

# Using metagenomic methods to detect organismal contaminants in microbial materials.

Nathan D. Olson<sup>1</sup>, Justin Zook<sup>1</sup>, Jayne Morrow<sup>1</sup>, and Nancy Lin<sup>1</sup>

<sup>1</sup>Material Measurement Laboratory, National Institute of Standards and Technology

## ABSTRACT

High sensitivity methods as next generation sequencing and PCR are adversely impacted by organismal and DNA contaminants. Current methods for detecting contaminants in microbial materials (genomic DNA and cultures) are not sensitive 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 (WGS) data and a metagenomic taxonomic classification algorithm for assessing the organismal purity of a microbial material. With this method for assessing organismal purity we characterized the types of false positive contaminants reported by the method and how the dependence of detectable contaminant concentration on material genome, contaminant genome, and contaminant proportion using simulated WGS 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, and to 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 polymerase chain reaction (PCR) and next generation sequencing (NGS) 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 negative controls (Jervis-Bardy et al., 2015), improved methods for removing contaminants (Woyke et al., 2011; Motley et al., 2014), 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).

Microbial materials essentially 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), to validate biodetection assays (Ieven et al., 2013; Coates et al., 2011), and for 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 a reference material report of analysis that allows users to properly determine whether the material is suitable for their application. Current methods for detecting contaminants in microbial materials rely on traditional methods such as culture, microscopy, and PCR. Culture and microscopy-based methods lack the required sensitivity for NGS and PCR applications, are not appropriate for genomic DNA materials, and assume the contaminants are phenotypically distinct from the material isolate they are contaminating. While PCR-based methods can detect contaminants in genomic DNA, the methods are limited as they can only detect targeted contaminants and are 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 all DNA in a sample, 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) must be in the database. As taxonomic classification algorithms are constantly improving, reference databases are expanding, and the cost of sequencing is dropping, 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 its 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 and metagenomic taxonomic classification methods were used to detect and identify foreign DNA in microbial materials. Simulated data from individual prokaryotic genomes were used to characterize how well the method correctly classifies reads to the material species or higher. To evaluate contaminant detection we used datasets comprised of pairwise combinations of simulated reads from individual genomes.

### Simulating Sequencing Data

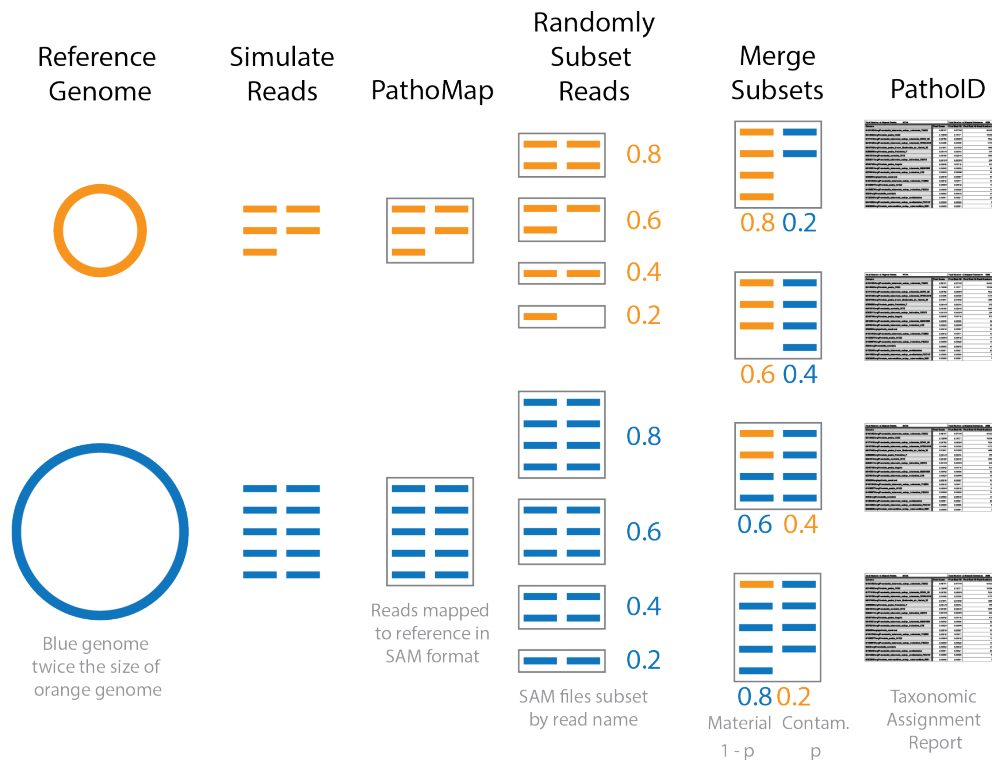
To approximate real sequencing data reads were simulated using an empirical error model and insert size distribution. Whole genome sequencing data were simulated using the ART sequencing read simulator (Huang et al., 2012). Reads were simulated with the Illumina MiSeq error model for  $2 \times 230$  base pair (bp) paired-end reads with an insert size of  $690 \pm 10$  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), and (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](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). These genera were select based on relevance to public health and biothreat detection. We will refer to the genome used to generate the reads as the material genome. The genomes included in



**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. 0.8 material and 0.2 contaminant) 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.

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* 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 *Escherichia* and *Shigella*, as the genus *Shigella* and species *Escherichia coli* are not phylogenetically resolved (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 (Fig. 1). The simulated datasets were randomly subsampled at defined proportions, with  $p$  representing the proportion of reads from the contaminant, and  $1 - p$  representing the proportion of reads from the material 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 con-

115 taminant datasets. This approach simulates the proportions of cells in a contaminated material and not  
116 the amount of DNA, assuming unbiased DNA extraction. Organisms with larger genomes therefore have  
117 more simulated reads.

118 To generate the simulated contaminant datasets single organism simulated datasets were first gener-  
119 ated for the 8 representative genomes using the same methods as used in the baseline assessment (Fig.  
120 1). The resulting simulated sequencing data was first processed using the PathoQC and PathoMap steps  
121 in the PathoScope pipeline. The output from the PathoMap step (SAM file, sequence alignment file  
122 <https://samtools.github.io/hts-specs/SAMv1.pdf>) for the material and contaminant  
123 datasets were subsampled as described above then combined. The resulting SAM file was processed by  
124 PathoID, the third step in the PathoScope pipeline. Subsampling the sam files instead of the simulated  
125 sequence files greatly reduces the computational cost of the analysis, as the simulated reads were only  
126 processed by the first two steps in PathoScope pipeline once rather than for every simulated contami-  
127 nant dataset. Quantitative accuracy of the contaminant detection method was assessed by comparing the  
128 defined contaminant proportion, true proportion, to the PathoScope contaminant proportions, estimated  
129 proportions, for simulated datasets with contaminant proportions greater than  $10^{-5}$ . Pearson's correla-  
130 tion coefficient was used to evaluate agreement between the true and estimated proportions. The nor-  
131 malized contaminant proportion residuals,  $(estimated - true)/true$ , were compared performance across  
132 material and contaminant combinations.

### 133 Bioinformatics Pipeline

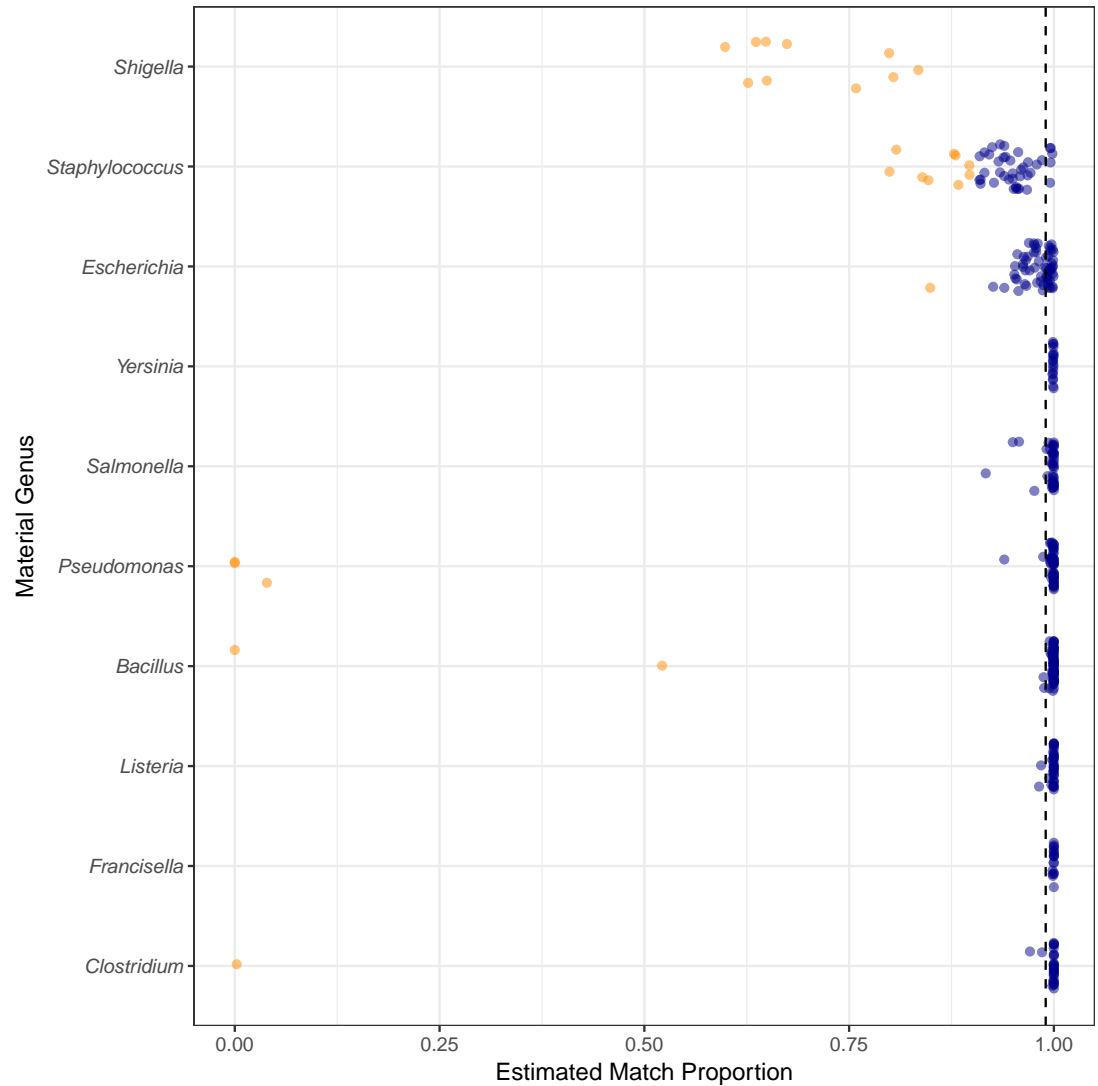
134 To facilitate repeatability and transparency, a Docker ([www.docker.com](http://www.docker.com)) container is available with  
135 pre-installed pipeline dependencies ([www.registry.hub.docker.com/u/natedolson/docker-pathos](http://www.registry.hub.docker.com/u/natedolson/docker-pathos))  
136 The scripts used to run the simulations are available at [https://github.com/nate-d-olson/](https://github.com/nate-d-olson/genomic_purity)  
137 `genomic_purity`. Additionally, seed numbers for the random number generator were randomly  
138 assigned and recorded for each dataset so the simulated datasets used in the study could be regener-  
139 ated. PathoScope results were processed and analyzed using the statistical programming language R (R  
140 Core Team, 2016), and intermediate analysis and data summaries were organized using ProjectTemplate  
141 (White, 2014) and archived in a GitHub repository ([https://github.com/nate-d-olson/genomic\\_](https://github.com/nate-d-olson/genomic_purity_analysis)  
142 `purity_analysis`) along with the source files for this manuscript.

## 143 RESULTS

### 144 Baseline Assessment Using Individual Genomes

145 First, we assessed baseline performance of the proposed contaminant detection method. We applied out  
146 contaminant DNA detection method to simulated sequencing data from individual genomes. All read  
147 assigned to a different taxa than the genome the reads were simulated from were defined as false positive  
148 contaminants, this assumes the genome sequence is contaminant free. Our analysis included taxonomic  
149 classification results for simulated sequencing data from 406 genomes, representing 10 different genera  
150 (Table 1). The method was evaluated using estimated proportion of species level matches. The estimated  
151 match proportion is the sum of the Final Guess values in the PathoScope output for all correct species  
152 level matches. For 105 of the 406 genomes, PathoScope estimated that less than 99% of the material  
153 was the expected species (Fig. 2). Of these 105 genomes, the estimated proportion of the sequencing  
154 data identified as the correct species varied by genus. None of the *Shigella* genomes and only five of the  
155 49 *Staphylococcus* genomes had estimated proportions for the correct species greater than 0.9. 87 of the  
156 105 genomes with estimated species level match proportions less than 0.99 come from *Shigella*, *Staphy-*  
157 *lococcus*, or *Escherichia*. Excluding *Shigella*, *Escherichia*, and *Staphylococcus*, the median estimated  
158 proportion matching at the species level or higher is 0.9996. We characterized false positive contaminants  
159 detected in genomes from the genera *Shigella*, *Escherichia*, and *Staphylococcus*, as well as genomes of  
160 other species, match proportions less than 0.9. Two types of false positive contaminants were identified:  
161 (1) contaminants that were genomically indistinguishable from the material and (2) contaminants due to  
162 errors in the reference database.

163 Two genomes can be genomically indistinguishable if the majority of the two genome sequences  
164 are highly similar. Phylogenetically closely related organisms are expected to have large genomic re-  
165 gions with high levels of similarity. Phylogenetic similarity is at least partially responsible for the low  
166 species level match proportion for *Shigella* and *Escherichia*, as *Shigella* is not phylogenetically distinct  
167 from *E. coli* (Lan and Reeves, 2002). When including matches to *E. coli* as species level matches, the



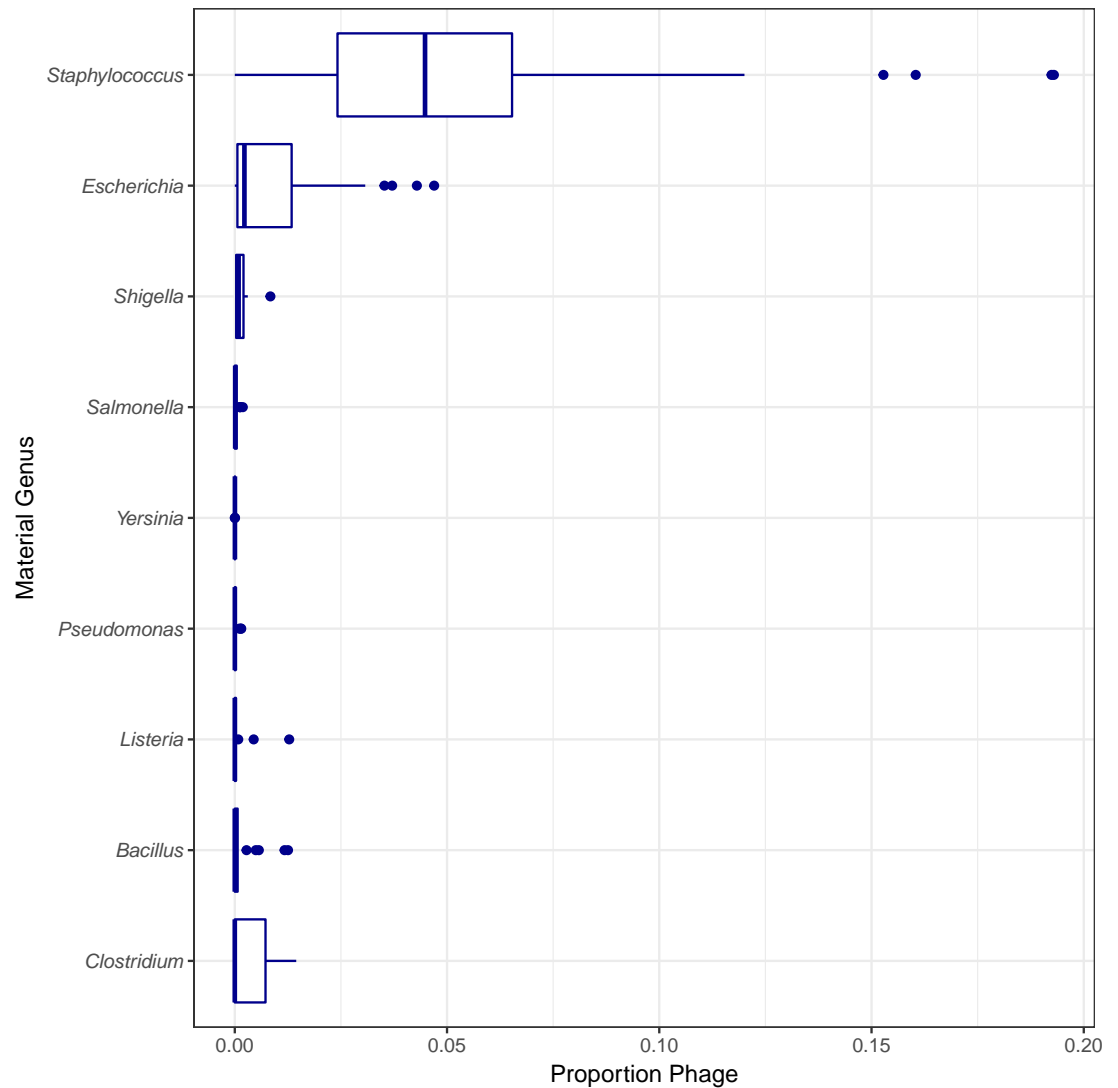
**Figure 2.** Species level or higher estimated match proportion varies by material genus. The estimated match proportion is the total proportion of the material with correct taxonomic assignments to the genome species, subspecies, strain, or isolate level. The proportions used are the Final Guess values in the PathoScope results table. 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 or equal to 0.90

Genus	N	Genome Size (Mb)
<i>Bacillus</i>	76	5.05 (3.07-7.59)
<i>Clostridium</i>	32	4.02 (2.55-6.67)
<i>Escherichia</i>	62	5.11 (3.98-5.86)
<i>Francisella</i>	18	1.89 (1.85-2.05)
<i>Listeria</i>	39	2.97 (2.78-3.11)
<i>Pseudomonas</i>	57	6.18 (4.17-7.01)
<i>Salmonella</i>	44	4.88 (4.46-5.27)
<i>Shigella</i>	10	4.74 (4.48-5.22)
<i>Staphylococcus</i>	49	2.82 (2.69-3.08)
<i>Yersinia</i>	19	4.73 (4.62-4.94)

**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).

median match proportions for *Shigella* genomes increases 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, where *Clostridium ljungdahlii* strain DSM13528 was assigned the top proportion of reads (0.998) instead of *C. autoethanogenum*. False positive contaminants due to phylogenetic similarity are not limited to a closely related species or genus. *Escherichia coli* strain UMNK88 low match proportions were due to two bacteria in the same family as *E. coli* (Enterobacteriaceae): *Providencia stuartii* and *Salmonella enterica* subsp. *enterica* serovar Heidelberg, which had estimated proportions of 0.11 and 0.03, respectively. False positives were also due to shared genetic material between bacteria and their phage. Phage were identified as false positive contaminants at varying proportions for genomes from all genera investigated, excluding *Francisella* (Fig. 3). 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 and the presence of genome assemblies containing in the database including sequence data from organismal or reagent contaminants. Low estimated 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) and is a false positive due to a misclassification. The *Bacillus subtilis* BEST7613 genome in the database is the genome of a synthetic chimeric genome constructed from *Bacillus subtilis* BEST7613 and *Synechocystis* sp. PCC 6803 substr. PCC-P not *Bacillus subtilis* BEST7613 (Watanabe et al., 2012). The *Bacillus subtilis* BEST7613 genome assembly (GenBank Accession GCA\_000328745.1) was flagged as an anomalous assembly and removed from the RefSeq database. 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 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 eukaryotic DNA extract or reagents used to generate the sequencing data for the assembly (Parks et al., 2015).



**Figure 3.** Estimated total proportion of phage in the simulated single genome datasets by genera. Final Guess values reported by PathoScope used to calculate estimated total proportions. No phage were reported by PathoScope in any of the *Francisella* genomes.

Representative Strain	Species	Aligned Reads	C Mb	C Acc
<i>Bacillus anthracis</i> str. Ames	1.00	227270	5.23	AE01
<i>Clostridium botulinum</i> A str. Hall	1.00	163500	3.76	CP00
<i>Escherichia coli</i> O157:H7 str. EC4115	0.98	247990	5.57	CP00
<i>Francisella tularensis</i> subsp. <i>tularensis</i> SCHU S4	1.00	82290	1.89	AJ74
<i>Pseudomonas aeruginosa</i> PAO1	1.00	272360	6.26	AE00
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium str. D23580	1.00	212140	4.88	FN42
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> ED133	0.98	123150	2.83	CP00
<i>Yersinia pestis</i> CO92	1.00	209970	4.65	AL59

**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. Aligned Reads is the number of simulated reads aligned to the database by PathoMap. DNA size (Mb) and Genbank accession numbers (Acc) are indicated for chromosomes (C) and plasmids (P).

## Contaminant Detection Assessment

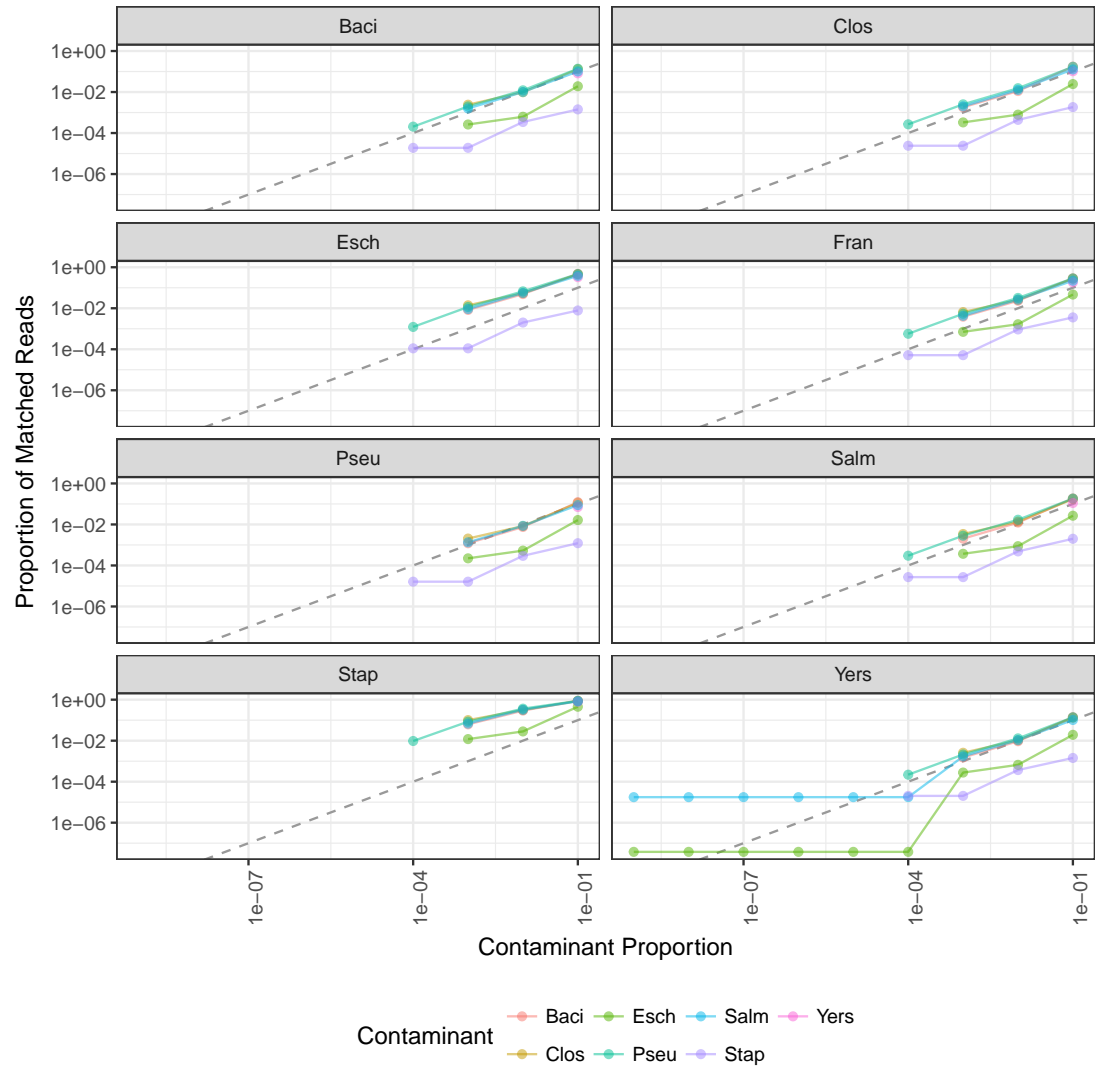
Finally, contaminant detection was assessed using simulated sequencing data from individual genomes. Contaminant datasets were developed by combining subsets of simulated data from two organisms at defined proportions, with the larger proportion representing the microbial material and smaller proportion 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, Species column).

The minimum contaminant proportion detected was  $10^{-3}$  and  $10^{-4}$  for most pairwise comparisons with a few 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. A few contaminants were detected at proportions as low as  $10^{-8}$ , such as when *Yersinia* contaminated with *E. coli* or *S. enterica*. However, contaminants detected at lower proportions were due to reads simulated from the material genome incorrectly assigned to the contaminant. The simulated contaminant-free *Y. pestis* material dataset had reads assigned to two of the contaminants resulting in artificially low contaminant detection proportions for *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18 with an estimated proportion of  $1.76 \times 10^{-5}$  and *Escherichia coli* O104:H4 str. 2011C-3493 with an estimated proportion of  $3.77 \times 10^{-8}$ .

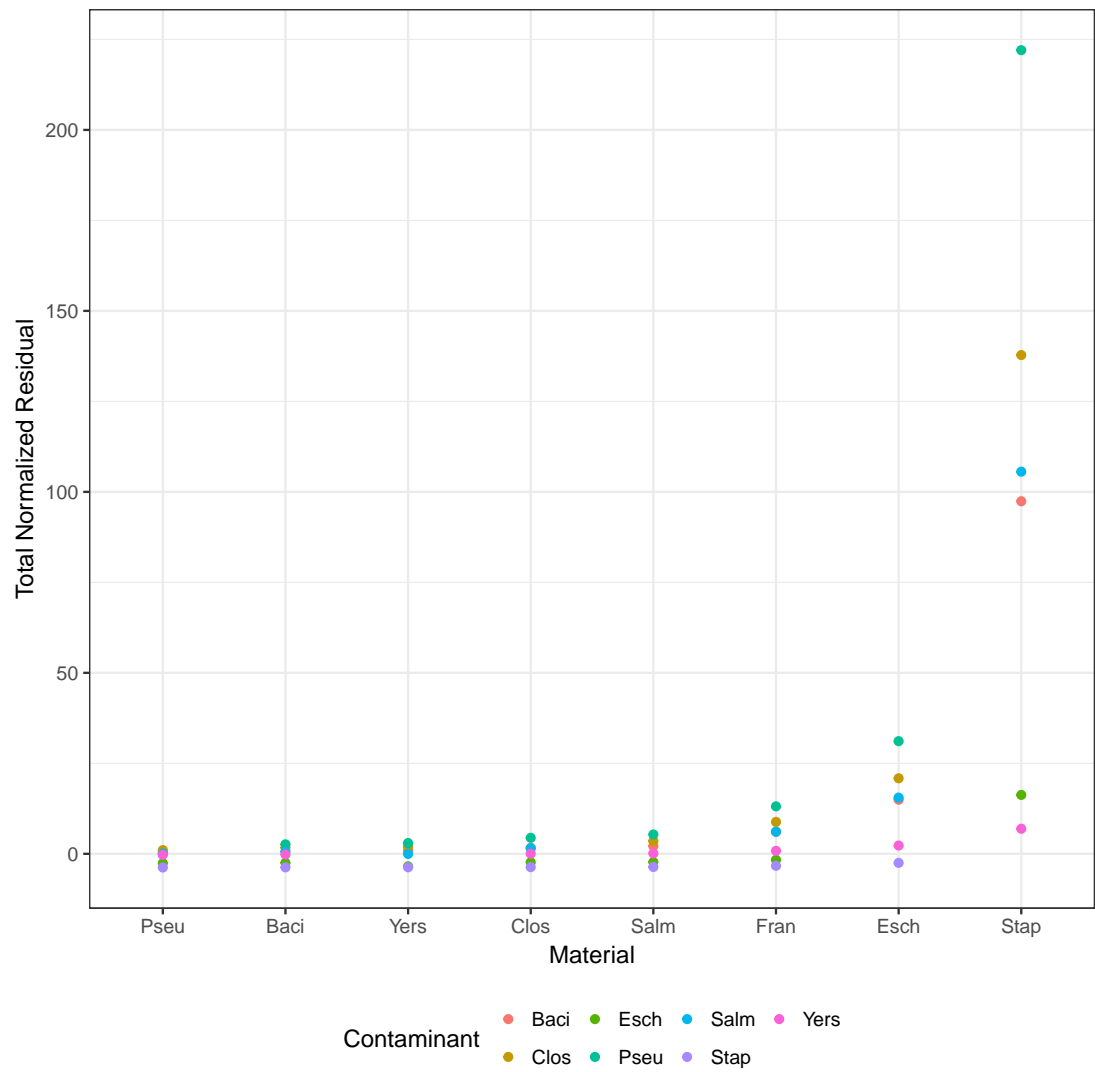
The minimum detected contaminant proportion ranged from  $10^{-3}$  to  $10^{-4}$  for most simulated contaminant datasets. As the individual datasets were simulated at 20X coverage, <300,000 reads were simulated for each dataset, and on average <3 reads were spiked into the material datasets for simulated contaminant proportions  $\leq 10^{-5}$  (Table 2). Unexpectedly low contaminant proportions,  $10^{-8}$ , were detected for *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,  $10^{-1}$  versus  $10^{-3}$ , 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 in the database for these two genera. Additionally, the *F. tularensis* dataset is much smaller relative to the other genera, less than 90,000 reads. Therefore, 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. In addition to the minimum detected contaminant proportion we also evaluated the quantitative accuracy the contaminant detection method.

Pearson's correlation coefficient was used to determine the correlation between the estimated contaminant and true contaminant proportions for simulated contaminant proportions greater than  $10^{-6}$ . The estimated and true proportions were strongly correlated for all pairwise comparisons, with an overall median and 95% confidence interval of 0.99945 (0.96945 - 1) (Fig. 4). Eight of the pairwise comparisons have correlation coefficients below 0.99, all of which have *S. aureus* as either the contaminant





**Figure 4.** The relationship between the proportion of reads matching the contaminant species and the proportion of simulated contaminant reads. Plots are split by the material species with line and point color indicating contaminant species. Dashed line indicate a 1:1 correlation between matched and expected contaminant reads. The contaminant proportion was underestimated for points below the dashed line and overestimated for points above the dashed line.



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

243 or the material. Two coefficients were below 0.98: *S. aureus* contaminated with *P. aeruginosa* and *S.*  
244 *enterica*, 0.952 and 0.969 respectively. Normalized contaminant proportion residuals, were used to as-  
245 sess the accuracy of the PathoScope contaminant proportion estimates (Fig. 5). The material genome  
246 strongly influenced the total normalized residuals with *E. coli* and *S. aureus* having consistently higher  
247 total normalized residuals compared to the other genomes.

## 248 DISCUSSION

249 The potential for using whole genome sequencing data and taxonomic sequence classification algorithms  
250 to detect contaminant DNA in microbial materials was evaluated. The method requires no *a priori* infor-  
251 mation about the contaminant and can identify organisms that are common contaminants and unexpected  
252 contaminants. Additionally, as whole genome sequencing can be performed on genomic DNA and cell  
253 cultures (after DNA extraction), the method is appropriate for both types of microbial materials. A  
254 baseline assessment of the contaminant detection method using simulated sequencing data from indi-  
255 vidual genomes was performed to identify common types of classification errors that would result in  
256 false positive contaminants. The false positive contaminants were split into two categories (1) those due  
257 to an inability of the method to differentiate the material genome from the contaminant genome, and  
258 (2) errors in the reference database. Contaminant detection performance was characterized for different  
259 material, contaminants, and contamination level. Overall the method was able to identify contaminant  
260 proportions at  $10^{-3}$  for most pairwise contaminant-material combinations. A contaminant proportion of  
261  $10^{-3}$  is equivalent to 1 contaminant cell per 1,000 cells in a microbial material, or 1,000 contaminant  
262 cells/mL in a  $10^6$  cells/mL culture. The estimated contaminant proportion accuracy for the simulated  
263 contaminated material varied by contaminant and material strain.

264 There are three basic steps to using this method to detect contaminants in a microbial material. Base-  
265 line assessment is the first step in the contaminant detection method. For a baseline assessment, reads  
266 are simulated from the reference genome of the organism of interest and process the simulated reads  
267 using a taxonomic classification algorithm. Performing a baseline assessment allows one to identify the  
268 false positive contaminants you can expect to observe due to limitations in the method. Simulating data  
269 with similar characteristics error profiles, read length, and fragment distribution is likely to yield results  
270 more representative of what one would expect from real sequencing data. Next process sequencing data  
271 generated from the microbial material using the same taxonomic classification algorithm as was used  
272 in the baseline assessment. The last step is a critical evaluation of results, potential for false positives.  
273 For all settings, research, clinical, regulatory, and attribution the contaminant detection method should  
274 be validated for the intended application. Appropriate validation methods may include experiments with  
275 simulated contaminants like those performed as part of this study and potentially sequencing of genomic  
276 DNA or cells spiked with varying concentrations of known contaminants.

277 It is important to evaluate the results in the context of the intended application. Quantitative accu-  
278 racy in contaminant proportions is important for applications where acceptable contaminant proportion  
279 thresholds are established. For example, a microbial material with a contaminant proportion of  $10^{-5}$   
280 may be acceptable for use in an assay where the contaminant adversely impacts an assay when present  
281 in proportions greater than  $10^{-4}$ . Quantitative accuracy is also relevant when performing a general char-  
282 acterization of the microbial material. General contaminant characterization is appropriate for reference  
283 materials with more than one use case such as the NIST microbial genomic reference materials (NIST  
284 RM8375)(Olson et al., 2016). Similar to the false positive contaminant baseline assessment, simulated  
285 data can be used to evaluate the minimal detectable contaminant proportion for specific organisms using  
286 different taxonomic assignment algorithms and databases.

287 A primary limitation of the proposed method is the observed false positive contaminants identified in  
288 the baseline assessment. The reference database and taxonomic assignment algorithm used are likely to  
289 impact the number and types of false positives. There are a number of different types of taxonomic read  
290 classification algorithms sequence similarity search, sequence composition methods, and phylogenetic  
291 methods (Bazin et al., 2012). The algorithm used in this study, PathoScope, is a type of  
292 sequence similarity search algorithm. Evaluating different types of algorithms using simulated data for  
293 the material genome of interest, similar to what was done in the baseline assessment part of this study,  
294 would help determine the optimal classification algorithm for a specific microbial material. Furthermore,  
295 recent advances in taxonomic classification algorithms have led to the development of faster methods,  
296 including Kaiju a sequence composition type method and Centrifuge a sequence similarity search type

297 method (Menzel et al., 2016; Kim et al., 2016). Application of these new methods would lower the  
298 computational cost of the proposed contaminant detection method.

299 A number of the observed false positive were due to errors in the database and inability of the taxo-  
300 nomic classification algorithm to correctly identify the source of the sequence when it matches multiple  
301 organisms in the database. Removing sequences from the database for irrelevant contaminants, such as  
302 phage, plasmids, vectors, and multicellular eukaryotes would reduce the proportion of false positives.  
303 By excluding irrelevant contaminants from the database sequencing reads aligning to these irrelevant  
304 sequences would no longer result in false positive contaminants. Methods for excluding sequence data  
305 from a reference database is dependent on the classification algorithm used. For example, user-specified  
306 sequence data could be removed from the reference database by PathoScope using the PathoDB function.  
307 Users should be cautious when removing sequences from a reference database when analyzing real data.  
308 For example, vector sequences from contaminants in sequencing reagents may be incorrectly classified  
309 as an organism contaminant. Similarly, using a curated database free of misclassified and unclassified se-  
310 quence data would further help to reduce the proportion of false positive contaminants (Tennessen et al.,  
311 2015). The *Bacillus subtilis*-*Synechocystis* chimeric genome appeared to have a high false positive con-  
312 taminant rate in the baseline assessment part of this study due to the genome being incorrectly classified  
313 as *Bacillus subtilis* and not a chimeric genome.

## 314 CONCLUSIONS

315 Identification and characterization of low abundance contaminants in a non-targeted manner is critical  
316 for a material used in high sensitivity assays such as PCR. With the continual decline in the cost of  
317 sequencing and the advances in sequence analysis methods, whole genome sequencing combined with  
318 taxonomic assignment algorithms provides a viable alternative to commonly used organismal contami-  
319 nant detection methods such as culturing, microscopy, and PCR. The method presented here is suitable  
320 for detecting organismal contaminants in both genomic DNA and whole cell microbial materials, with  
321 the only *a priori* assumptions about the contaminants is that they are present in the reference database.  
322 Furthermore, the method was shown to detect most contaminants when present at  $10^{-3}$  proportion of  
323 cells in a high-throughput manner. The false positive contaminant detection is a primary limitation of  
324 the proposed method. As false positive contaminants are database and taxonomic assignment algorithm  
325 dependent, additional work is needed to improve database curation and data authentication efforts as well  
326 as characterization of taxonomic assignment algorithm performance. With the rapid decrease in sequenc-  
327 ing cost and ability to detect unknown contaminants at low concentrations, whole genome sequencing is  
328 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.
- Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, 26(12):1721–1729.
- 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.

381 Olson, N. D., Zook, J. M., Samarov, D. V., Jackson, S. A., and Salit, M. L. (2016). Pepr: pipelines for  
 382 evaluating prokaryotic references. *Analytical and bioanalytical chemistry*, 408(11):2975–2983.  
 383 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Checkm:  
 384 assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.  
 385 *Genome research*, 25(7):1043–1055.  
 386 R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for  
 387 Statistical Computing, Vienna, Austria.  
 388 Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets.  
 389 *Bioinformatics*, 27(6):863–864.  
 390 Scott Chamberlain and Eduard Szocs (2013). taxize - taxonomic search and retrieval in r.  
 391 *F1000Research*.  
 392 Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When Is a Microbial Culture Pure  
 393 ? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing.  
 394 Tennessen, K., Andersen, E., Clingenpeel, S., Rinke, C., Lundberg, D. S., Han, J., Dangel, J. L., Ivanova,  
 395 N., Woyke, T., Kyrpides, N., et al. (2015). Prodege: a computational protocol for fully automated  
 396 decontamination of genomes. *The ISME journal*.  
 397 Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis.  
 398 *Microbial informatics and experimentation*, 2(1):3.  
 399 Watanabe, S., Shiwa, Y., Itaya, M., and Yoshikawa, H. (2012). Complete sequence of the first chimera  
 400 genome constructed by cloning the whole genome of synechocystis strain pcc6803 into the bacillus  
 401 subtilis 168 genome. *Journal of bacteriology*, 194(24):7007–7007.  
 402 White, J. M. (2014). *ProjectTemplate: Automates the creation of new statistical analysis projects*. R  
 403 package version 0.6.  
 404 Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R.,  
 405 and Cheng, J.-F. (2011). Decontamination of mda reagents for single cell whole genome amplification.  
 406 *PloS one*, 6(10):e26161.