Using metagenomic methods to detect organismal contaminants in microbial materials.

Nathan D. Olson

Justin Zook

Jayne Morrow

Nancy Lin

# Introduction

Shotgun metagenomic sequencing is used to characterize environmental samples and detect pathogens in complex samples. The same method can also be used to detect contaminants in microbial materials, cell cultures and genomic DNA from clinical or environmental isolates. Microbial materials free of contaminants are needed for biodetection assay validation and basic research using model systems . 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 contaminanting. 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. 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. For genomic DNA materials, the material itself can be sequenced, whereas the 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. All methods require a reference database for classification. In order to detect a contaminant within a microbial material, the contaminating 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 detecting contaminants in microbial materials.

In this work, we present the results for a proof of concept in-silico method to evaluate the suitability of whole genome sequencing data combined with a taxonomic assignment algorithm for detecting organismal contaminants in microbial materials. 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 organismal contaminants in microbial materials. Simulated data from single genomes was used to characterize the rate at which the method correctly classifies reads as the test material. To characterize the method’s ability to detect contaminants, simulated contaminant datasets comprised of pairwise combinations of single genomes spiked with a defined proportions of contaminant reads, reads simulated from a different genome.

Error Models

To best approximate real sequencing data reads were simulated using an empirical error model and insert size distribution. Whole genome sequencing data was simulated using the ART sequencing read simulator . 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 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 .

Taxonomic Composition Assessment

The taxonomic composition of simulated datasest was assessed using the Pathoscope sequence taxonomic classifier . 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 , (2) PathoMap - mapping reads to a reference database using the bowtie2 algorithm , (3) PathoID - expectation-maximization classification algorithm. The annotated Genbank nt database provided by the PathoScope developers was used as the reference database (<ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz>).

## Single Genome - Baseline Assessment

Simulated sequencing data from individual genomes was used to characterize the false positive contaminants reported by Pathoscope. Sequence data was simulated for 406 strains, from 9 genera (Table [tab:single\_org]). We will refer to the genome used to generate the reads as the target genome. The genomes included in the simulation study were limited to the number of closed genomes in the Genbank database (<http://www.ncbi.nlm.nih.gov/genbank/>, accessed 10/18/2013) belonging to the genera of interest (Table [tab:single\_org]). 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 target genome and simulated read assignment match levels were determined using the R package, Taxize .

## Simulated Contaminants

Simulated contaminated datasets were used to evaluate how contaminant dection varied by material and contaminant strain over a range of contaminant concentrations. Representative genomes for 8 of the 9 genus were used to generate the simulated contaminant datasets (Table [tab:contam\_table]). An *Escherichia coli* strain was selected as a representative of both and *Shigella* as the genus *Shigella* phylogenetically resides within the species *Eschericha coli* . For each pairwise combination of representative genomes the simulated contaminant dataset was comprised of a randomly selected subset of reads from the target and contaminant simulated single genome sequence dataset. The simulated datasets were randomly subsampled at defined proportions, with representing the proportion of reads from the contaminant single genome dataset, and representing the proportion of reads from the target genome dataset. *Make Sure to Revise for Clarity - Maybe include a figure/diagram.* A range of contaminant proportions at 10-fold increments were simulated with ranging from to , 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 have 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-pathoscope/>). The script used to run the simulations are available at <https://github.com/nate-d-olson/genomic_purity>. Additionally, seed numbers for the random number generator were randomly assigned and recorded for each dataset so that the same simulated datasets could be regenerated. Pathoscope results were processed using the statistical programing language R , and intermediate analysis and data summaries were organized using ProjectTemplate and archived in a github repository (<https://github.com/nate-d-olson/genomic_purity_analysis>) along with the source file for this manuscript.

# Results

## Single Genome - Baseline Assessment

We first assessed baseline performance of the proposed method for detecting organismal contaminants in microbial materials. Our analysis included taxonomic classification results for simulated sequencing data from 388 genomes, representing 9 different genera (Table [tab:single\_org]). For 105 out of 388 genomes, Pathoscope estimated that 99% of the material was the same species as the genome the sequencing data was simulated from (Fig. [fig:species\_prop]). The estimated proportion of the material identified as the correct species varied by genus. All of the pure *Shigella* genomes and 44 of the 49 pure *Staphylococcus* genomes were inaccurately estimated to compose less than 90% of the sample. 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 of *Shigella*, *Escherichia*, and *Staphylococcus* were due to reads being incorrectly identified as belonging to organisms from other species.

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

|  |  |  |
| --- | --- | --- |
| 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) |

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

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

We characterized the false positive contaminants responsible for the genera *Shigella*, *Escherichia*, and *Staphylococcus*, as well as genomes of other genera with species match proportions less than 0.9. The false positive contmainants were grouped into three types, taxonomic ambiguities, phage, and method artifacts. Taxonomic ambiguities were defined as contaminants with highly similar genome sequences to the material genome sequence but taxonomically classified as different species. For example the low match percentage for *Clostridium autoethanogenum* strain DSM10061 was due to *Clostridium ljungdahlii* strain DSM13528 assigned the top proportion (**VALUE**) instead of *C. autoenthanogenum*. Similarly, *Escherichia coli* strain UMNK88 low match, 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 **VALUE** and **VALUE** respectively. Taxonomic ambiguities can be due to a species being incorrectly assigned to the wrong taxonomic group for example the *Bacillus* genome with species match proportion close to zero, *Bacillus infantis* string NRRL B 14911. While the *B. infantis* strain was originally classified as *Bacillus*, the species is phylogenetically distinct from other members of the genus . Taxonomic ambiguities are 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. Though considerably higher, this match proportion is still low relative to the other genera **INCLUDE VALUE DIFFERENCE**.

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

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

Phage, the second type of false positive contaminant, were reported by Pathoscope as present in varing proportions for genomes from all 9 genera investigated (Fig. [fig:phage\_prop]). 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 (Fig. [fig:phage\_prop]). All of the false postive phage contmaninants were specific to the taxonomy of the genome the sequence data was simulated from. The phage contaminants may represent errors in the database, where sequence data from the host organisms genome is missassembled into the phage genome, or genomic sequence is shared between the phage and the host, such as CRISPR, and lysogenic phage (REF).

**Method Artifacts** The third type of false positive contaminant, are method artifacts. Method artifact false positive contaminants are defined as false positives due to missclassified or unclassified sequences in the database, genome assemblies in the database including sequence data from organismal or reagent contaminants, and errors associated with the classification algorithm. *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. The high match proportion is not due to taxonomic ambuguities but potentially 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 species, sp. HF-1. The eukaryotic false positive contaminants are likely either due to similarities between the genome sequences or errors in the assembly **REF**. Plasmids and vectors are common used molecular biology tools and well known contmainants of sequencing data (REF). The plasmids and vectors identified as false positive contaminants due to errors in the genome assemblies where artifacts of the sequencing process were not properly removed from the sequencing data prior to assembly. Alternatively, the misclassification could be due to high similarity between the genome sequence the reads were simulated from and the plasmid and vector sequence which is not unexpected as most plasmid and vectors have microbial origins (REF). The low species proportion of species level matches for *Pseudomonas* strain TKP was likely due to both missclassified sequences (*Thioalkalivibrio sulfidophilus* strain HL-EbGr7 match proportion 0.0648), and contminanted genome sequences in the database (wheat - *Triticum aestivum* match proportion 0.087). Errors associated with the classification algorithm are likely resposible for low species level match path proportion for *Pseudomonas* strain VLB120 is likely due to errors in the database as of the matches assigned to the unrelated species *Xanthomonas* sp. W17. The representative database sequence is  9 kb (https://www.ncbi.nlm.nih.gov/nuccore/710572), the Pathoscope expectation maximization step include a length normalization step which is likely responsible for the classification error.

## Simulated Contaminants - 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. 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 [tab:contam\_table]). 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 [tab:contam\_table]).

The minimum contaminant proportion at which any reads were assigned to the contaminant was and for most pairwise comparisons with a few notable exceptions. When *Yersina* was the simulated contaminant the minimum detected proportion was 0.1 for all material strains (Table [tab:contam\_min\_table]). Conversely, contaminants were detected at lower proportions, , when *Yersinia* was contaminated with *E. coli* as well as when *S. enterica* and *E. coli* contaminated with *B. anthracis*.

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

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

Pearson correlation coefficients for estimated and true contaminant proportions.

Pearson correlation coefficients for estimated and true contaminant proportions.

Total normalized residuals for pairwise combinations of material and contaminants.

Total normalized residuals for pairwise combinations of material and contaminants.

The Pearson’s correlation coefficient was used to measure the correlation between the estimated contaminant and true contaminant proportions. The estimated and true proportions were strongly correlated for all pairwise comparisons, with an overall median and 95% confidence interval of 0.99945 (0.96943 - 0.99999 (Fig. [fig:contam\_fig]). 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. Contaminant proportion residuals were used to assess the accuracy of the Pathoscope contaminant proportion estimates (Fig. [fig:contam\_resid]). The material genome strongly influenced the total normalized residuals with *E. coli* and *S. aureus* having consistently higher total normalized residuals compared to the other genomes.

# Discussion

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

The primary limitation of the proposed method is the observed false positive contamiants for single genome simulated sequencing data. Performing a baseline assessment of false positive contaminant using simulated sequence data from the microbial material’s genome sequence, and choosing the appropriate database and taxonomic assignment algorithm can help reduce the impact of false positive on the method’s ability to identify organismal contmainants. 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. Differentiating between two organisms with highly similar sequences will vary by taxonomic classification algorithm. Pathoscope was used for this proof of concept study as the method uses the full reads and paired end information for taxonomic classification rather then shorter sequence fragment, -mers. Our assumption is that the longer sequence allows for better discrimination between highly simillar 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.

The ability to identify contaminants present at low concentrations is critical for when the material will be 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 . Similar to the false positive contaminant baseline assessment. This study shows, simulated 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 constant decline in the cost of sequencing, advances in sequence analysis methods, whole genome sequencing combined with a taxonomic assignment algorithm provide a viable alternative to commonly used organismal contaminant dection methods such as culturing, microscopy, and PCR. The method is suitable for detecting organismal contaminants in both genomic DNA and whole cell microbial materials without requiring any *a priori* assumptions about the contaminant. Furthermore, the method was shown to detect contaminants making up proportion of cells in a high-throughput manner. Finally, the sequence data generated for organismal contaminant detection can 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.