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

High sensitivity methods such as next generation sequencing and PCR are adversely impacted by organismal and DNA contaminants. Current methods such for detecting contaminants in microbial materials (genomic DNA and cultures) are not sensitivity enough and require either a known or culturable contaminant. Therefore, higher sensitivity methods not requiring *a priori* assumptions about the contaminant are needed. We demonstrate the use of whole genome sequencing data and a metagenomic taxonomic classification algorithm for assessing the organismal purity of a microbial material. Using whole genome sequencing and a taxonomic classification algorithm we characterized the types of false positive contaminants reported by the method and how the detectable contaminant concentration varies with material genome, contaminant genome, and contaminant proportion using simulated whole genome sequencing data. Using this method to characterize microbial material purity will help to ensure that the materials used to validate pathogen detection assays, generate genome assemblies for database submission, or benchmarking sequencing methods are free of contaminants adversely impacting measurement results.

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

INTRODUCTION

High sensitivity methods such as PCR and next generation sequencing require higher material and reagent purity than traditional microbiology methods such as culturing, biochemical tests, and microscopy. Issues related to reagent contaminants have been well documented and addressed with improved methods for removing contaminants (Woyke et al., 2011; Motley et al., 2014), negative controls (Jervis-Bardy et al., 2015), and post processing of sequence data (Mukherjee et al., 2015). However, contaminants in microbial materials such as non-axenic cellular materials and genomic materials with foreign DNA contaminants have only been addressed in data processing (Shrestha et al., 2013; Tennessen et al., 2015).

High sensitivity methods to detect and characterize contaminants in microbial materials are needed. Microbial materials free of contaminants are needed; to populate sequence databases (Parks et al., 2015), for mock communities used to validate metagenomic methods (Bokulich et al., 2016), biodetection assay validation (Ieven et al., 2013; Coates et al., 2011), basic research using model systems (Shrestha et al., 2013). General contaminant assessment is also needed for the characterization of microbial reference materials (Olson et al., 2016). The inclusion of contaminant characterization results in the reference material report of analysis allows users to properly determine whether the material is suitable for use in their application. Current methods for detecting contaminants in microbial materials use traditional methods such as culture, microscopy, and polymerase chain reaction (PCR). Culture and microscopy-based methods lack the required sensitivity for NGS and PCR applications, 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 detect contaminants in genomic DNA, the methods are limited as they can only detect targeted contaminants 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 while only requiring the contaminant has sequencing reads that differentiate it from the material strain.

Shotgun metagenomic sequencing is used to characterize environmental samples and detect pathogens in clinical samples and is also suitable for detecting contaminants in microbial materials. 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 materials, the material itself is sequenced, whereas 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 is a variety of classification algorithms with varying accuracy and computational performance (Bazinet and Cummings, 2012; Menzel et al., 2016). All methods require a reference database. In order to detect a contaminant in a microbial material, the contaminating organism (or an organism more closely related to the contaminant than the material) is 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 available alternative to current methods for detecting 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 contaminant DNA. This work first provides a baseline assessment of the method using simulated sequencing data from single microorganisms characterizing 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 using whole genome sequence data and metagenomic taxonomic classification methods foreign DNA in microbial materials. Simulated data from individual prokaryotic genomes was used to characterize the rate at which the method correctly classifies reads to the material species. 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) (NCBI Biosample accession SAMN02854573).

Assessing Taxonomic Composition

The taxonomic composition of simulated datasets was determined 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 contaminants not represented in the database and (2) it leverages 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`).

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

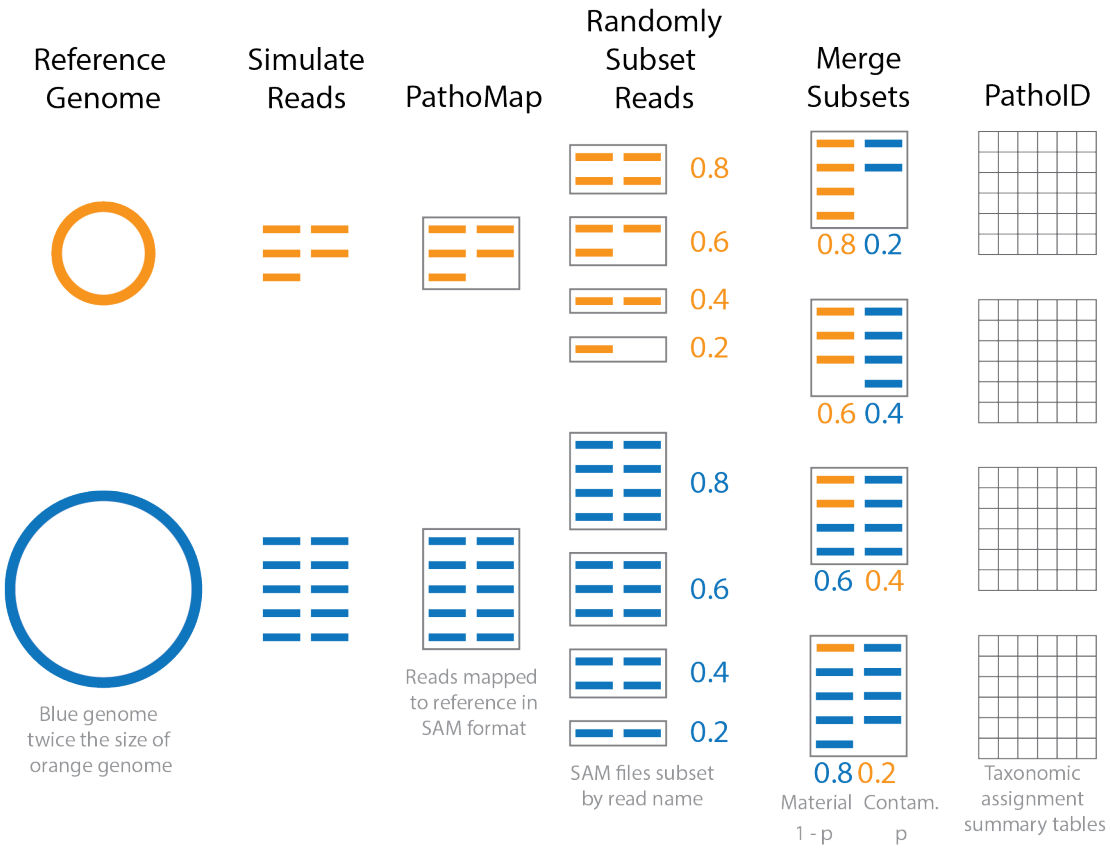

Figure 1. Diagram of the simulated contaminant dataset workflow for two individual genomes. Contaminant proportions 0.2 and 0.4 are used for demonstration purposes. The reads were initially simulated from individual genomes. The blue genome is twice the size of the orange genome and twice as many reads 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.

(`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* respectively. The taxonomic 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 genome over a range of contaminant concentrations. Representative genomes for 9 of the 10 genera 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 material 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 was 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 contaminated 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 the baseline assessment (Fig. 1). 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.

Bioinformatics Pipeline

To facilitate repeatability and transparency, a Docker (`www.docker.com`) container is available with pre-installed pipeline dependencies (`www.registry.hub.docker.com/u/natedolson/docker-pathosc`
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 the simulated datasets used in the study 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 files for this manuscript.

RESULTS

Baseline Assessment Using Individual Genomes

First, we 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 406 genomes, representing 10 different genera (Table 1). For 105 out of the 406 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 sequencing data 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 species level match proportions less than 0.99 come from *Shigella*, *Staphylococcus*, or *Escherichia*. Excluding *Shigella*, *Escherichia*, and *Staphylococcus* the median estimated proportion matching at the species level or higher is 0.9996. We characterized false positive 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) contaminants due to errors in the reference database.

Two genomes can be genomically indistinguishable if the majority of two genome sequences are highly similar. Phylogenetically closely related organisms are expected to have large genomic regions high levels of similarity. 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* (Lan and Reeves, 2002). When including matches to *E. coli* as species level matches, the median match proportions for *Shigella* genomes increase from 0.66 to 0.92. Another example of false positives at the species level due to phylogenetic similarity was low match percentage for *Clostridium autoethanogenum* strain DSM10061 which was due to *Clostridium ljungdahlii* strain DSM13528 assigned the top proportion (0.998) instead of *C. autoenthanogenum*. False positive contaminants due to phylogenetic similarity are not limited to closely related species or genus. *Escherichia coli* strain UMNK88 low match proportions, 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, such as

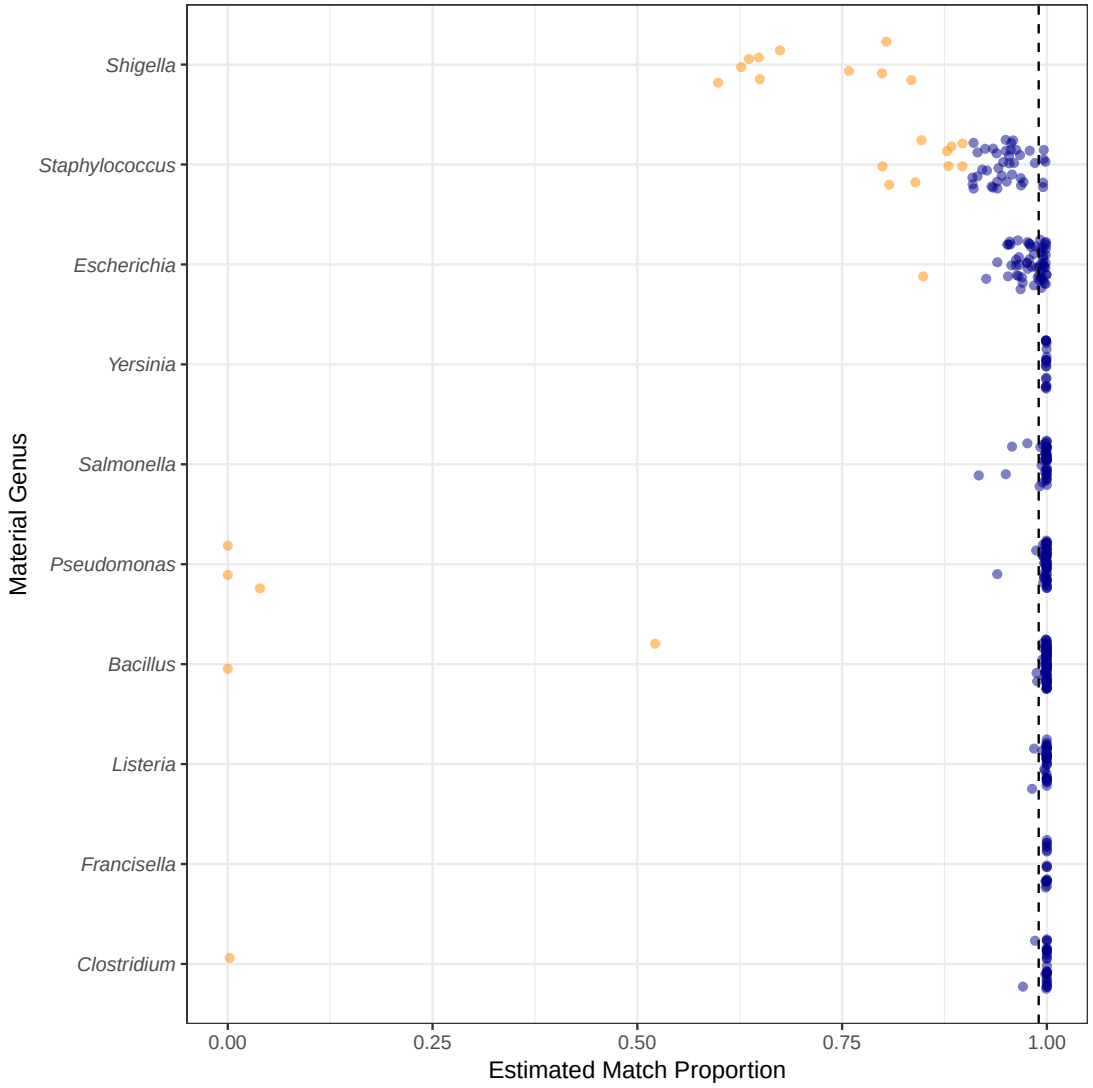

Figure 2. Species level estimated match proportion varies by material genus. The proportion of the material, simulated sequence data from individual genomes, was estimated by Pathoscope. The estimated match proportion is the total proportion of the material with taxonomic assignments to the genome species,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

the sharing of genetic material between bacteria and their phage. Phage was identified as false positive contaminants at varying proportions for genomes from all genera investigated, excluding *Francisella* (Fig. 3). Most notably, the low proportions of species level matches for *E. coli* and *Staphylococcus* are partly due to relatively higher proportions of matches to phage, compared to the other genera investigated. Based on phage names all of the false positive phage contaminants were specific to the taxonomy of the genome the sequence data was simulated from.

False positive contaminants were also due to potential errors in the database such as unclassified or misclassified sequences in the database and genome assemblies in the database including sequence data from organismal or reagent contaminants. Low species level match proportions can also be due to the database containing unclassified sequence data for organisms with genomic regions that are highly similar to regions of 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 potentially misclassified sequences (*Thioalkalivibrio sulfidophilus* strain HL-EbGr7 match proportion 0.0648). *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. *Synechocystis* is in a different phylum compared to *Bacillus*, cyanobacteria versus firmicutes is a false positive due to a misclassification. The *Bacillus subtilis* BEST7613 genome is a synthetic chimeric genome constructed from *Bacillus subtilis* BEST7613 and *Synechocystis* sp. PCC 6803 substr. PCC-P (Watanabe et al., 2012). The genome sequences used to populate the reference database can contain contaminants themselves (Parks et al., 2015). These database contaminants are responsible for additional false positive contaminants. The low species proportion of species level matches for *Pseudomonas* strain TKP was partially due to contaminated genome sequences in the database (wheat - *Triticum aestivum* match proportion 0.087). The eukaryotic false positive contaminants are likely due to contaminants in the material or reagents used to generate the sequencing data used in the assembly (Parks et al., 2015).

Contaminant Detection Assessment

Finally, contaminant detection was assessed using simulated sequencing data from individual genomes. Contaminant datasets were developed by combining subsets of simulated 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 greater than 0.98 (Table 2).

The minimum contaminant proportion detected was 1^{-3} and 1^{-4} for most pairwise comparisons with a few notable exceptions. When *Y. pestis* was the simulated contaminant the minimum detected proportion was 0.1 for all material strains. For all simulated datasets where *F. tularensis* was the contaminant the contaminant was not detected. Conversely, a few contaminants were detected at lower proportions,

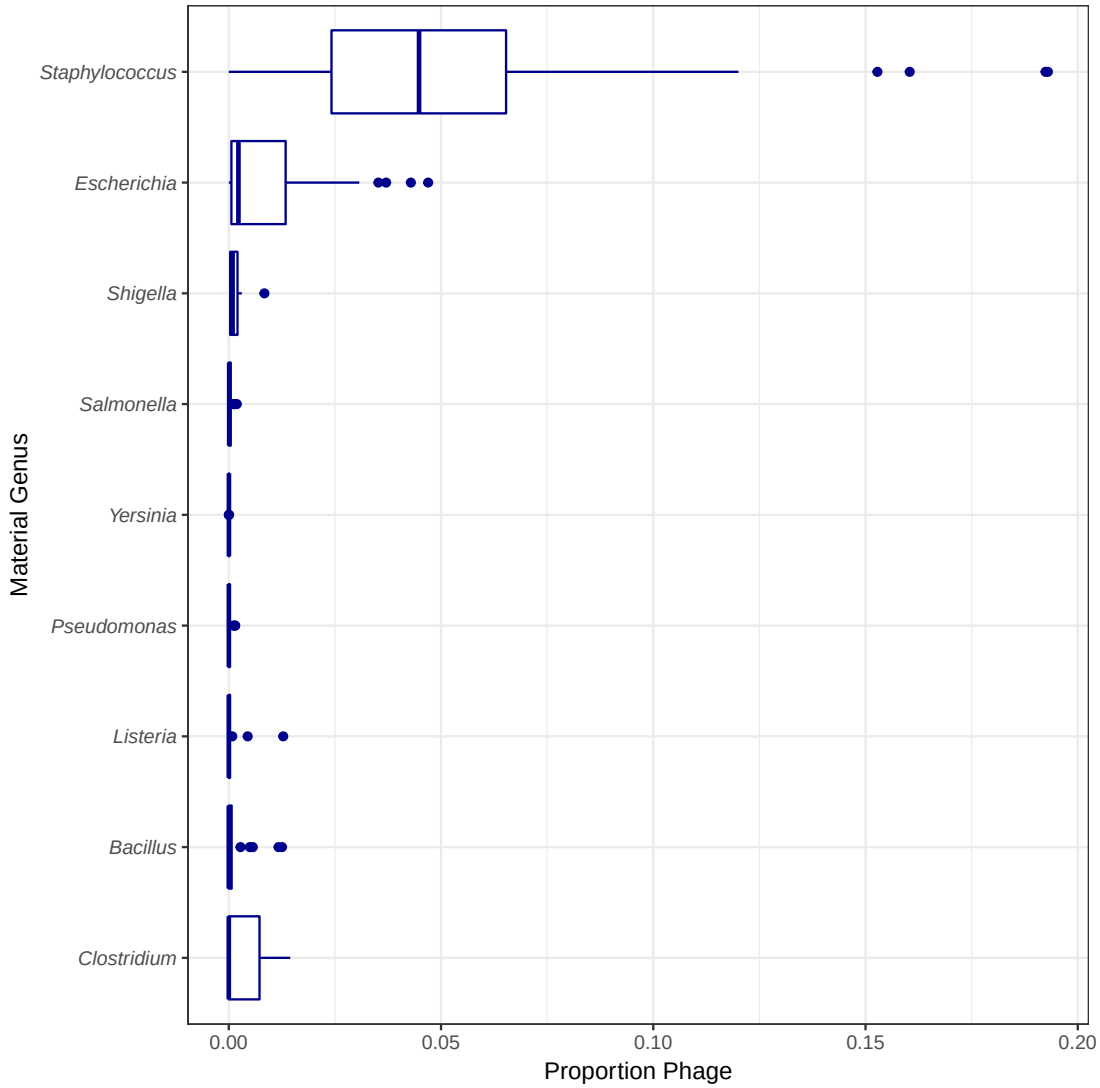

Figure 3. Estimated proportion of phage in the simulated single genome datasets by genera. Proportions based on the final estimated proportions for all phage.

Representative Strain	Species	Aligned Reads	C Mb	C Acc	P Mb	P Acc
Bacillus anthracis str. Ames	1.00	227270	5.23	AE016879.1		
Clostridium botulinum A str. Hall	1.00	163500	3.76	CP000727.1		
Escherichia coli O157:H7 str. EC4115	0.98	247990	5.57	CP001164.1	0.13	CP001163.1, CP001165.1
Francisella tularensis subsp. tularensis SCHU S4	1.00	82290	1.89	AJ749949.2		
Pseudomonas aeruginosa PAO1	1.00	272360	6.26	AE004091.2		
Salmonella enterica subsp. enterica serovar Typhimurium str. D23580	1.00	212140	4.88	FN424405.1		
Staphylococcus aureus subsp. aureus ED133	0.98	123150	2.83	CP001996.1		
Yersinia pestis CO92	1.00	209970	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). Aligned Reads is the number of simulated reads aligned to the database by PathoMap.

1^{-8}, when *Yersinia* was contaminated with *E. coli* as well as when *S. enterica* and *E. coli* contaminated with *B. anthracis*. The contaminants detected at lower proportions were false positives in the material single genome simulated datasets. For the *E. coli* material dataset with no simulated contaminants, Bacillus sp. SXB had an estimated proportion of 9.2×10^{-6} resulting in an artificially low contaminant detection proportion. The simulated contaminant-free Y. pestis material dataset had two false positives resulting in artificially low contaminant detection proportions Salmonella enterica subsp. enterica serovar Typhi str. CT18 with an estimated proportion of 1.76×10^{-5} and Escherichia coli O104:H4 str. 2011C-3493 with an estimated proportion of 3.77×10^{-8}.

Pearson's correlation coefficient was used to measure the correlation between the estimated contaminant and true contaminant proportions for simulated contaminant proportions greater than 0.1×10^{-5}. 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 contaminant proportion residuals, $(estimated - true)/true$, were 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

The potential for using whole genome sequencing data and taxonomic sequence classification algorithm *Pathoscope* to detect contaminant DNA in microbial materials was evaluated. The method requires no *a priori* information about the contaminant, therefore the method is able to identify organisms that are known contaminate the type of material being analyzed as well as previously unknown contaminants. Additionally, as whole genome sequencing can be performed on genomic DNA and culture (after DNA extraction) the method is appropriate for both types of microbial material. A baseline assessment of the contaminant DNA detection method using simulated sequencing data generated from individual genomes to characterize the types of false positive contaminants identified by the method was 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) those 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 1×10^{-3} for most pairwise contaminant-material combinations. A contaminant proportion of 1×10^{-3} is equivalent to 1 contaminant cells per 1,000 cells in a microbial material, or 1,000 contaminant cells in 1 mL of a 1×10^{6} culture. 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 contaminants for single genome simulated sequencing data. Baseline assessments using simulated sequence data from the microbial material's genome sequence is a first step in determining the impact of false positive on the method's ability to detect contaminant DNA. Additionally, choosing the appropriate database and taxonomic assignment algorithm may help reduce the number of false positive. Removing sequences from

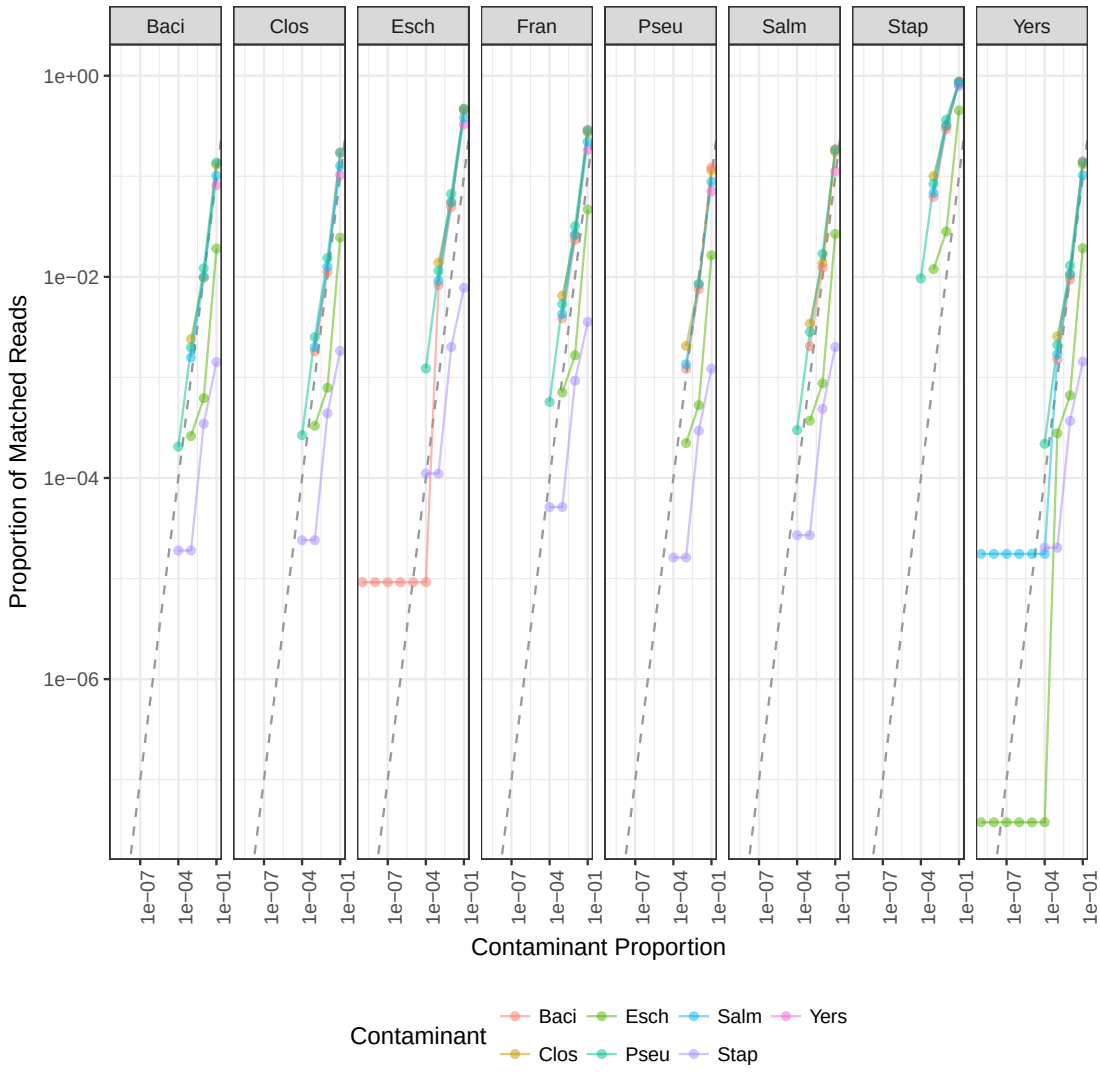

Figure 4. The relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus. Plots are split by the material genus with line and point color indicating contaminant genus.

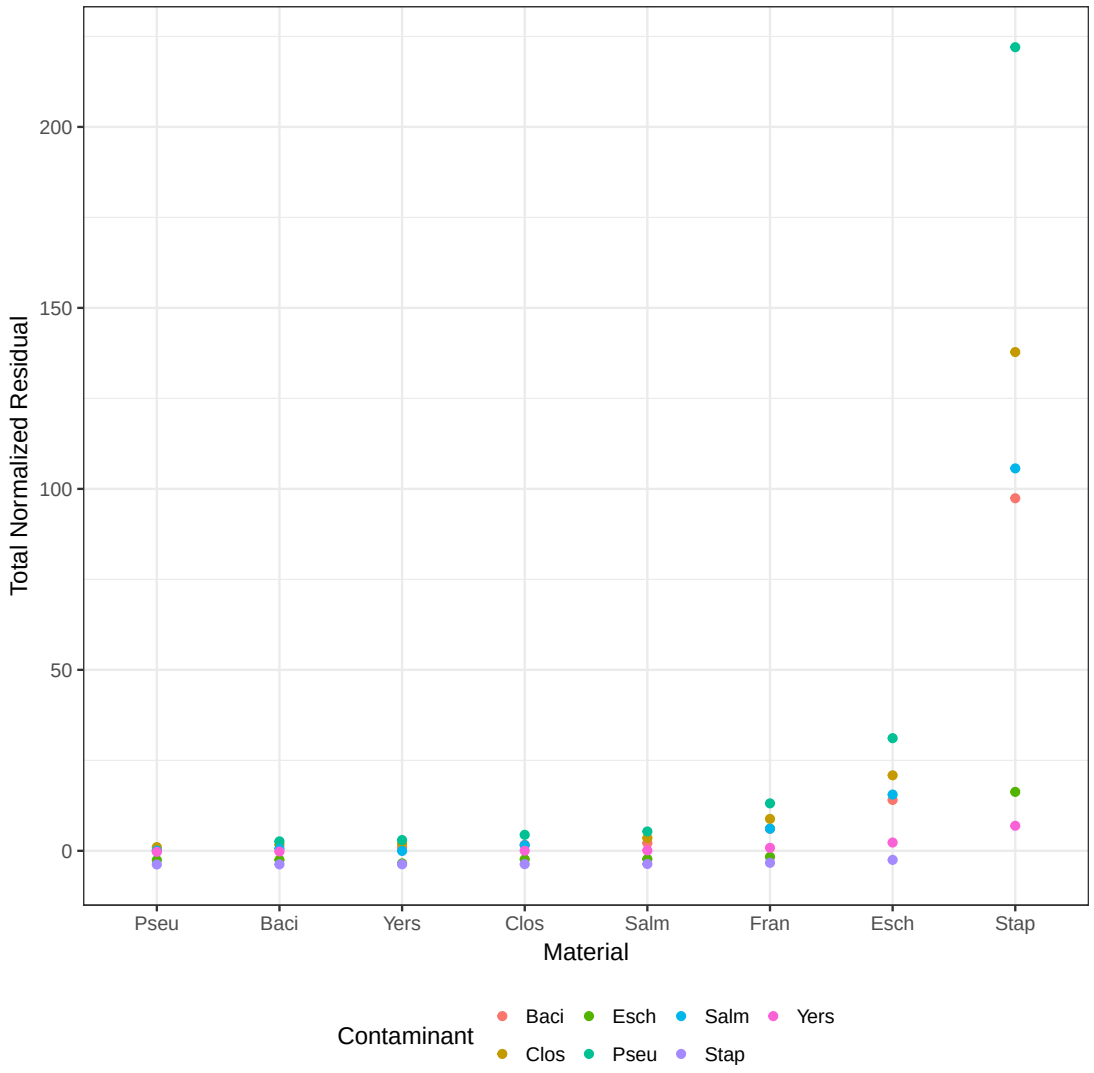

Figure 5. Total normalized residuals for pairwise combinations of material and contaminants.

the database for irrelevant contaminants, such as phage, plasmids, vectors, and multicellular eukaryotes would reduce the proportion of false positives. By excluding irrelevant contaminants from the database sequencing reads aligning to these irrelevant sequences would no longer result in false positive contaminants. Methods for excluding sequence data from a reference database is dependent on the classification algorithm used. For example, user-specified sequence data from the reference database by PathoScope using the PathoDB function. Users should be cautious when removing sequences from a reference database when analyzing real data. For example vector sequences from contaminants in sequencing, reagents may be incorrectly classified as an organism contaminant. Similarly, using a curated database free of misclassified and unclassified sequence data would further help to reduce the proportion of false positive contaminants (Tennessen et al., 2015). The *Bacillus subtilis-Synechocystis* chimeric genome which appeared to be a false positive contaminant in the baseline assessment part of this study. 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. Longer sequences allow for better discrimination between highly similar sequences as the sequences contain more information. However, evaluating multiple algorithms using simulated data for the material genome of interest, similar to what was done in the baseline assessment part of this study, can help determine the optimal classification algorithm for a specific microbial material. Regardless of the method and database used, contaminants identified by the method should be evaluated considering the impact of the contaminant on the intended application or the likelihood the contaminant is a false positive. For all settings, research, clinical, regulatory, and attribution the contaminant detection method should be validated for the intended application. As false positive contaminants are database and taxonomic assignment algorithm dependent, additional work is needed to improve database curation and data authentication efforts as well as characterization of taxonomic assignment algorithm performance.

Identification and characterization of low abundance contaminants in a non-targeted manner is critical for a material used in high sensitivity assays such as PCR. The presence of contaminants that are known to interfere with an assay can be tested using specific assays for that contaminant. However, specific assays cannot be developed or used to contaminants that are not known to adversely impact an assay. The minimum detected contaminant proportion ranged from 1^{-3} to 1^{-4} for most simulated contaminant datasets. As the individual datasets were simulated at 20X coverage less than 300,000 reads were simulated for each dataset on average less than 3 reads were spiked into the material datasets for simulated contaminant proportions less than 1^{-4} (Table 2). Unexpectedly low contaminant proportions, 1^{-8}, were detected for *E. coli* contaminated with *B. anthracis* and *Y. pestis* contaminated with *S. enterica* and *E. coli*. The low detection proportions were due to false positive contaminants present in the simulated material single genome dataset used to generate the contaminant mixtures. For datasets with *Y. pestis* as the simulated contaminant the minimum detected contaminant proportion was 0.1 and *F. tularensis* was not detected in any simulated contaminant datasets. It is unclear why *Y. pestis* was detected at a higher proportion relative to the other datasets, 1^{-1} versus 1^{-3}, and *F. tularensis* was not detected at all. One possible reason for the lower contaminant detect for these two organisms is that there are fewer genomes is the database for these two genera. Additionally, the *F. tularensis* dataset is much smaller relative to the other genera, less than 90,000 reads. With fewer reads in the dataset and genomes in the database, the probability that the randomly selected subset of reads spiked into the simulated material dataset contains reads allowing for contaminant detection is lower. While the minimum detected contaminant proportion is important for assessing the suitability of microbial materials for specific applications, quantitative accuracy of the contaminant detection method is important for general material characterization.

The quantitative accuracy of the method varied by material and contaminant. For all material-contaminant pairs, the Pathoscope estimated and true contaminant proportions were highly correlated. Quantitative accuracy in contaminant proportions is important for applications where acceptable contaminant proportion thresholds are established. For example, a microbial material with a contaminant proportion of 1^{-5} may be acceptable for use in an assay where the contaminant adversely impact an assay when present in proportions greater than 1^{-4}. Quantitative accuracy is also relevant when performing a general characterization of the microbial material. General contaminant characterization is appropriate for reference materials with more than one use case such as the NIST microbial genomic reference materials (NIST RM8375)(Olson et al., 2016). Similar to the false positive contaminant baseline assessment, simulated data can be used to evaluate the minimal detectable contaminant proportion for specific organisms using different taxonomic assignment algorithms and databases.

CONCLUSIONS

With the continual decline in the cost of sequencing, advances in sequence analysis methods, whole genome sequencing combined with taxonomic assignment algorithms provides a viable alternative to commonly used organismal contaminant detection 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 are that it is present in the reference database. Furthermore, the method was shown to detect contaminants making up 1^{-3} proportion of cells in a high-throughput manner. With the rapid decrease in sequencing cost and ability to detect unknown contaminants at low concentrations, whole genome sequencing is a viable alternative to culture and PCR-based contaminant detection methods.

ACKNOWLEDGMENTS

The authors would like to thank 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-15-T-00019 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.

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., Dutton, R. J., Turnbaugh, P. J., Knight, R., and Caporaso, J. G. (2016). mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems*, 1(5).

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.

Jervis-Bardy, J., Leong, L. E., Marri, S., Smith, R. J., Choo, J. M., Smith-Vaughan, H. C., Nosworthy, E., Morris, P. S., OLeary, S., Rogers, G. B., et al. (2015). Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of illumina miseq data. *Microbiome*, 3(1):1.

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

Motley, S. T., Picuri, J. M., Crowder, C. D., Minich, J. J., Hofstadler, S. A., and Eshoo, M. W. (2014). Improved multiple displacement amplification (imda) and ultraclean reagents. *BMC genomics*, 15(1):1.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by illumina phix control. *Standards in genomic sciences*, 10(1):1.

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.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Checkm:

assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055.

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.

Tennessen, K., Andersen, E., Clingenpeel, S., Rinke, C., Lundberg, D. S., Han, J., Dangl, J. L., Ivanova, N., Woyke, T., Kyrpides, N., et al. (2015). Prodege: a computational protocol for fully automated decontamination of genomes. *The ISME journal*.

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

Watanabe, S., Shiwa, Y., Itaya, M., and Yoshikawa, H. (2012). Complete sequence of the first chimera genome constructed by cloning the whole genome of synechocystis strain pcc6803 into the bacillus subtilis 168 genome. *Journal of bacteriology*, 194(24):7007–7007.

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

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., and Cheng, J.-F. (2011). Decontamination of mda reagents for single cell whole genome amplification. *PloS one*, 6(10):e26161.