Pseudomonas* sp. HF-1. The eukaryotic false positive contaminants are likely either due to simi-
 164 larities between the genome sequences or errors in the assembly **REF**. Plasmids and vectors are common
 165 used molecular biology tools and well known contaminants of sequencing data (REF). The plasmids
 166 and vectors identified as false positive contaminants due to errors in the genome assemblies where arti-
 167 facts of the sequencing process were not properly removed from the sequencing data prior to assembly.
 168 Alternatively, the misclassification could be due to high similarity between the genome sequence the
 169 reads were simulated from and the plasmid and vector sequence which is not unexpected as most plas-
 170 mid and vectors have microbial origins (REF). The low species proportion of species level matches for
 171 *Pseudomonas* strain TKP was likely due to both misclassified sequences (*Thioalkalivibrio sulfidophilus*
 172 strain HL-EbGr7 match proportion 0.0648), and contaminated genome sequences in the database (wheat
 173 - *Triticum aestivum* match proportion 0.087). Errors associated with the classification algorithm are
 174 likely responsible for low species level match path proportion for *Pseudomonas* strain VLB120 is likely
 175 due to errors in the database as 2.6×10^{-3} of the matches assigned to the unrelated species *Xanthomonas*
 176 sp. W17. The representative database sequence is 9 kb (<https://www.ncbi.nlm.nih.gov/nucleotide/710572>),
 177 the Pathoscope expectation maximization step include a length normalization step which is likely respon-
 178 sible for the classification error.

179 Simulated Contaminants - Detection Assessment

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 the material assigned to the correct species. 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.

180 Next we evaluated how well contaminants were detected. Using simulated sequencing data from
 181 individual genomes we generated contaminant datasets by combining subsets of simulated data from two
 182 organisms at defined proportions, with the larger proportion representing the microbial material and
 183 smaller proportion the contaminant. We simulated contaminant datasets as pairwise combinations of
 184 representative genomes from 8 of the genera used in the baseline assessment section of the study (Table
 185 2). For all of the genomes selected for the detection assessment study, the estimated proportion of
 186 material assigned to the correct species was 0.98 (Table 2).

187 The minimum contaminant proportion detected was 10×10^{-3} and 10×10^{-4} for most pairwise
 188 comparisons with a few notable exceptions. When *Yersinia* was the simulated contaminant the minimum

contam.label	Baci	Clos	Esch	Fran	Pseu	Salm	Stap	Yers
Bacillus anthracis Ames		1.0E-03	1.0E-08	1.0E-03	1.0E-03	1.0E-03	1.0E-03	1.0E-03
Clostridium botulinum A Hall	1.0E-03		1.0E-03	1.0E-03	1.0E-03	1.0E-03	1.0E-03	1.0E-03
Escherichia coli O157 H7 EC4115	1.0E-03	1.0E-03		1.0E-03	1.0E-03	1.0E-03	1.0E-03	1.0E-08
Pseudomonas aeruginosa	1.0E-04	1.0E-04	1.0E-04	1.0E-04		1.0E-04	1.0E-04	1.0E-04
Salmonella enterica serovar Typhimurium	1.0E-03	1.0E-03	1.0E-03	1.0E-03	1.0E-03		1.0E-03	1.0E-08
Staphylococcus aureus ED133	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04		1.0E-04
Yersinia pestis CO92	1.0E-01	1.0E-01	1.0E-01	1.0E-01	1.0E-01	1.0E-01	1.0E-01	

Table 3. Lowest proportion of contaminant in each pairwise combination of representative genomes detected.

detected proportion was 0.1 for all material strains (Table 3). Conversely, contaminants were detected at lower proportions, 10×10^{-8} , when *Yersinia* was contaminated with *E. coli* as well as when *S. enterica* and *E. coli* contaminated with *B. anthracis*.

The quantitative accuracy of contaminant proportions estimated by Pathoscope varied by material and contaminant strain. The Pearson's correlation coefficient was used to measure the correlation between the estimated contaminant and true contaminant proportions. The estimated and true proportions were strongly correlated for all pairwise comparisons, with an overall median and 95% confidence interval of 0.99945 (0.96943 - 0.99999 (Fig. 3). Eight of the pairwise comparisons have correlation coefficients below 0.99, all of which have *S. aureus* as either the contaminant or the material strain. Two coefficients were below 0.98, *S. aureus* contaminated with *P. aeruginosa* and *S. enterica*, 0.952 and 0.969 respectively. Normalized contaminant proportion residuals, $(estimated - true)/true$, was used to assess the accuracy of the Pathoscope contaminant proportion estimates (Fig. 5). The material genome strongly influenced the total normalized residuals with *E. coli* and *S. aureus* having consistently higher total normalized residuals compared to the other genomes.

DISCUSSION

We demonstrated the potential for using the taxonomic sequence classification algorithm *Pathoscope* to detect genomic contaminants in microbial materials using whole genome sequencing data. Using simulated sequencing data generated from individual genomes we provided a baseline assessment of the method to characterize the types of false positive contaminants identified the method. The false positive contaminants were split into three types; taxonomic ambiguities, phage, and method artifacts. We then characterized how contaminant detection varied by the material organism, the contaminant strain, and level of contamination. Overall the method was able to identify contaminant proportions at 10×10^{-3} for most pairwise contaminant/material combinations. However, the accuracy of the estimated proportion of the contaminant in the simulated contaminated material varied by contaminant and material strain.

The primary limitation of the proposed method are the observed false positive contaminants for single genome simulated sequencing data. Performing a baseline assessment of false positive contaminant using simulated sequence data from the microbial material's genome sequence, and choosing the appropriate database and taxonomic assignment algorithm can help reduce the impact of false positive on the method's ability to identify organismal contaminants. Additionally, false positive contaminants are likely database and taxonomic assignment algorithm dependent. Removing sequences from the database for irrelevant contaminants, such as phage, plasmids, vectors, and multicellular eukaryotes can help reduce the proportion of false positives. Though the relevance of these contaminants is application specific. Similarly, using a curated database free of misclassified and unclassified sequence data would further help to reduce the proportion of false positive contaminants. Differentiating between two organisms with highly similar sequences will vary by taxonomic classification algorithm. Pathoscope was used for this proof of concept study as the method uses the full reads and paired end information for taxonomic classification rather than shorter sequence fragment, *k*-mers. Our assumption is that the longer sequence allows for better discrimination between highly similar sequences. As there are numerous taxonomic classification algorithms, evaluating multiple methods using sequence data simulated from the material genome of interest can help to determine the optimal contaminant detection method for a specific microbial material.

The ability to identify contaminants present at low concentrations is critical for when the material

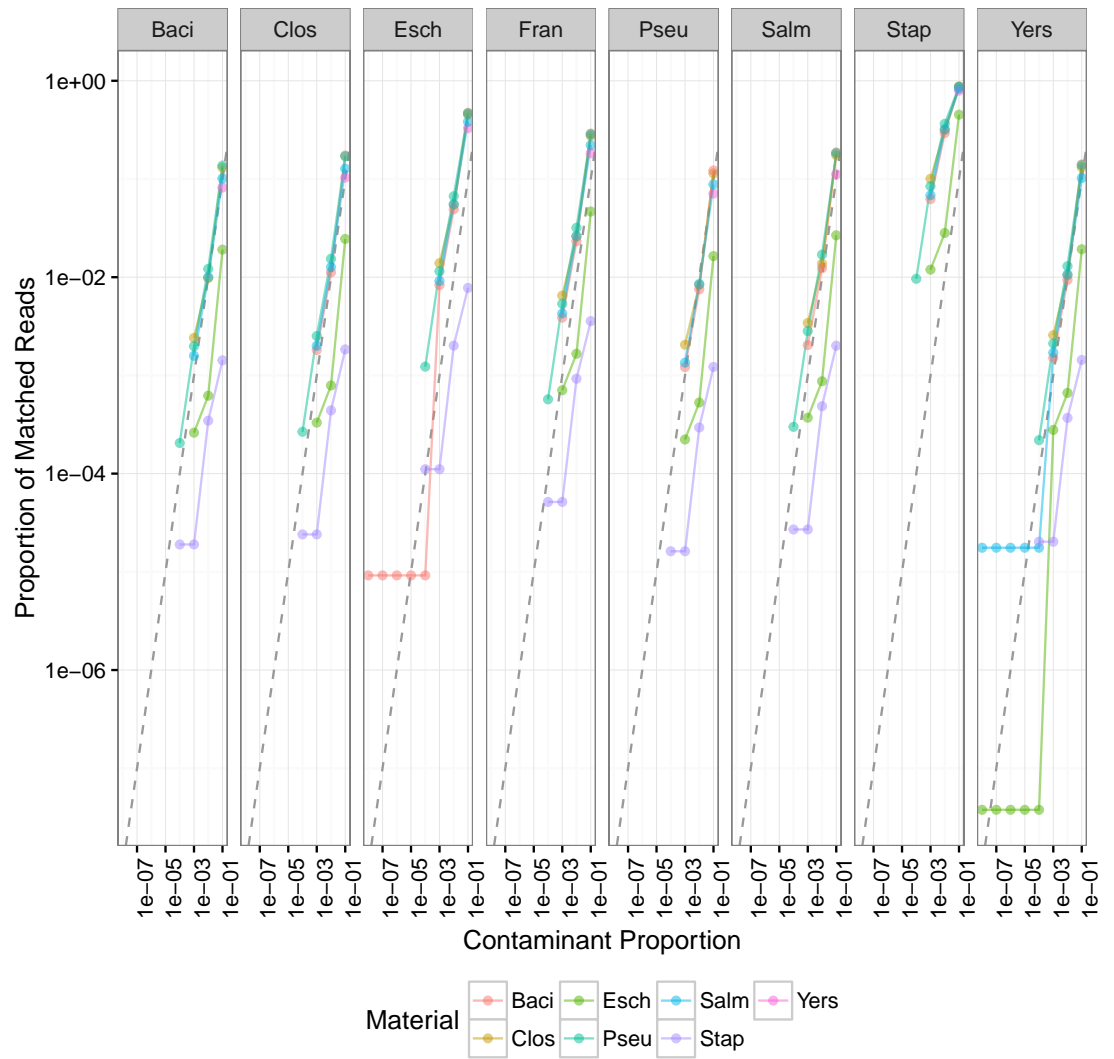


Figure 3. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

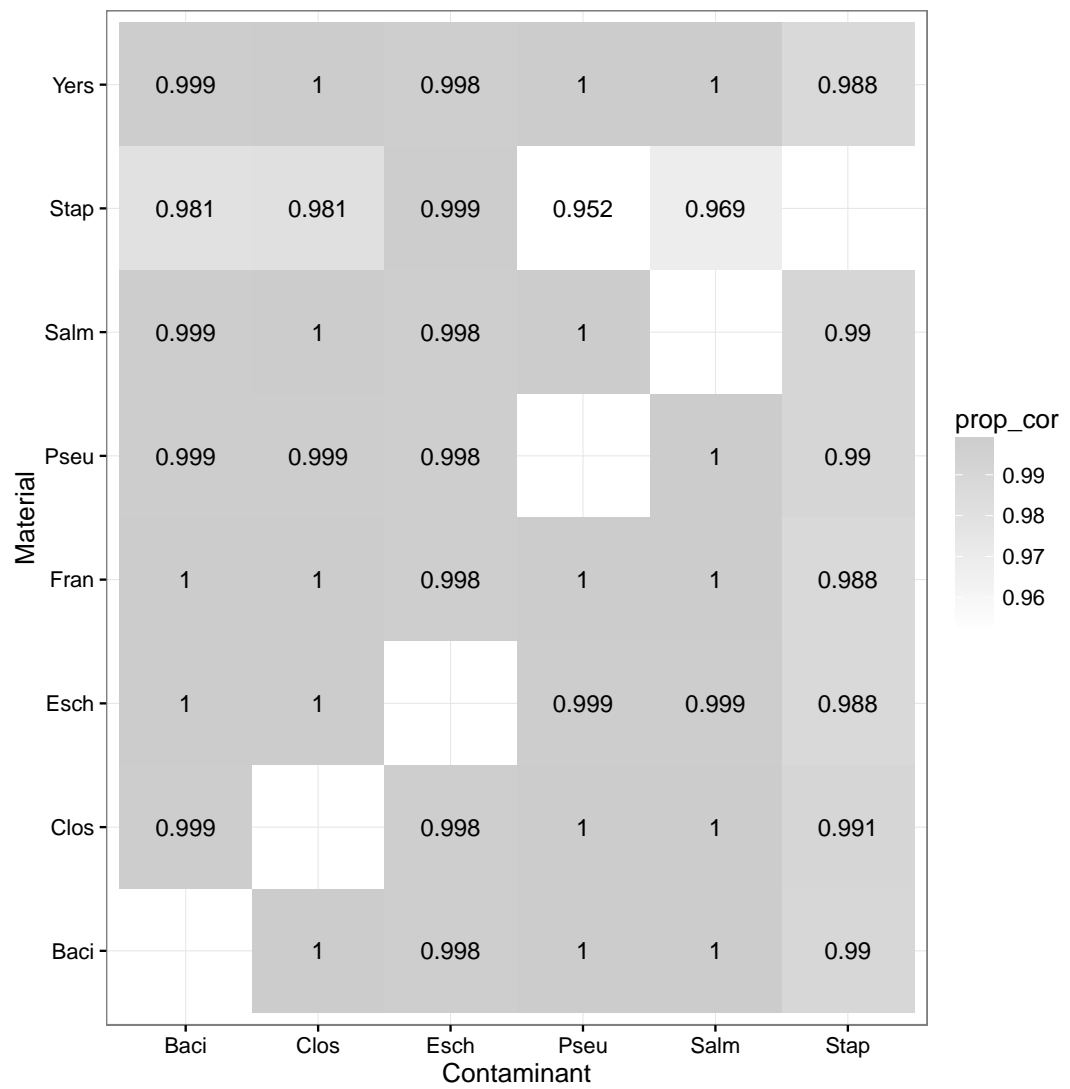


Figure 4. Pearson correlation coefficients for estimated and true contaminant proportions.

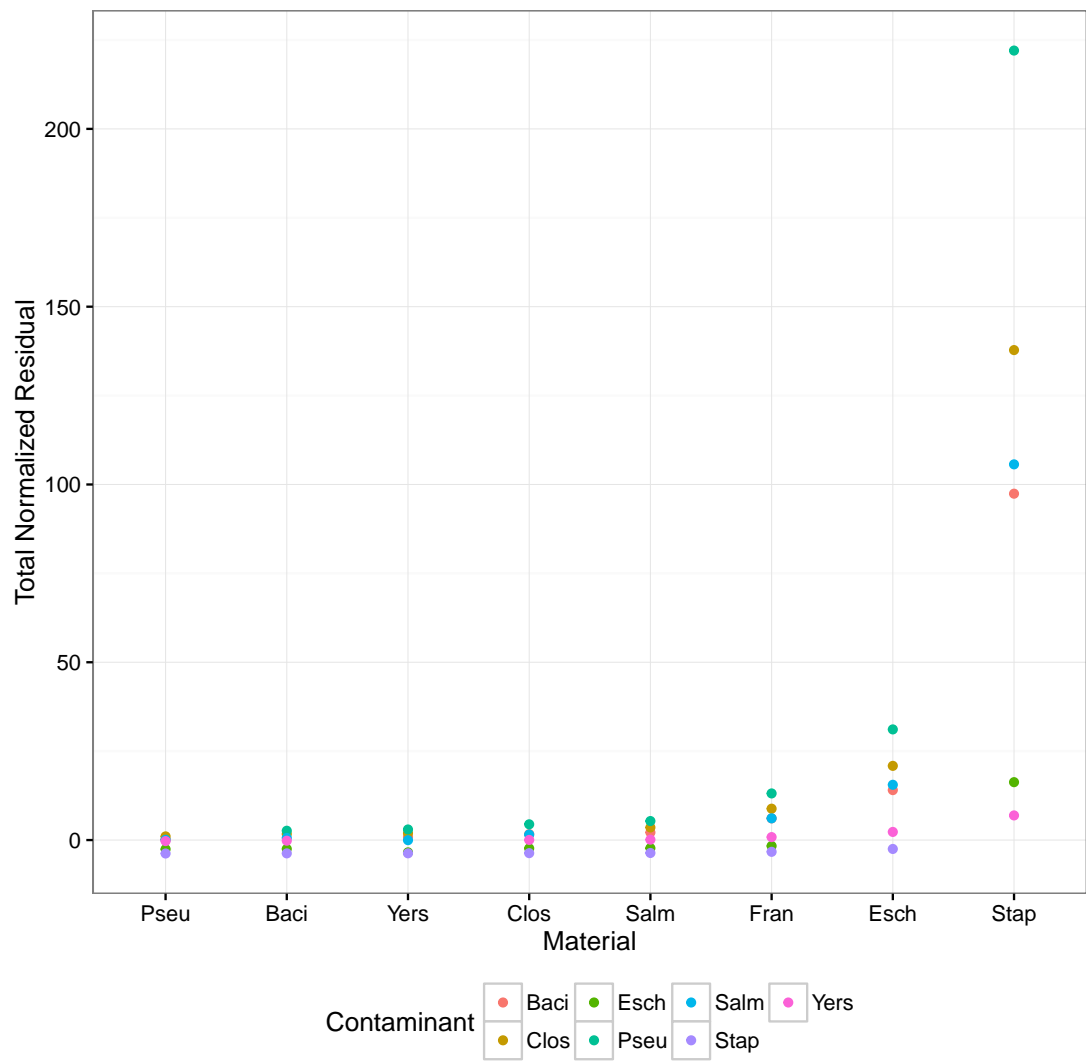


Figure 5. Total normalized residuals for pairwise combinations of material and contaminants.

231 will be used for high sensitivity assays such as PCR. The minimum contaminant proportion was con-
232 sistent for most simulated contaminant datasets, *Yersinia* was only detected at a proportion of 0.1. The
233 quantitative accuracy of the method varied by material and contaminant, but for all material-contaminant
234 pairs, the Pathoscope estimated and true contaminant proportions were highly correlated. While not nec-
235 essarily relevant to organismal contaminant detection, quantitative accuracy is relevant if the contaminant
236 proportion is included in the report of analysis characterizing a material (Olson et al., 2016). Similar to
237 the false positive contaminant baseline assessment. Simulated data can be used to evaluate the minimal
238 detectable contaminant proportion for specific contaminants of interest using different taxonomic assign-
239 ment algorithms and databases. Additionally, sequencing at a higher depth would theoretically result in
240 lower minimum detectable contaminant proportions and potentially increased quantitative accuracy due
241 to the larger sample size (number of sequences).

242 CONCLUSIONS

243 With the constant decline in the cost of sequencing, advances in sequence analysis methods, whole
244 genome sequencing combined with a taxonomic assignment algorithm provide a viable alternative to
245 commonly used organismal contaminant detection methods such as culturing, microscopy, and PCR. The
246 method is suitable for detecting organismal contaminants in both genomic DNA and whole cell microbial
247 materials without requiring any *a priori* assumptions about the contaminant. Furthermore, the method
248 was shown to detect contaminants making up 10×10^{-3} proportion of cells in a high-throughput manner.
249 Finally, the sequence data generated for organismal contaminant detection can be used to further charac-
250 terize the material's genome and potentially identify other characteristics of interest such as the presence
251 of virulence or antibiotic resistance genes.

252 ACKNOWLEDGMENTS

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

262 REFERENCES

- 263 Bazinet, A. L. and Cummings, M. P. (2012). A comparative evaluation of sequence classification pro-
264 grams. *BMC Bioinformatics*, 13(1):92.
- 265 Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., Foster, Z., and
266 O'Donnell, J. (2016). *taxize: Taxonomic information from around the web*. R package version 0.7.4.
- 267 Coates, S. G., Brunelle, S. L., and Davenport, M. G. (2011). Development of standard method perfor-
268 mance requirements for biological threat agent detection methods. *Journal of AOAC International*,
269 94(4):1328–37.
- 270 Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q.,
271 Schaalje, G. B., Clement, M. J., Crandall, K. a., and Johnson, W. E. (2013). Pathoscope: Species
272 identification and strain attribution with unassembled sequencing data. *Genome research*.
- 273 Heck, K., Machineski, G. S., Alvarenga, D. O., Vaz, M. G. M. V., de Mello Varani, A., and Fiore,
274 M. F. (2016). Evaluating methods for purifying cyanobacterial cultures by qpcr and high-throughput
275 illumina sequencing. *Journal of Microbiological Methods*.
- 276 Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read
277 simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4.
- 278 Ieven, M., Finch, R., and van Belkum, a. (2013). European quality clearance of new microbiological
279 diagnostics. *Clinical microbiology and infection : the official publication of the European Society of*
280 *Clinical Microbiology and Infectious Diseases*, 19(1):29–38.
- 281 Kanesaki, Y., Shiwa, Y., Tajima, N., Suzuki, M., Watanabe, S., Sato, N., Ikeuchi, M., and Yoshikawa, H.
282 (2012). Identification of substrain-specific mutations by massively parallel whole-genome resequenc-
283 ing of synechocystis sp. pcc 6803. *DNA research*, 19(1):67–79.
- 284 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).
285 *Bacillus infantis* sp. nov. and *bacillus idriensis* sp. nov., isolated from a patient with neonatal sepsis.
286 *International journal of systematic and evolutionary microbiology*, 56(11):2541–2544.
- 287 Lan, R. and Reeves, P. R. (2002). *Escherichia coli* in disguise: molecular origins of shigella. *Microbes*
288 *and infection*, 4(11):1125–1132.
- 289 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*,
290 9(4):357–9.
- 291 Marron, A. O., Akam, M., and Walker, G. (2013). A Duplex PCR-Based Assay for Measuring the
292 Amount of Bacterial Contamination in a Nucleic Acid Extract from a Culture of Free-Living Protists.
293 *PloS one*, 8(4):e61732.
- 294 Menzel, P. and Krogh, A. (2016). Kaiju : Fast and sensitive taxonomic classification for metagenomics.
295 *Nature communications*, 7(11257):1–9.
- 296 Olson, N. D., Zook, J. M., Samarov, D. V., Jackson, S. A., and Salit, M. L. (2016). Pepr: pipelines for
297 evaluating prokaryotic references. *Analytical and bioanalytical chemistry*, 408(11):2975–2983.
- 298 R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for
299 Statistical Computing, Vienna, Austria.
- 300 Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets.
301 *Bioinformatics*, 27(6):863–864.
- 302 Scott Chamberlain and Eduard Szocs (2013). *taxize - taxonomic search and retrieval in r*.
303 *F1000Research*.

- 304 Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When Is a Microbial Culture Pure
305 ? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing.
- 306 Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis.
307 *Microbial informatics and experimentation*, 2(1):3.
- 308 White, J. M. (2014). *ProjectTemplate: Automates the creation of new statistical analysis projects*. R
309 package version 0.6.