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] **Biosystems and Biomaterials Division, National Institute of Standards and Technology**

ABSTRACT

Dummy abstract text.

Keywords: Biodetection, Test material, Purity, Bioinformatics

INTRODUCTION

Rapid, sensitive and accurate assays for detecting bacterial pathogens in food, water, clinical samples, and suspicious biothreats is critical to public health and safety (REF). Biodetection assays must be evaluated for assay sensitivity and specificity prior to deployment as well in the hands of the user to instill confidence in the actions made based on the assay results (Ieven et al., 2013; Coates et al., 2011; EPA, 2004; ISO/TS, 2010; Guide, 1998; Feldsine et al., 2002). Test materials are used to validate assay performance. Test materials can be either purified cultures, genomic DNA or whole cells spiked into a matrix (EPA, 2004; ISO/TS, 2010; CLSI, 2010). Before being used to evaluate a biodetection assay the test material itself must be validated in terms of purity and identity to eliminate false positive results due to test material contaminants or false negative due to the test material being the wrong strain (CLSI, 2010). There are a number of potential sources of microbial contaminants including the stock culture, the preservation medium, as well as airborne and laboratory contaminants (Marron et al., 2013; Shrestha et al., 2013; Tanner et al., 1998).

Current methods for evaluating test material purity include polymerase chain reaction (PCR) assays, metagenomics, and whole genome sequencing based approaches. One PCR assay was developed to analyze protist cultures. This PCR assay uses endpoint PCR for prokaryotes and eukaryotes with template dilutions (Marron et al., 2013). The benefit to PCR-based approaches is that they can be cost effective and fast if an applicable protocol exists. However, PCR assays can only target specific contaminants. While PCR assays can detect contaminants, this approach does not scale effectively for multiple contaminants and test materials. The bioinformatics tools developed to identify contaminants in metagenomic datasets, which include sequencing data from all organisms in a sample, can also be used to evaluate test material purity. For example DeconSeq (Schmieder and Edwards, 2011) and a similar method QC-Chain (Zhou et al., 2013) were developed to identify contaminants based on analysis of 16S ribosomal ribonucleic acid (rRNA) gene sequences or comparison of a subset of reads to a reference database using Basic local alignment search tool (BLAST). Metagonomic-based methods are able to identify contaminants without any prior knowledge or assumptions regarding the identity of the organism(s). However, methods based on 16S rRNA gene identification have limited resolution, as 16S rRNA sequences can only provide genus level taxonomic resolution at best. Methods using BLAST-based searches represent a broader scale approach but are limited by the accuracy of the BLAST classification method. The benefit to using metagenomic tools developed is that prior knowledge of the identity of the contaminant is not required; however this method is unable to identify contaminants to the strain level.

Another approach to evaluating test material purity is through shotgun whole genome sequencing, sequence all DNA in a single organism sample. A recently published bioinformatics method, *pathoscope*, was developed to detect pathogens and identify strains using whole genome sequencing data (Francis et al., 2013). This method benefits from the large sample size obtained using next generation sequencing for higher sensitivity and leverages algorithm advances for whole genome sequence mapping. Mapping algorithms determine the optimal placement of reads relative to a reference sequence (Schbath et al., 2012). Reads are either uniquely or ambiguously mapped. For uniquely mapped reads only a single optimal mapping location is identified, whereas for ambiguously mapped reads multiple optimal mapping locations are identified. Pathoscope uses the number of reads that uniquely map to different genomes in the reference database to assign ambiguously mapped reads, reads that align equally well to multiple reference sequences. The primary benefits to shotgun whole genome sequencing and subsequent pathoscope analysis approach are that prior knowledge of the contaminant is not required and it has the potential for higher sensitivity compared to other methods. However, the main limitation to this method is the size of the reference database, namely that the genome of the contaminant or a closely related organism must be present in the database for the contaminant to be detected. In this work, we present the results of a proof of concept study to measure the purity of single organism test materials built upon the pathoscope software for pathogen detection. This method is based on whole genome sequencing and utilizes *pathoscope* with an expanded reference database. We will first present the specificity of the method using simulated data for single organisms. Then, evaluate sensitivity of the method using simulated datasets generated to represent contaminated test material.

METHODS

To test the suitability of using whole genome sequence data and metagenomic taxonomic classification methods to evaluate the genomic purity of a test material we first used simulated whole genome sequence data from single genomes and simulated contaminanted datasets. Simulated data from single genomes was used to assess method specificity and simulated contaminant datasets method sensitivity. Simulated datasets were generated using the ART sequencing read simulator (Huang et al., 2012). The datasets were generated using the Illumina MiSeq error models for 230 paired end base pair reads and 20 X mean coverage with an average insert size of 690 base pairs with standard deviation of 10 bp for each strain, the seed number for the random number generator was randomly assigned and recorded for each dataset.

The taxonomic composition of simulated datasest was assessed using the Pathoscope metagenomic taxonomic classifier (Francis et al., 2013). This method uses an expectation maximization algorithm where the sequence data are first mapped to a database comprised on all sequence data in the Genbank nt database, then through an iterative process re-assigns ambiguously mapped reads to based on the proportion of reads mapped unambigously to individual taxa in the database. The PathoScope 2.0 taxonomic read classification pipeline includes an initial read filtering step (PathoQC), followed by mapping reads to a reference database (PathoMap) - a wrapper for bowtie2 (Langmead and Salzberg, 2012)), then an expectation-maximization classification algorithm (PathoID). 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`).

Specificity

Sequence data was simulated for strains 406, from 9 genus (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). Genomes included in the study were limited to those belonging to the genus *Pseudomonas*, *Listeria*, *Clostridium*, *Yersinia*, *Francisella*, and *Shigella*. Due to the large number of closed genomes the follow species were used instead of genus *Bacillus cereus*, *Escherichia coli*, *Salmonella enteria*. Method specificity was defined as the proportion of reads assigned to a different taxonomy then the taxonomy of the target genome. 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).

Sensitivity

We simulated datasets with genomic contaminants to evaluate method sensitivity. Representative genomes for the 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 *Eschericha coli* and *Shigella*. For each pairwise combination of representative genomes for the simulated simulated contaminant dataset was subsampled at 0.1 to 10^{-8}, representing 10 fold dilutions, the target genome dataset was subsamples at 1 - contaminant proportion, resulting in 512 simulated contaminant datasets. This apporach simulates the proportions of cells in a test material and not the amount of DNA, assuming unbiased DNA extraction. To speed up processing the aligned sequence files were subsampled instead of the simulated sequence files.

Reproducibility

To facility reusability 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 were available at `https://github.com/nate-d-olson/genomic_purity`. Pathoscope results were processed using 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`).

RESULTS AND DISCUSSION

Specificity

Simulated sequence data from individual isolates was used to assess the specificity of the genomic purity assessment method. We defined specificity as ability of the method to assign reads to taxonomy of the genome the sequencing reads were simulated from, the target genome. Method specificity was evaluated by characterizing the read assignment results based on the level of agreement between the genome and assigned taxonomy (Fig. 1). Overall high proportions of matches at species and genus level. The cumulative match proportions do not always reach 1.00, for example *Staphylococcus* genomes. This might be due to exclusion of unclassified and unknown matches from match level analysis or reads that bowtie (mapping algorithm) was unable to align to any sequece in the reference database. Some genus have high levels of family and higher matches. For *Shigella* most likely due to matches with *Escherichia* (NEED TO VARIFY).

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 simultated datasets. N indicates the number of genomes, and Genome Size is presented as the median and range (minimum to maximum) genome size

Most of the genus had genus level or higher match proportions excluding a few outliers (Fig. 2). *Escherichia*, *Shigella*, and *Staphylococcus* are noteable exceptions. As discussed previously the taxonomic ambiguities for *Shigella* and *Escherichia* are responsible for the overall lower genus level match proportions. It is important to consider the strain and genome being characterized as taxonomic ambiguities (e.g. *Shigella* and *Escherichia*) can lead to lower than expected specificity and the identification of false positive contaminants.

Sensitivity

To evaluate the genomic purity assessment method we generated simulated contaminant datasets as pairwise combinations of representative genomes from 8 of the genus used in the specificity section of the study (did not include Shigella in this component of the study). Due to the overall high proportion of reads matched to the correct genome in the method specificity study the simulated contaminant datasets

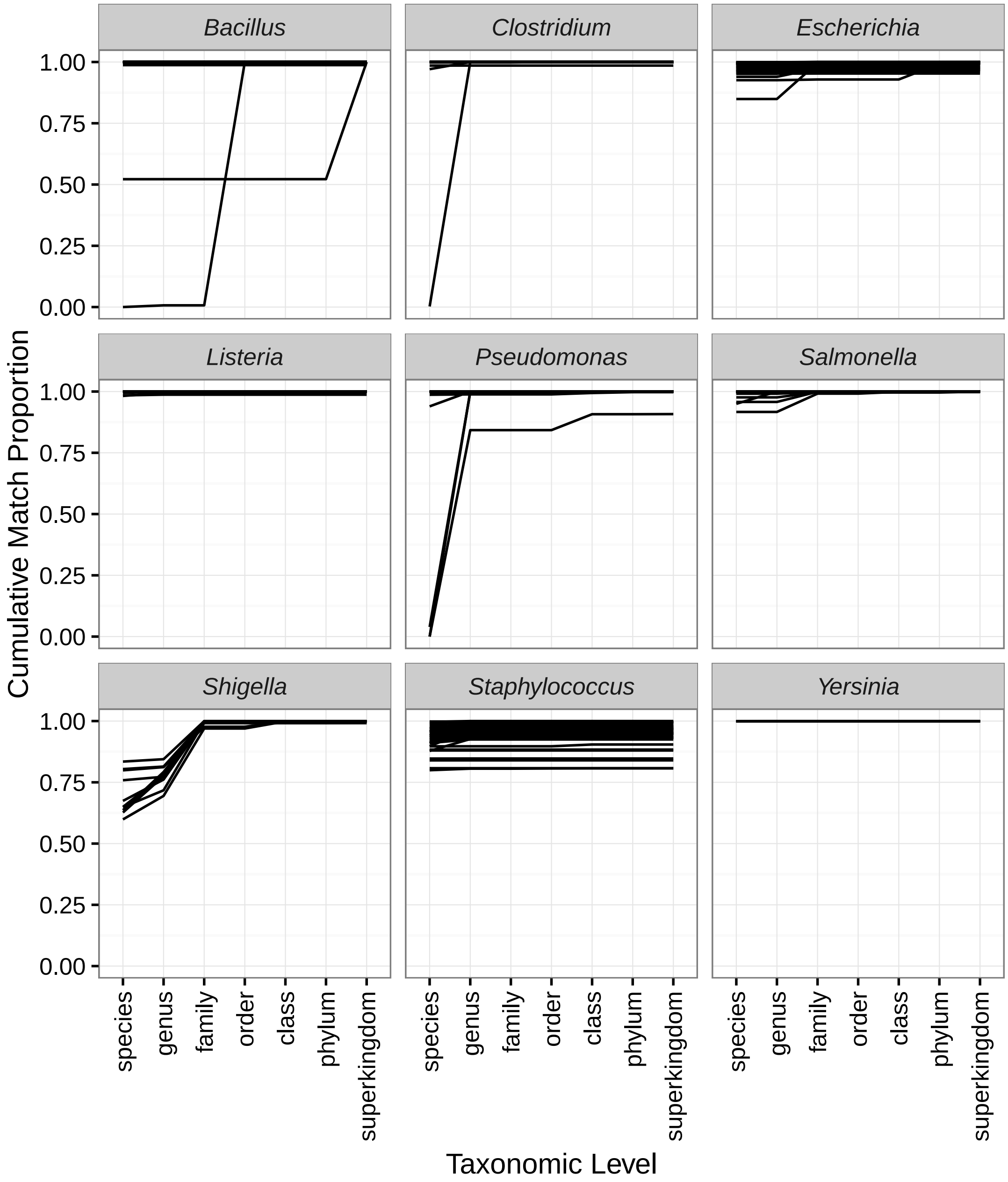

Figure 1. Cumulative taxonomic match results for genomic purity assements of simulated sequence data from single genomes. Each line represents the cumulative proportion of simulated reads with taxonomic assignments matching at or above the specified taxonomic level. Genomes are grouped by genus.

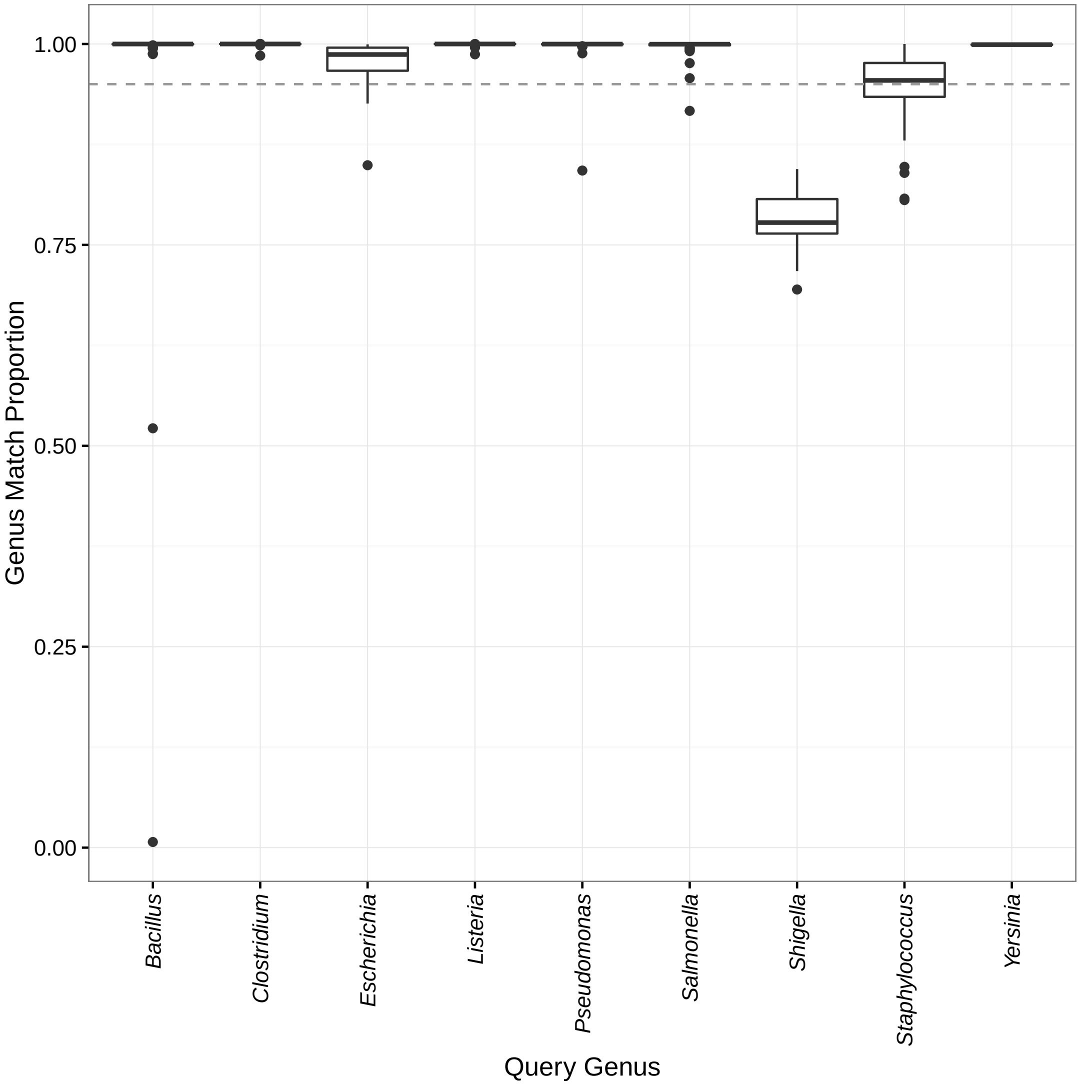

Figure 2. Distribution of the proportion of reads assigned to the source genome at or above the genus level. Horizontal grey line highlights a match proportion of 0.95. Boxplots hinges represent the 25th and 75th percentiles, line through box represent is the median, whiskers are the 95% confidence interval, and the black dots are outliers.

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. Represenative 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 respecitively.

were evaluated at the genus level. For all the uncontaminated representative set of target genomes the proportion of simulated reads that matched at species level or higher was 0.98 (Table 2).

The proportion of reads assigned to the contaminant genus was comparable to the expected proportion (Fig. 3).

The lowest proportion of simulated contaminant detected varied by both contaminant and taget genome. All organisms had comparable minimum contamination levels for which reads were assigned to the contaminat genome. Two notable exceptions are *Escherichia* and *Yersinia*, where *Bacillus*, and *Salmonella* and *Escherichia* were detected at the lowest contaminant levels respectively. As the results are from simulated data and based on proportions of simulated reads, these values do not indicate a limit of dection for the method.

CONCLUSIONS

- Proof of concept study additional work required to validate use in assessing the purity of a test material.
- Use of other taxonomic classification methods are likely to have different sensitivity and specificity results.
- Need to evaluate the suitablility of the reference database for used the genome and contaminant of interest.
- Work to further expand the taxonomic database to include genomes from uncultured organism using either metagenome datasets for single cell datasets along with efforts to address issues related to taxnomic ambiguities will help to improve the method applicability.

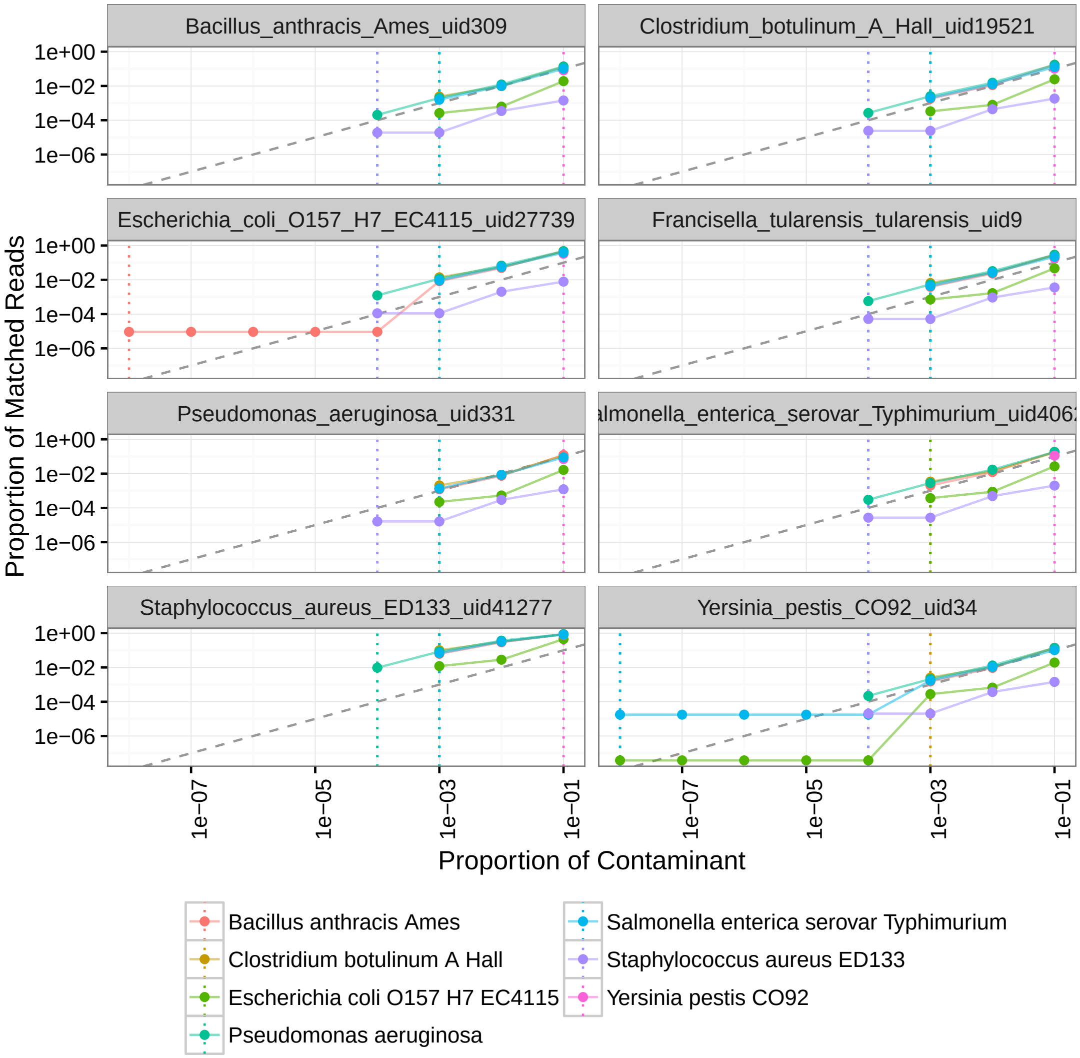

Figure 3. Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

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

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.

CLSI (2010). Characterization and Qualification of Commutable Reference Materials for Laboratory Medicine ; Approved Guideline. Technical Report 22.

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.

EPA (2004). Quality Assurance/Quality Control Guidance for Laboratories Performing PCR Analyses on Environmental Samples October 2004. *October*, (October).

Feldsine, P., Abeyta, C., and Andrews, W. (2002). AOAC international methods committee guidelines for validation of qualitative and quantiative food microbiological official methods of analysis. Technical Report May.

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*.

Guide, E. (1998). The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics. Technical report.

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.

ISO/TS (2010). Microbiology of food and animal feeding stuffs - specific requirements and quidance for proficiency testing by interlaboratory comparison. Technical report.

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.

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

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of computational biology : a journal of computational molecular cell biology*, 19(6):796–813.

Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3):e17288.

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.

Tanner, M. A., Goebel, B. M., Dojka, M. A., and Pace, N. R. (1998). Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. *Appl. Envir. Microbiol.*, 64(8):3110–3113.

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

Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE*, 8(4):e60234.