Using metagenomic methods to detect organismal contaminants in microbial materials.

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

TODO

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

INTRODUCTION

Presence of contamiannts that interfer with measurements is a common problem. The required purity level changes with measurement method sensitivity. High sensitivity methods suchs as PCR and next generation sequencing require higher material and reagent purity compared to traditional microbiology methods such as culturing, biochemical tests, and microscopy. Issues realated to reagent contamiannts have been well doucmented and addressed with improved methods for removing contmainants (Eschoo, MDA), negative controls (Microbiome), and post processing of sequence data (Illumina PhiX). However, contmainants in microbial materials such as non-anexic cellular materials (Shrestha et al., 2013) and genomic materials with foreign DNA contaminants have only been addressed in data processing (ProGeDe??). High sensitivity methods are needed to detect and characterize contaminants in microbial materials.

Shotgun metagenomic sequencing is used to characterize environmental samples and detect pathogens in complex samples. Shotgun metagenomic sequencing can also be used to detect contaminants in microbial materials. Microbial materials free of contaminants are needed for to populate sequence databases (REF), mock communities used to validate metagenomic methods (mockrobiota), biodetection assay validation (Ieven et al., 2013; Coates et al., 2011), 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 methods are limited as they can only detect particular targeted contaminants and are not ammenable 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 has sequencing reads that differentiate it 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 taxonommic assignment algorithm (Thomas et al., 2012). For genomic DNA materials, the material itself is sequenced, whereas genomic DNA must be extracted from from cell cultures prior to sequencing. After sequencing, a taxonomic assignment algorithm is used to characterize the sequencing data. There is a variety 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 contaminanting organism (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 de-

tecting contaminants in microbial materials.

In this work, we present the results from an *in-silico* assessment method to evaluate the suitability of whole genome sequencing data combined with a taxonomic assignment algorithm for detecting foreign DNA. This work provides a baseline assessment of the method using simulated sequencing data from single microorganisms to characterize the types of false positive contaminants the method may report. Then, the method was challenged for the 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 foreign DNA in microbial materials. Simulated data from individual prokaryotic genomes was used to characterize the rate at which the method correctly classifies reads as the test material. To evaluate contaminant detection we used datasets comprised of pairwise combinations of simulated reads from individual genomes.

Simulating Sequencing Data

To 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).

Assessing Taxonomic Composition

The taxonomic composition of simulated datasest was assessed using the Pathoscope sequence taxonomic classifier (Francis et al., 2013). Pathoscope was selected for two reasons: (1) it uses a large reference database reducing potential biases due to contaminant sequences not present in the database and (2) it uses 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, 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`).

Baseline Assessment Using Individual Genomes

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 material 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 taxononomic hierarchy for the material genome and simulated read assignment match levels were determined using the R package, Taxize (Scott Chamberlain and Eduard Szocs, 2013; Chamberlain et al., 2016).

Contaminant Detection Assessment

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

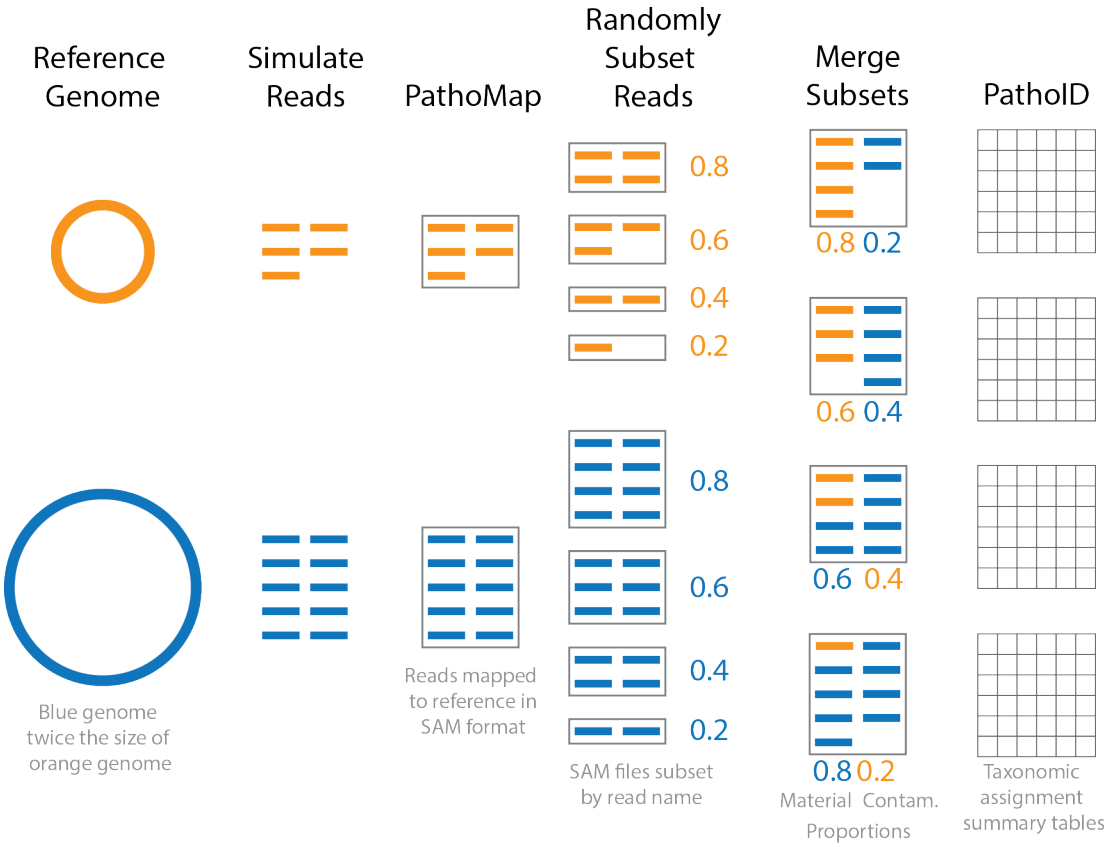

Figure 1. Diagram of the simulated contaminant dataset workflow for two individual genomes. Contaminant proportions 0.2 and 0.4 are used for demonstration purposes. The reads were initial simulated from individual genomes. The blue genome is twice the size of the orange genome and twice as many reads are simulated for the blue genome compared to the orange in order to obtain the same coverage. The simulated reads were aligned to the reference database using PathoMap. The resulting alignment file, in SAM file format, was randomly subset based on the desired proportions. Complementary subsets of SAM files (e.g. 0.2 contaminant and 0.8 material) from the two genomes were merged to create individual simulated contaminant datasets. Due to the different sized genomes the simulated contaminant datasets have different numbers of reads. Taxonomic assignment summary tables were generated from simulated contaminant datasets using PathoID.

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 (Fig. 1). The simulated datasets were randomly subsampled at defined proportions, with p representing the proportion of reads from the contaminant single genome dataset, and $1-p$ representing the proportion of reads from the material genome simulated dataset. A range of contaminant proportions at 10-fold increments were simulated with p ranging from 10^{-1} to 10^{-8}, resulting in 512 simulated contaminant datasets. This approach simulates the proportions of cells in a contaminanted material and not the amount of DNA, assuming unbiased DNA extraction. This results in organisms with larger genomes having more simulated reads.

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 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, 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 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

Baseline Assessment Using Individual Genomes

We first assessed baseline performance of the proposed method for detecting contaminant DNA 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 less than 99% of the material was the same species as the genome the sequencing data was simulated from (Fig. 2). The estimated proportion of the material identified as the correct species varied by genus. None of the *Shigella* genomes and five of the 49 *Staphylococcus* genomes had estimated proportions greater than 0.9 for the correct species. 87 of the 105 genomes with estimated match proportions less than 0.99 at the species level come from *Shigella*, *Staphlyococcus*, or *Escherichia*. 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 for pure simulations *Shigella*, *Staphlyococcus*, and*Escherichia* were due to reads being incorrectly identified as belonging to organisms from other species. We characterized false postive contaminants detected in genomes from the genera *Shigella*, *Escherichia*, and *Staphylococcus*, as well as genomes of other species match proportions less than 0.9. Two types of false positive contaminants were identified (1) contaminants that were genomically indistinguishable from the material and (2) contamiants due to errors in the reference database.

Two genome sequences can be genomically indistinguishable as they are either phylogenetically closely related or share parts of their genome. Phylogenetic similarity is at least partially responsible for the low species level match proportions for *Shigella* and *Escherichia*, as *Shigella* is not phylogenetically distinct from *E. coli* (**REF**). When including matches to *E. coli* as species level matches, the median match proportions increase from 0.66 to 0.92. Another example of false positives at the species level due to phylogentic similarity was low match percentage for *Clostridium autoethanogenum* strain DSM10061

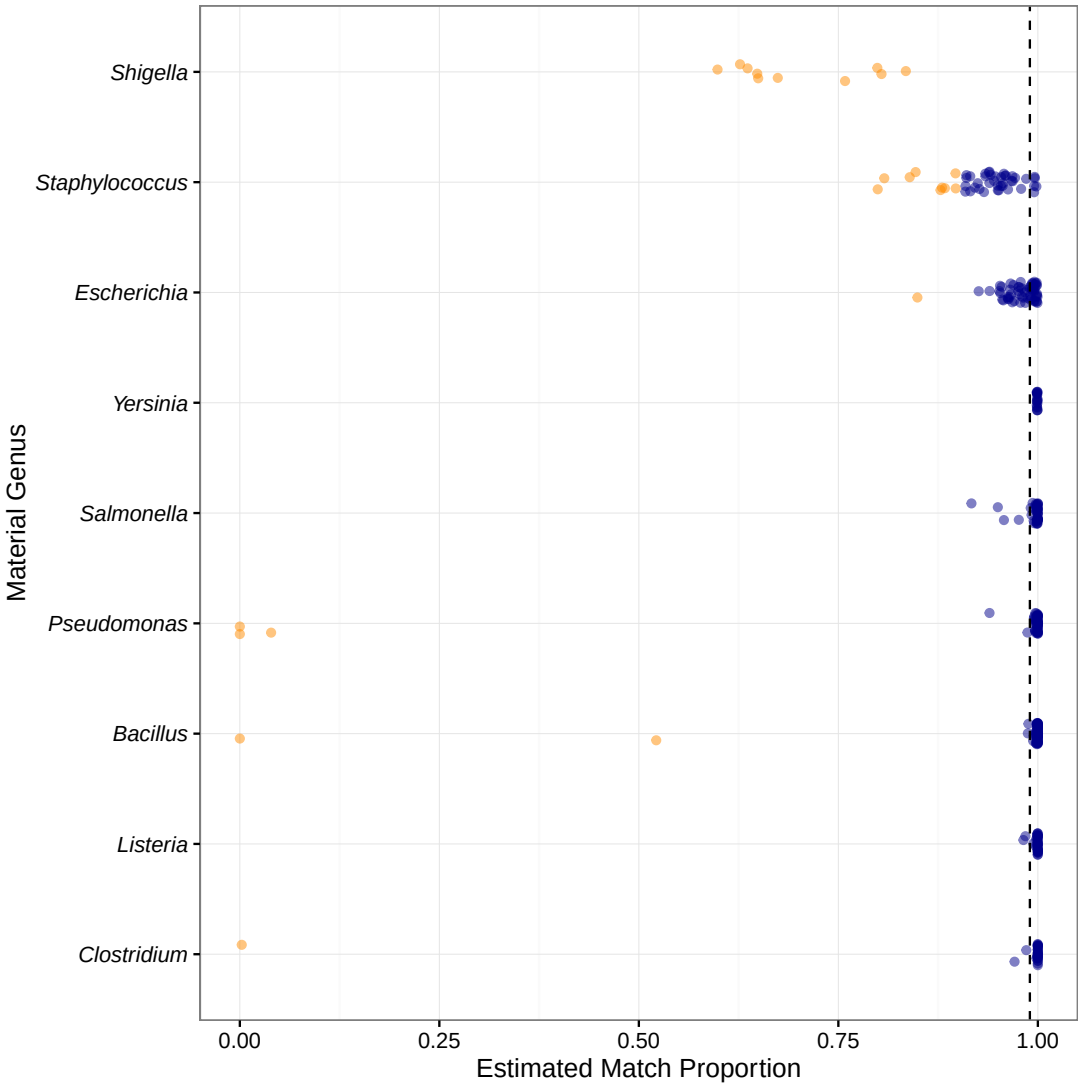

Figure 2. Species level estimated match proportion varies by material genus. The proportion of the material, simulated sequence data from individual genomes, was estimated by Pathoscope. The estimated match proportion is the total proportion of the material with taxonomic assignments to the genome species, 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

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

which was due to *Clostridium ljungdahlii* strain DSM13528 assigned the top proportion (0.998) instead of *C. autoenthanogenum*. False positive contmainants due to phylogentic similarity is not limited to closely related species or genus. *Escherichia coli* strain UMNK88 low match, was due to two bacteria in the same family as *E. coli*, Enterobacteriaceae, *Providencia stuartii* and *Salmonella enterica* subsp. enterica serovar Heidelberg with estimated proportions of 0.11 and 0.03 respectively.

False positives were also due to sharing of genetic material between organisms. An example of this type of false positive contaminant was phage. Phage were identified as a false positive contaminants at varing proportions for genomes from all 9 genera investigated (Fig. 3). Most noteably, low proportions of species level matches for *E. coli* and *Staphylococcus* is partly due to relatively higher proportions of matches to phage, compared to the other genera investigated. All of the false postive phage contaminants were specific to the taxonomy of the genome the sequence data was simulated from.

False positive contaminants were also due to errors in the database such as missclassified or unclassified sequences in the database, genome assemblies in the database including sequence data from organismal or reagent contaminants. *Bacillus subtilis* BEST7613 genome had low estimated species level match proportion due to *Synechocystis* sp. PCC 6803 substr. PCC-P being estimated as comprising 47% of the material (Kanesaki et al., 2012). *Synechocystis* is in a different phylum compared to *Bacillus*, cyanobacteria versus firmicutes. The high match proportion is potentially due to an error in the database. Low species level match proportions can also be due to the database containing unclassified sequence data for organisms highly similar to the material genome. For example the low match proportion for *Pseudomonas* strain FGI182 was due to matches to unclassified bacteria, bacterium 142412, and unclassified *Pseudomonas* species, *Pseudomonas* sp. HF-1. The low species proportion of species level matches for *Pseudomonas* strain TKP was also due to missclassified sequences (*Thioalkalivibrio sulfidophilus* strain HL-EbGr7 match proportion 0.0648).

The genome sequences used to populate the reference database can contain contaminants themselves. These database contaminants are responsible for additional false positive contaminants. The eukaryotic false positive contaminants are likely due to contaminants in the material or reagents used to generate the sequencign data used in the assembly **REF**. The low species proportion of species level matches for *Pseudomonas* strain TKP was also likely due to contminanted genome sequences in the database (wheat - *Triticum aestivum* match proportion 0.087).

Contaminant Detection Assessment

Finally, contaminant detection was assessed using simulated sequencing data from individual genomes. Contaminant datasets were developed by combining subsets of simulatd data from two organisms at defined proportions, with the larger proportion representing the microbial material and smaller proportion the contaminant (Fig. 1). We 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). For all of the genomes selected for the detection assessment study, the estimated proportion of material assigned to the correct species was 0.98 (Table 2).

The minimum contaminant proportion detected was 10×10^{-3} and 10×10^{-4} for most pairwise

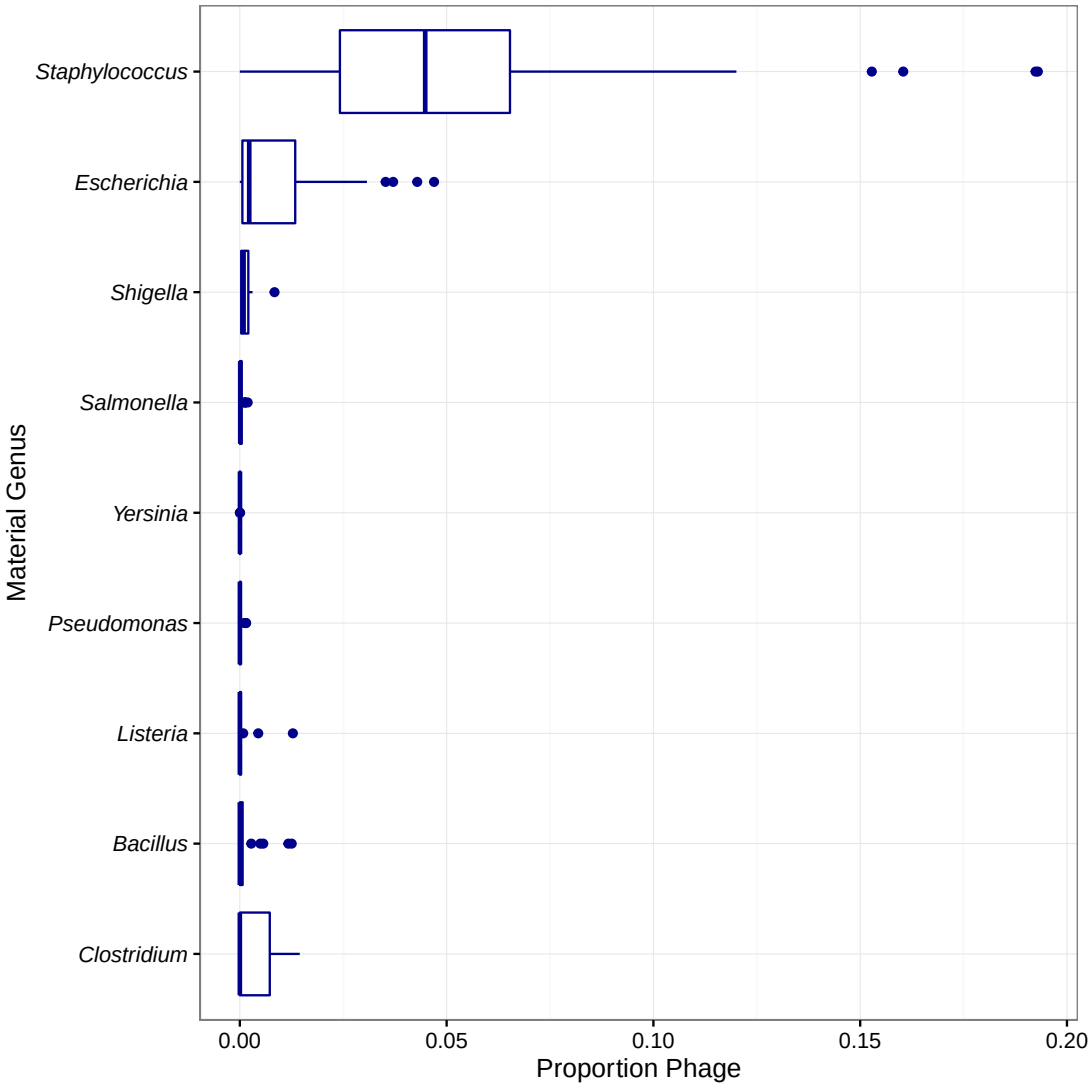

Figure 3. Estimated proportion of phage in the simulated single genome datasets by genera.

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. When available type strains were selected as the representative genome. 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.

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.

comparisons with a few notable exceptions. When *Yersina* was the simulated contaminant the minimum 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 Pearson's correlation coeficient 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. 4). 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 contamiant proportion residuals, $(estimated - true)/true$, were used to assess the accuracy of the Pathoscope contaminant proportion estimates (Fig. 6). 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

The potential for using the taxonomic sequence classification algorithm *Pathoscope* to detect contaminant DNA in microbial materials using whole genome sequencing data was evaluated. A baseline assessement of the contaminant DNA detection method using simulated sequencing data generated from individual genomes to characterize the types of false postive contaminants identified by the method was initially performed. The false positive contaminants were split into two categories (1) those due to an inability of the method to differentiate the material genome from the contaminant genome and (2) false positives due to errors in the reference database. Variation in contaminant detection was characterized by varying the material, the contaminant, 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.

A primary limitation of the proposed method is the observed false positive contamiants for single genome simulated sequencing data. Performing a baseline assement 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

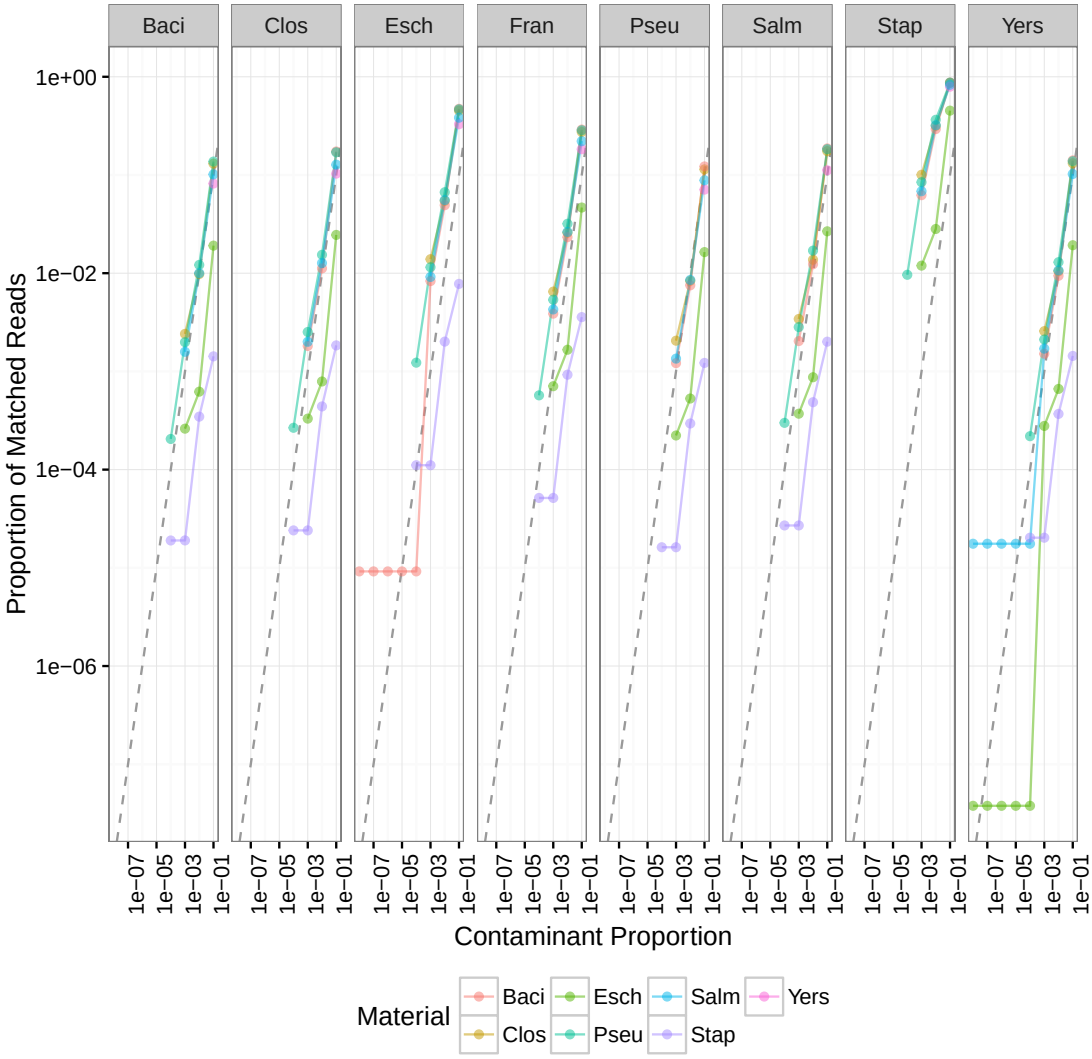

Figure 4. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

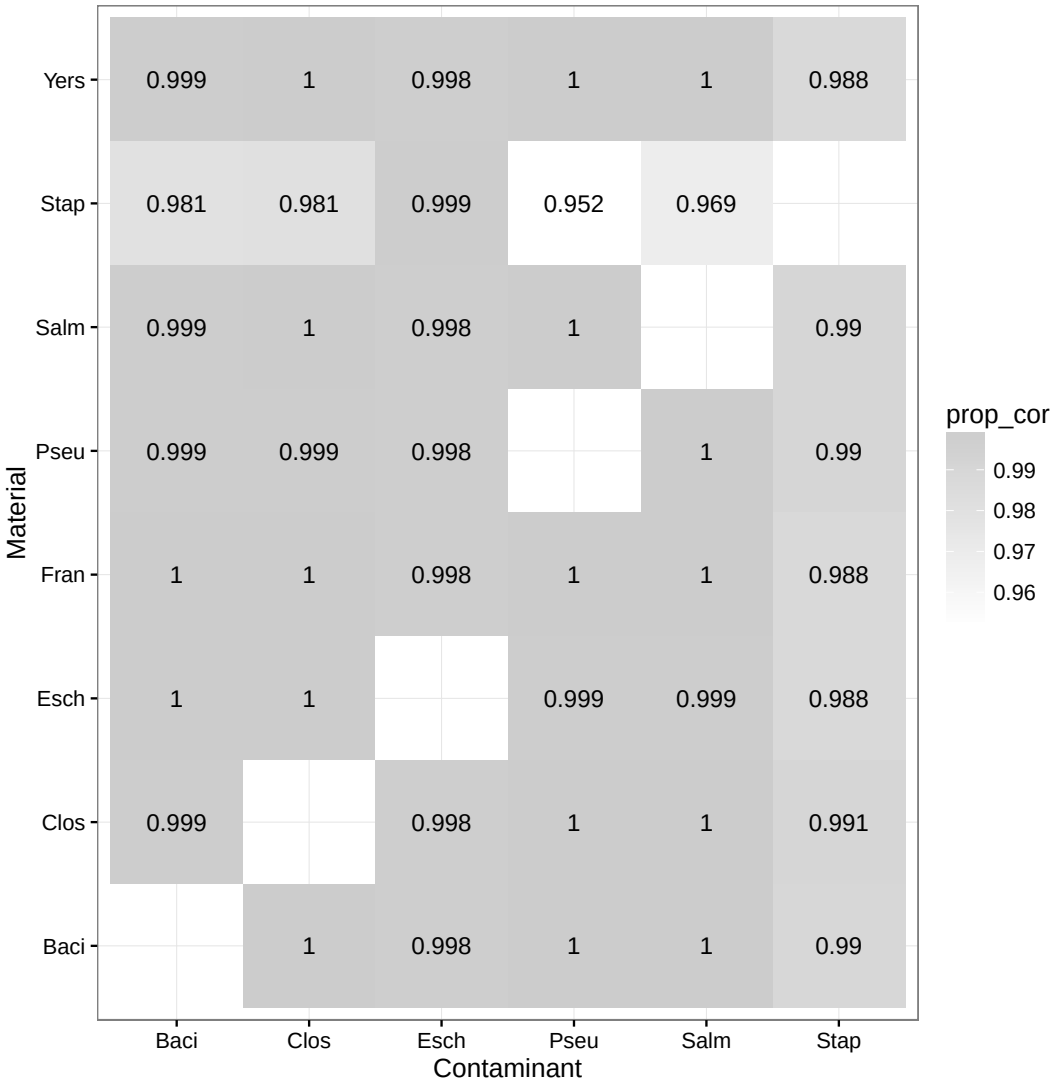

Figure 5. Pearson correlation coefficients for estimated and true contaminant proportions.

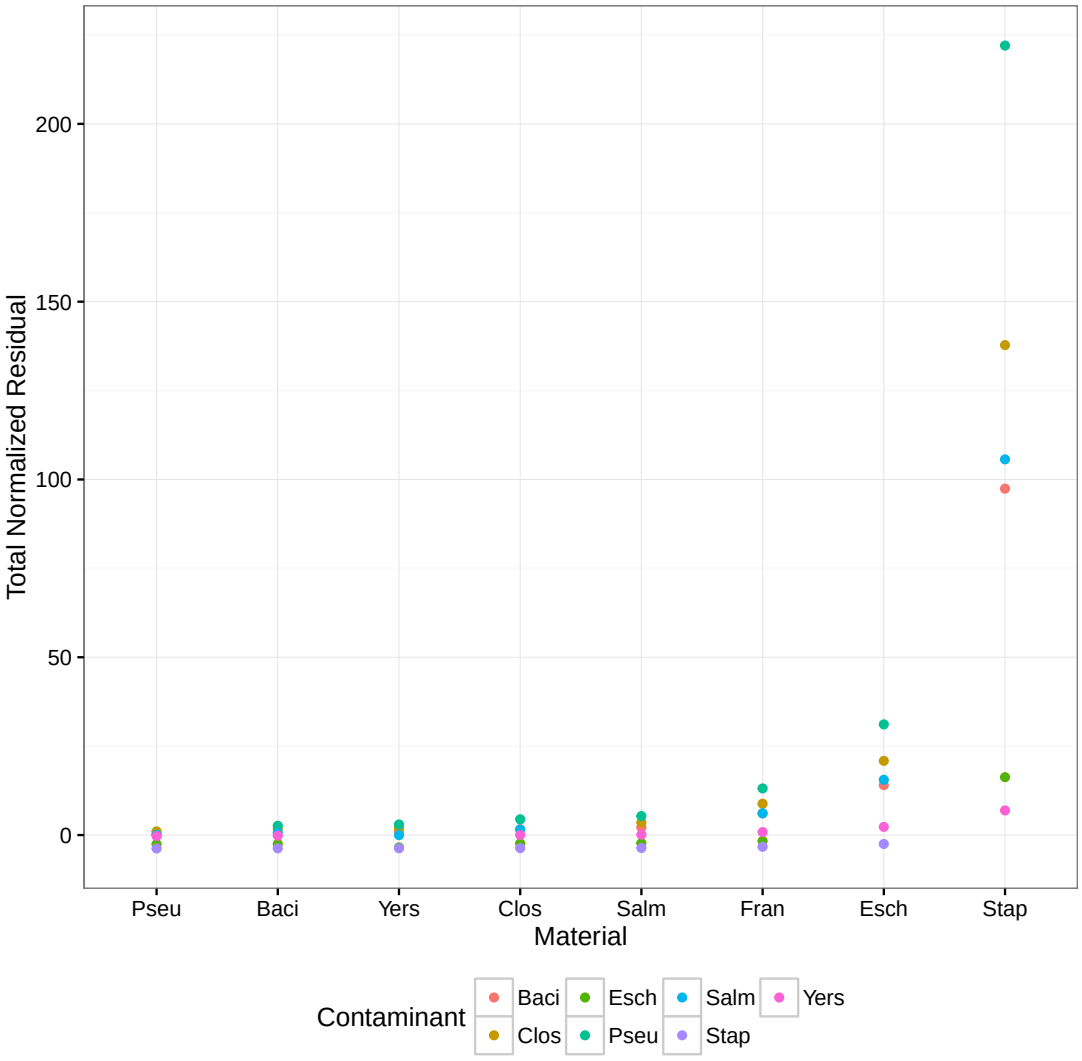

Figure 6. Total normalized residuals for pairwise combinations of material and contaminants.

the method's ability to detect contaminant DNA. 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 missclassified and unclassified sequence data would further help to reduce the proportion of false positive contaminants (REFERENCE ProDeGen). Pathoscope was used for this proof of concept study as the method uses the full reads and paired end information for taxonomic classification rather then 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 comtainant detection method for a specific microbial material.

Identification and characterization of low abundance contaminants is critical for when the material is used for high sensitivity assays such as PCR. The minimum contaminant proportion was consistent for most simulated contaminant datasets, *Yersinia* was only detected at a proportion of 0.1. The quantitative accuracy of the method varied by material and contaminant, but for all material-contaminant pairs, the Pathoscope estimated and true contaminant proportions were highly correlated. While not necessarily relevant to organismal contaminant detection, quantitative accuracy is relevant if the contaminant proportion is included in the report of analysis characterizing a material (Olson et al., 2016). Similar to the false positive contaminant baseline assessment. Similated data can be used to evaluate the minimal detectable contamiant proportion for specific contaminants of interest using different taxonomic assignment algorithms and databases. Additionally, sequencing at a higher depth would theoretically result in lower minimum detectable contaminant proportions and potentially increased quantitative accuracy due to the larger sample size (number of sequences).

CONCLUSIONS

With the continual decline in the cost of sequencing, advances in sequence analysis methods, whole genome sequencing combined with a taxonomic assignment algorithms provides a viable alternative to commonly used organismal contaminant dection methods such as culturing, microscopy, and PCR. The method presented here is suitable for detecting organismal contaminants in both genomic DNA and whole cell microbial materials with the only *a priori* assumptions about the contaminant is that it is present in the reference database. Furthermore, the method was shown to detect contaminants making up 10×10^{-3} proportion of cells in a high-throughput manner. Even with the rapid decrease in sequencing cost, whole genome sequencing is more expensive than culture and PCR based contaminant detection methods. However unlike culture and PCR based contaminant detection methods the data generated when whole genome sequencing data is used for organismal contaminant detection can also be used to further characterize the material's genome and potentially identify other characteristics of interest such as the presence of virulence or antibotic resistence genes.

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

Bazinet, A. L. and Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92.

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.

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.

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.

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.

Kanesaki, Y., Shiwa, Y., Tajima, N., Suzuki, M., Watanabe, S., Sato, N., Ikeuchi, M., and Yoshikawa, H. (2012). Identification of substrain-specific mutations by massively parallel whole-genome resequencing of synechocystis sp. pcc 6803. *DNA research*, 19(1):67–79.

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.

Menzel, P. and Krogh, A. (2016). Kaiju : Fast and sensitive taxonomic classification for metagenomics. *Nature communications*, 7(11257):1–9.

Olson, N. D., Zook, J. M., Samarov, D. V., Jackson, S. A., and Salit, M. L. (2016). Pepr: pipelines for evaluating prokaryotic references. *Analytical and bioanalytical chemistry*, 408(11):2975–2983.

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

Schmieder, R. and Edwards, R. (2011). 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*.

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.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3.

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