Method for evaluating genomic material purity using whole genome sequencing data.

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

- basic introduction - more detailed background - general problem - main result summary - explanation of what main result reveals - general context - broader perspective

Keywords: Biodetection, Test material, Reference material, Purity, Bioinformatics

INTRODUCTION

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

Shotgun metagenomics consist of two main steps, whole genome sequencing on genomic DNA, and analyzing the resulting sequencing data most commonly using a taxonommic assignment algorithm (Thomas et al., 2012). For genomic DNA material, 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 taxonomically characterize the sequencing data. There are a number of different types of classification algorithms with varying classification accuracy and computational performance (Bazinet and Cummings, 2012; Menzel and Krogh, 2016). All methods require a reference database for classification. Inorder for a contaminant to be detectable within a microbial material, the contaminant or an organism more closely related to the contaminant than the material must be present in the database. As taxonomic classification algorithms are constantly improving, reference databases are expanding, and the cost of sequencing drops, shotgun metagenomic sequencing provides an alternative method for detecting contaminants in microbial materials over current methods.

In this work, we present the results of a proof of concept study evaluating the suitability of whole genome sequencing data combined with a metagenomic read classification algorithm for detecting organismal contaminants in microbial materials. We used *Pathoscope*, a taxonomic classification algorithm orginally developed for strain level pathogen detection. We will first provide a baseline assessment of the method using simulated sequencing data for single organisms to characterize the types of false positive contaminants the method may report. Then, we evaluate the methods ability to detect organismal contaminants in microbial material strains using sequence data simulated to replicate microbial materials with different organismal contaminant strains and 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 validating test material purity. 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 ability of the method to detect contaminants, simulated contaminant datasets comprised of pairwise combinations of single genomes spiked with a defined proportion of contaminant reads, reads simulated from a different genome.

To best approximate real sequencing data reads were simulated using an empirically determined error model and insert size distributions. The whole genome sequencing data was simulated using the ART sequencing read simulator (Huang et al., 2012). Reads were simulated with ART simulator using the Illumina MiSeq error model for 2×230 base pair (bp) paired end reads with an insert size of 690 ± 10 bp (average $\pm$ standard deviation) and 20 X mean coverage. The insert size parameters were defined based on the observed average and standard deviation insert size of the NIST RM8375-MG002 MiSeq sequencing data (Olson et al., 2016).

The taxonomic composition of simulated datasest was assessed using the Pathoscope metagenomic taxonomic classifier (Francis et al., 2013). This method was selected as it combines the use of a large reference database reducing potential biases due to contaminant sequences not present in the database and efficient whole genome read mapping algorithms. This method uses an expectation maximization algorithm where the sequence data are first mapped to a database comprised of all sequence data in the Genbank nt database. Then, through an iterative process, it re-assigns ambiguously mapped reads based on the proportion of reads mapped unambigously to individual taxa in the database. The PathoScope 2.0 taxonomic read classification pipeline has three steps; (1) PathoQC - read quality filtering and trimming using the PRINSEQ algorithm (Schmieder and Edwards, 2011), (2) PathoMap - mapping reads to a reference database using the bowtie2 algorithm (Langmead and Salzberg, 2012), (3) PathoID - expectation-maximization classification algorithm. The annotated Genbank nt database provided by the PathoScope developers was used as the reference database (`ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz`).

Single Genome - Baseline Assessment

Method specificity was first assessed to characterize the baseline accuracy of the read classifier. Method specificity was defined as the proportion of reads in a single organism simulated dataset incorrectly assigned to a taxonomy different from the test material taxonomy. Sequence data was simulated for 406 strains, from 9 genera (Table 1). We will refer to the genome used to generate the reads as the target genome. The genomes included in the simulation study were limited to the number of closed genomes in the Genbank database (`http://www.ncbi.nlm.nih.gov/genbank/`, accessed 10/18/2013) belonging to the genera of interest (Table 1). Due to the large number of closed genomes from the genera *Bacillus*, *Escherichia*, and *Salmonella*, these genera were limited to the species *Bacillus cereus*, *Escherichia coli*, and *Salmonella enterica*. The taxononomic heirarchy for the target genome and simulated read assignment match levels were determined using the R package (Scott Chamberlain and Eduard Szocs, 2013; Chamberlain et al., 2016).

Simulated Contaminants

Method sensitivity was assessed using simulated contaminated datasets to evaluate at how well the method is able to detect genomic contaminants at a range of contaminant concentrations. Representative genomes for 8 of the 9 genus were used to generate the simulated contaminant datasets (Table 2). An *Escherichia coli* strain was selected as a representative of both and *Shigella* as the genus *Shigella* phylogenetically resides within the species *Eschericha coli* (Lan and Reeves, 2002). For each pairwise combination of representative genomes the simulated contaminant dataset was comprised of a randomly selected subset of reads from the target and contaminant simulated single genome sequence dataset. The simulated datasets were subsampled at defined proportions with p representing the proportion of reads from the contaminant single genome dataset subsampled and $1 - p$ the proportion of reads from the target genome simulated dataset. *Make Sure to Revise for Clarity - Maybe include a figure/diagram.* A 10 fold range of contaminant proportions were simulated with p ranging from 0.1 to 10^{-8}, resulting in 512 simulated contaminant datasets. This approach simulates the proportions of cells in a test material and not the amount of DNA, assuming unbiased DNA extraction.

To generate the simulated contaminant datasets single organism simulated datasets were first generated for the 8 representative genomes using the same methods as used in baseline assessment. The resulting simulated sequencing data was first processed using the PathoQC and PathoMap steps in the PathoScope pipeline. The output from the PathoMap step (sam file, sequence alignment file `https://samtools.github.io/hts-specs/SAMv1.pdf`) for the target and contaminant datasets were subsampled as described above 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 rather then for every simulated contaminant dataset.

Bioinformatic Pipeline

To facilitate repeatability and transparency, a Docker (`www.docker.com`) container is available with installed pipeline dependencies (`www.registry.hub.docker.com/u/natedolson/docker-pathoscop`
The script used to run the simulations are available at `https://github.com/nate-d-olson/genomic_purity`. Additionally, seeds number for the random number generator was randomly assigned and recorded for each dataset so that the same simulated datasets could be regenerated. Pathoscope results were processed using the statistical programing language R (R Core Team, 2016), and intermediate analysis and data summaries were organized using ProjectTemplate (White, 2014) and archived in a github repository (`https://github.com/nate-d-olson/genomic_purity_analysis`) along with the source file for this manuscript.

RESULTS

Single Genome - Baselines Assessment

We first assessed baseline performance of the method proposed method for characterizing organismal contaminants of microbial materials. Our analysis included taxonomic classification results for sequencing data simulated from 406 genomes, representing 9 different genera (Table 1). For 105 out of 388 genomes, Pathoscope estimated that 99% of the material was the same species as the genome the sequencing data was simulated from (Fig. 1). The estimated proportion of the material identified as the correct species varies by genus, with none of the *Shigella* genomes having estimated proportions greater than 99% and five of the 49 *Staphylococcus* genomes having proportions greater than 99%. *Shigella* and *Staphlyococcus* along with *Escherichia* represent 87 of the 105 genomes with less than 99% estimated match proportions at the species level. Excluding *Shigella*, *Escherichia*, and *Staphylococcus* the median estimated proportion matching at the species level or higher is 0.9995037. The low species level match proportions were due to false positive contaminants as the input sequencing data were simulated from individual genome sequences.

Genus	N	Genome Size (Mb)
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

We characterized the false postive contaminants responsible for the observed low match proportions for the *Shigella*, *Escherichia*, and *Staphylococcus* genus, as well as genomes of other genera with species match proportions less than 90%. The false positive contmainants were split into three types, taxonomic

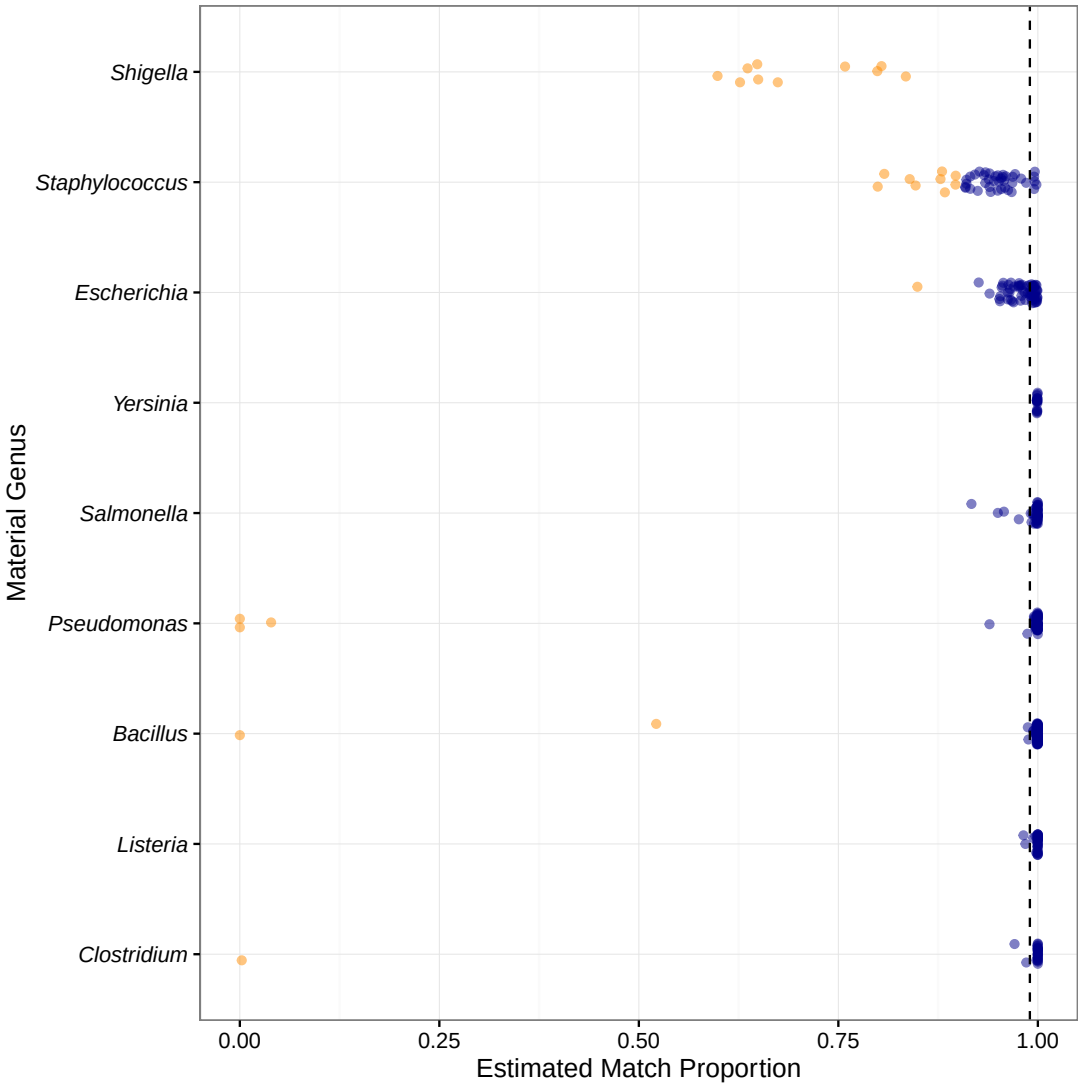

Figure 1. Species level estimated match proportion varies by material genus. The proportion of the material, simulated sequence data from individual genomes, was estimated by Pathoscope. The estimated match proportion is the total proportion of the material with taxonomic assigment to the genome species, subspecies, strain, or isolate levels. The vertical dashed line indicates the 99% match proportion. Orange points are genomes with species level match proportions less than 90% and blue points greater than 90%

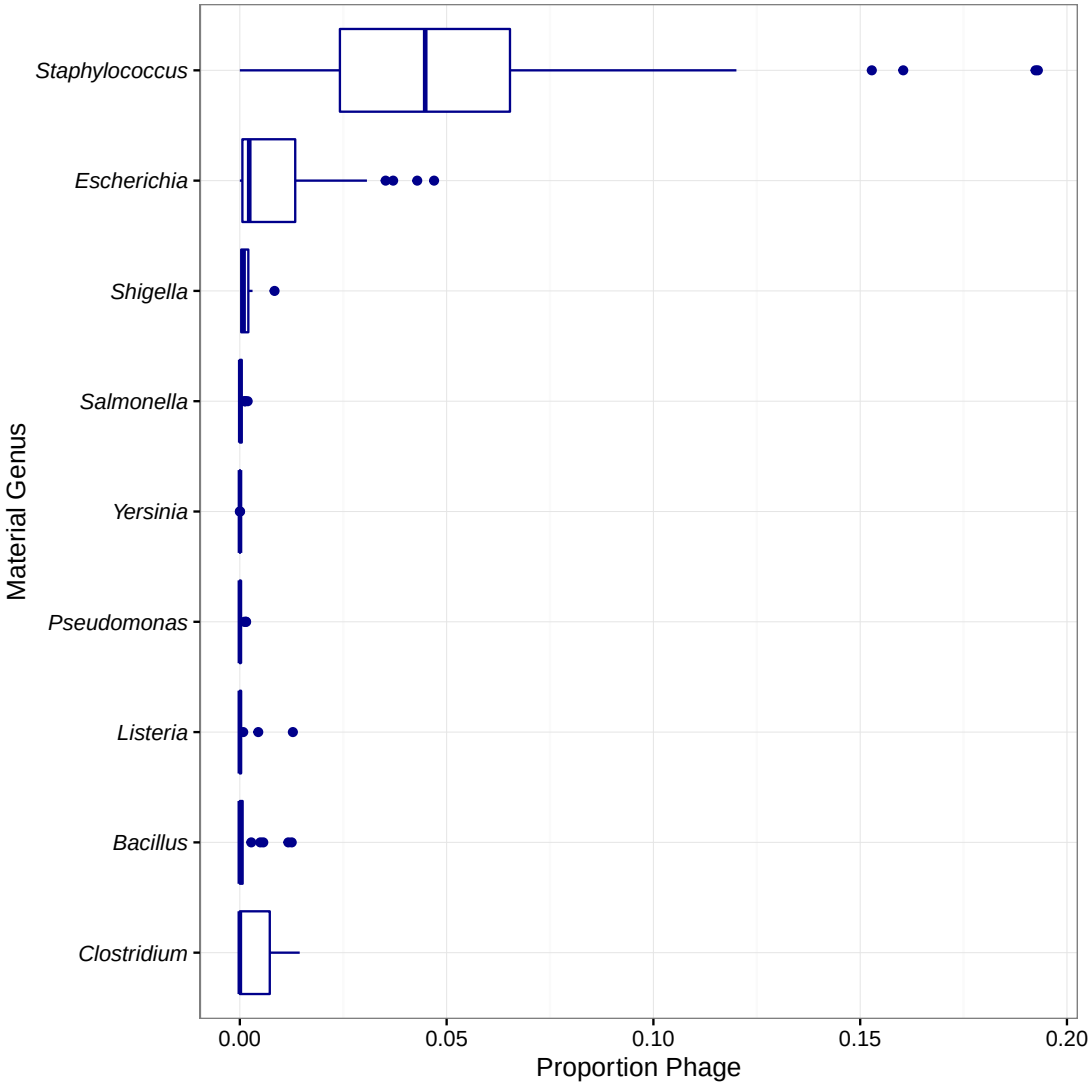

Figure 2. Estimated proportion of phage in the simulated single genome datasets by genera.

ambiguities, phage, and **OTHERS**. Taxonomic ambiguities were defined as contaminants with highly similar genome sequences but taxonomic classified as different species. For example the low match percentage for *Clostridium autoethanogenum* strain DSM10061 was due to *Clostridium ljungdahlii* strin DSM13528 had the top proportion 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 0.1 and 0.02 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 (Ko et al., 2006). Taxonomic ambiguities are at least partially responsible for the low species level match proportions for *Shigella* and *Escherichia*. When including matches to E. coli as species level matches, the median match proprtions increases from 0.918609 to 0.6618406. Though considerably higher, this match proportion is still low relative to the other genera.

Phage, the second type of false positive contaminant, were reported by Pathoscope as present at varing proportions for genomes from all 9 genera (Fig. 2). Most noteably, low proportions of species level matches for *E. coli* and *Staphylococcus* can partially be attributed to relatively high proportions of

matches to phage, compared to the other genera investigated . All of the phage false postive 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 where sequence data is shared between the phage and the host, such as CRISPR, and lysogenic phage (REF).

Method Artifacts

Simulated Contaminants - Detection Assessment

Representative Strain	Species	C Mb	C Acc	P Mb	P Acc
Bacillus anthracis str. Ames	1.00	5.23	AE016879.1		
Clostridium botulinum A str. Hall	1.00	3.76	CP000727.1		
Escherichia coli O157:H7 str. EC4115	0.98	5.57	CP001164.1	0.13	CP001163.1, CP001165.1
Francisella tularensis subsp. tularensis SCHU S4	1.00	1.89	AJ749949.2		
Pseudomonas aeruginosa PAO1	1.00	6.26	AE004091.2		
Salmonella enterica subsp. enterica serovar Typhimurium str. D23580	1.00	4.88	FN424405.1		
Staphylococcus aureus subsp. aureus ED133	0.98	2.83	CP001996.1		
Yersinia pestis CO92	1.00	4.65	AL590842.1	0.18	AL109969.1, AL117189.1, AL117211.1

Table 2. Representative strains used in simulated contaminant datasets. Species indicates the proportion of simulated reads assigned to the correct taxa at the species level or higher. DNA size (Mb) and Genbank accession numbers (Acc) are indicated for chromosomes (C) and plasmids (P). *Escherichia coli* O157:H7 str. EC4115 and *Yersinia pestis* CO92 have two and three plasmids respectively.

Next we evaluated how well contaminants are detected. Again using simulated sequencing data from individual genomes we generated contaminant datasets by mixing subsets of datasets from two organisms at defined proportions, with the larger proportion representing the microbial material and smaller proportion the contaminant. 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 proportion of simulated reads that matched at species level or higher was 0.98 (Table 2).

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

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

The minimum proportion of contaminant detected was 10×10^{-3} and 10×10^{-4} for most pairwise comparisons except for when *Yersina* was the simulated contaminant (0.1 for all material strains) (Table **??**). Contaminants were also detected at lower proportions, 10×10^{-8}, when *Yersinia* was contaminated with *E. coli* as well as when *S. enterica* and *E. coli* contaminated with *B. anthracis*.

The quantitative accuracy of the method, linear regression - R^2?

DISCUSSION

- taxonomic ambiguites, need to perform a baseline assessment for your material -contaminants from DNA extraction The plasmids and vectors **Need to make clear from molecular biology**, false positives are likely either 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 eukaryotic false positive contaminants are likely either due to similarities between the genome sequences or errors in the assembly **reference eukaryote microbial genome assembly contaminants**. Validation of material purity prior to performing whole genome sequencing for assembly may

help to prevent this type of assembly errors. - Types of false positive contaminants will be database and classification algorithm specific.

- limitations of the method and false positive rate - how limitations might be addressed - different taxonomic assignment algorithm - database issues - defining the baseline for your material - taxonomic resolution of the method - strain level vs. genus level resolution - Detection limits - how detection limits vary by contaminant and organism - how these may vary by classification methods and sequencing depth - Quantitative nature of the method - is this relevant to the application - How the method can be applied - big picture conclusion

CONCLUSIONS

ACKNOWLEDGMENTS

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

REFERENCES

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

Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., Foster, Z., and O'Donnell, J. (2016). *taxize: Taxonomic information from around the web*. R package version 0.7.4.

Coates, S. G., Brunelle, S. L., and Davenport, M. G. (2011). Development of standard method performance requirements for biological threat agent detection methods. *Journal of AOAC International*, 94(4):1328–37.

Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. a., and Johnson, W. E. (2013). Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome research*.

Heck, K., Machineski, G. S., Alvarenga, D. O., Vaz, M. G. M. V., de Mello Varani, A., and Fiore, M. F. (2016). Evaluating methods for purifying cyanobacterial cultures by qpcr and high-throughput illumina sequencing. *Journal of Microbiological Methods*.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4.

Ieven, M., Finch, R., and van Belkum, a. (2013). European quality clearance of new microbiological diagnostics. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(1):29–38.

Ko, K. S., Oh, W. S., Lee, M. Y., Lee, J. H., Lee, H., Peck, K. R., Lee, N. Y., and Song, J.-H. (2006). Bacillus infantis sp. nov. and bacillus idriensis sp. nov., isolated from a patient with neonatal sepsis. *International journal of systematic and evolutionary microbiology*, 56(11):2541–2544.

Lan, R. and Reeves, P. R. (2002). Escherichia coli in disguise: molecular origins of shigella. *Microbes and infection*, 4(11):1125–1132.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Marron, A. O., Akam, M., and Walker, G. (2013). A Duplex PCR-Based Assay for Measuring the Amount of Bacterial Contamination in a Nucleic Acid Extract from a Culture of Free-Living Protists. *PloS one*, 8(4):e61732.

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

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

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

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

Scott Chamberlain and Eduard Szocs (2013). taxize - taxonomic search and retrieval in r. *F1000Research*.

Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When Is a Microbial Culture Pure ? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing.

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

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