1 **Method for evaluating genomic material**

2 **purity using whole genome sequencing**

3 **data.**

4 **Nathan D. Olson**1**, Justin Zook**1**, Jayne Morrow**1**, and Nancy Lin**1

5 1**Biosystems and Biomaterials Division, National Institute of Standards and Technology**

6 **ABSTRACT**

7

Dummy abstract text.

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

9 **INTRODUCTION**

10 Rapid, sensitive and accurate assays for detecting bacterial pathogens in food, water, clinical samples,

11 and suspicious biothreats are critical to public health and safety. Biodetection assays must be evaluated for

12 assay sensitivity and specificity prior to deployment and then in the hands of the user to instill confidence

13 in the actions made based on assay results ([Ieven et al.](#_bookmark15), [2013](#_bookmark15); [Coates et al.](#_bookmark9), [2011](#_bookmark9); [EPA](#_bookmark10), [2004](#_bookmark10); [ISO/TS](#_bookmark16),

14 [2010](#_bookmark16); [Guide](#_bookmark13), [1998](#_bookmark13); [Feldsine et al.](#_bookmark11), [2002](#_bookmark11)). Test materials are used to validate assay performance. Test

15 materials can be either purified cultures, genomic DNA or whole cells spiked into a matrix [(EPA](#_bookmark10), [2004](#_bookmark10);

16 [ISO/TS](#_bookmark16), [2010](#_bookmark16); [CLSI](#_bookmark8), [2010](#_bookmark8)). Before being used to evaluate a biodetection assay, the material itself must

17 be validated in terms of purity and identity to eliminate false positive results due to test material contami-

18 nants or false negatives due to the test material being the wrong strain [(CLSI](#_bookmark8), [2010](#_bookmark8)). There are a number

19 of potential sources of microbial contaminants including the stock culture, preservation medium, and airborne and laboratory contaminants [(Marron et al.](#_bookmark20), [2013](#_bookmark20); [Shrestha et al.](#_bookmark28), [2013](#_bookmark28); [Tanner et al.](#_bookmark29), [1998](#_bookmark29)).

21 Currently polymerase chain reaction (PCR) assays are the most commonly used method for evalu-

22 ating test material purity. Other methods to detect contaminants using whole genome sequencing

23 datasets have been developed, but they are not currently used to evaluate test material purity. A PCR assay was

24 developed to analyze protist cultures. This assay uses endpoint PCR for prokaryotes and eukaryotes

25 with template dilutions ([Marron et al.](#_bookmark20), [2013](#_bookmark20)). The benefit to PCR-based approaches is that they can

26 be cost effective and fast if an applicable protocol exists. While PCR assays can detect contaminants,

27 this approach does not scale to multiple contaminants and test materials. More importantly, PCR assays

28 can only target specific contaminants, which biases the purity assessment to known potential contaminants.

29 The bioinformatics tools developed to identify contaminants in metagenomic datasets, which include

30 sequencing data from all organisms in a sample, can also be used to evaluate test material purity. For

31 example DeconSeq ([Schmieder and Edwards](#_bookmark24), [2011](#_bookmark24)) and a similar method QC-Chain ([Zhou et al.](#_bookmark32), [2013](#_bookmark32))

32 were developed to identify contaminants based on analysis of 16S ribosomal ribonucleic acid (rRNA)

33 gene sequences or comparison of a subset of reads to a reference database using Basic local alignment

34 search tool (BLAST). Metagonomic-based methods are ideally able to identify contaminants without any prior

35 knowledge or assumptions regarding the identity of the organism(s). However, methods based on 16S

36 rRNA gene identification have limited resolution, as 16S rRNA sequences can only provide genus level

37 taxonomic resolution at best. The benefit to using metagenomic tools developed for 16S rRNA is that prior knowledge

38 of the identity of the contaminant is not required; however, this method is unable to identify contaminants

39 to the species level or higher.

40 Another approach to evaluating test material purity is through shotgun whole genome sequencing, i.e.,

41 sequence all DNA in a purportedly single organism sample. There are a number of metagenomic read classifica-

42 tion algorithms developed to determine the taxonomic composition of a sequence dataset of unknown

43 composition. These algorithms tend to use one of three primary strategies for taxonomic assignment.

44 The first method consist of aligning reads to a reference database that contains assemblies of microbial genomes ([Buchfink et al.](#_bookmark6), [2015](#_bookmark6); [Francis et al.](#_bookmark12),

45 [2013](#_bookmark12)). This approach, while exhaustive, is computationally expensive. The second type of method focuses

46 on marker genes, genes common to different phylogenetic groups, which reduces the computational cost

47 [(Segata et al.](#_bookmark27), [2012](#_bookmark27); [Liu et al.](#_bookmark19), [2011](#_bookmark19)). The disadvantage of using only marker genes is that infor-

48 mation required to discriminate closely related genomes may not be present in the marker genes. The

49 third method uses a *k*-mer based approach, where taxonomic composition is determined based the abun-

50 dance of DNA sequences of length *k* in the sequence dataset and a reference database ([Ounit et al.](#_bookmark22), [2015](#_bookmark22);

51 [Menzel et al.](#_bookmark21), [2016](#_bookmark21); [Wood and Salzberg](#_bookmark31), [2014](#_bookmark31)).

52 In this work, we present the results of a proof of concept study to measure the purity of single

53 organism test materials using whole genome sequencing data combined with a metagenomic read classi-

54 fication algorithm. We choose to use *Pathoscope*, a method that aligns sequences to a database of genome assemblies. It was developed to detect pathogens and identify strains using whole genome sequencing data [(Francis et al.](#_bookmark12), [2013](#_bookmark12)). The pathogen detection problem is similar to contaminant detection, as the organism of interest is likely

57 present at low concentrations. *Pathoscope* benefits from the large sample size obtained using all sequence

58 data for higher sensitivity (compared to marker gene based methods) and leverages algorithmic advances

59 for whole genome sequence mapping. We will first present the specificity of the method using simulated

60 data for single organisms. Then, we evaluate sensitivity of the method using simulated contaminanted test

61 material datasets.

62 **METHODS**

63 To test the suitability of using whole genome sequence data and metagenomic taxonomic classifica-

64 tion methods for evaluating material purity, we first simulated whole genome sequence data from single

65 genomes and also simulated contaminanted datasets. Simulated data from single genomes was used to as-

66 sess method specificity and simulated contaminant datasets method sensitivity. Simulated datasets were

67 generated using the ART sequencing read simulator [(Huang et al.](#_bookmark14), [2012](#_bookmark14)). The datasets were generated

68 using the Illumina MiSeq error models for 2x230 base pair paired end reads and 20 X mean coverage with

69 an average insert size of 690 base pairs with standard deviation of 10 bp for each strain. The seed number

70 for the random number generator was randomly assigned and recorded for each dataset.

71 The taxonomic composition of simulated datasest was assessed using the Pathoscope metagenomic

72 taxonomic classifier [(Francis et al.](#_bookmark12), [2013](#_bookmark12)). This method uses an expectation maximization algorithm

73 where the sequence data are first mapped to a database comprised of all sequence data in the Genbank

74 nt database. Then, through an iterative process, it re-assigns ambiguously mapped reads based on the

75 proportion of reads mapped unambigously to individual taxa in the database. The PathoScope 2.0 taxo-

76 nomic read classification pipeline includes an initial read filtering step (PathoQC), followed by mapping

77 reads to a reference database (PathoMap) - a wrapper for bowtie2 ([Langmead and Salzberg](#_bookmark18), [2012](#_bookmark18))), followed by

78 an expectation-maximization classification algorithm (PathoID). The annotated Genbank nt database pro-

[79](ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz) vided by the PathoScope developers was used as the reference database ([ftp://pathoscope.bumc.](ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz)

80 [bu.edu/data/nt\_ti.fa.gz](ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz)).

81 **Specificity**

82 Method specificity was defined as the proportion of reads in a single organism simulated dataset assigned

83 to a taxonomy different from the taxonomy of the test material. Sequence data was simulated for 406

84 strains, from 9 genera (Table [1](#_bookmark0)). We will refer to the genome used to generate the reads as the target

85 genome. The genomes included in the simulation study were limited to the number of closed genomes

86 in the Genbank database (<http://www.ncbi.nlm.nih.gov/genbank/>, accessed 10/18/2013)

87 belonging to the genera of interest (Table [1](#_bookmark0)). Due to the large number of closed genomes genomes from the gen-

88 era *Bacillus*, *Escherichia*, and *Salmonella*, these generawere limited to the species *Bacillus cereus*, *Escherichia*

89 *coli*, and *Salmonella enteria*. The taxononomic heirarchy for the target genome and simulated read assign-

90 ment match levels were determined using the R package ([Scott Chamberlain and Eduard Szocs](#_bookmark26), [2013](#_bookmark26);

91 [Chamberlain et al.](#_bookmark7), [2016](#_bookmark7)).

92 **Sensitivity**

93 We simulated datasets with contaminants to evaluate method sensitivity. Representative genomes for

94 8 of the 9 genus were used to generate the simulated contaminant datasets (Table [2](#_bookmark4)). An *Escherichia*

95 *coli* strain was selected as a representative of both *Eschericha coli* and *Shigella*. For each pairwise

96 combination of representative genomes, the simulated contaminant dataset was subsampled at 0*.*1 to

97 10*−*8 (contaminant proportion), at 10 fold dilutions. The target genome dataset was subsamples at 1 -

98 contaminant proportion, resulting in 512 simulated contaminant datasets. This approach simulates the

99 proportions of cells in a test material and not the amount of DNA, assuming unbiased DNA extraction.

100 To speed up processing the aligned sequence files were subsampled instead of the simulated sequence

101 files.

102 **Reproducibility**

103 To facilitate reproducibility and transparency, a Docker ([www.docker.com](http://www.docker.com/)) container is available with

104 installed pipeline dependencies ([www.registry.hub.docker.com/u/natedolson/docker-pathoscop](http://www.registry.hub.docker.com/u/natedolson/docker-pathoscope/)

[105](https://github.com/nate-d-olson/genomic_purity) The script used to run the simulations are available at [https://github.com/nate-d-olson/](https://github.com/nate-d-olson/genomic_purity)

106 [genomic\_purity](https://github.com/nate-d-olson/genomic_purity). Pathoscope results were processed using the statistical programing language

107 R [(R Core Team](#_bookmark23), [2016](#_bookmark23)), and intermediate analysis and data summaries were organized using ProjectTem-

[108](https://github.com/nate-d-olson/genomic_purity_analysis) plate ([White](#_bookmark30), [2014](#_bookmark30)) and archived in a github repository ([https://github.com/nate-d-olson/](https://github.com/nate-d-olson/genomic_purity_analysis)

109 [genomic\_purity\_analysis](https://github.com/nate-d-olson/genomic_purity_analysis)).

110 **RESULTS AND DISCUSSION**

111 **Specificity**

112 Simulated sequence data from individual isolates was used to assess the genomic purity assessment

113 method specificity. Here we use specificity as a measure of the ability of the method to correctly assign reads to

114 the taxonomy of the genome the sequencing reads were simulated from, the target genome. True negatives (TNs) are reads assigned to the target genome’s species, genus, family, etc., depending on the match stringency, and false positives (FPs) are reads incorrectly assigned to a different species, genus, family, etc., and specificity = TN/(FP+TN) (Fig. [1](#_bookmark1)). Overall high proportion of matches at species and

117 genus level.

Some genera have low specificity at the species and genus levels.

120 For *Shigella* most likely due to matches with *Escherichia* (Fig. [2](#_bookmark2)).

|  |  |  |
| --- | --- | --- |
| 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

121 Most of the genera had genus level or higher match proportions excluding a few outliers (Fig. [3](#_bookmark3)).

122 *Escherichia*, *Shigella*, and *Staphylococcus* are notable exceptions. As discussed previously the taxo-

123 nomic ambiguities for *Shigella* and *Escherichia* are responsible for the overall lower genus level match

124 proportions. Another example of low genus level matches is the *Bacillus* genome with genus match pro-

125 portion close to zero, *Bacillus infantis* string NRRL B 14911. While the *B. infantis* strain was originally

126 classified as *Bacillus* the species is phylogenetically distinct from other members of the genus ([Ko et al.](#_bookmark17),

127 [2006](#_bookmark17)). It is important to consider the strain and genome being characterized, as taxonomic ambiguities

128 (e.g. *Shigella* and *Escherichia*) can lead to lower than expected specificity and the identification of false

129 positive contaminants.

130 **Sensitivity**

131 To evaluate genomic purity assessment methods we generated simulated contaminant datasets as pair-

132 wise combinations of representative genomes from 8 of the genera used in the specificity section of the

1.00

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Bacillus* | | | | | | | |  | *Clostridium* | | | | | | | |  | *Escherichia* | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

0.75

0.50

0.25

0.00

1.00

Cumulative Match Proportion

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Listeria* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Pseudomonas* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Salmonella* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

0.75

0.50

0.25

0.00

1.00

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Shigella* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Staphylococcus* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Yersinia* | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

0.75

0.50

0.25

0.00

species genus family order class phylum

superkingdom

species genus family order class phylum

superkingdom

species genus family order class phylum

superkingdom

# Taxonomic Level

**Figure 1.** Cumulative taxonomic match results for genomic purity assessments 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.

1.00

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

0.95

Cumulative Match Proportion

0.90

0.85

genus

family

class

# Taxonomic Level

species

superkingdom

**Figure 2.** Cumulative taxonomic match results for genomic purity assement for *Shigella* considering matches to *E. coli* as species level matches. 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.

1.00

0.75

0.50

Genus Match Proportion

0.25

0.00

*Bacillus*

*Clostridium*

*Escherichia*

*Listeria*

*Salmonella*

*Shigella*

*Yersinia*

Query Genus

*Pseudomonas*

*Staphylococcus*

**Figure 3.** 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.

133 study (Table [2](#_bookmark4)). Due to the overall high proportion of reads matched to the correct genome in the method

134 specificity study, the simulated contaminant datasets were evaluated at the genus level for sensitivity. For all of the genomes selected for the sensitivity study, the proportion of simulated reads that matched at species

136 level or higher was 0.98 (Table [2](#_bookmark4)).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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.

137 To evaluate sensitivity, we plot the proportion of reads assigned to the contaminant genus or species vs. the proportion of reads simulated from the contaminating genome. While the proportion of contaminant reads in the simulated datasets was not equal to the defined con-

138 taminant proportion, the proportion of reads assigned to the contaminant genus was comparable to the

139 expected proportion (Fig. [4](#_bookmark5)). This was especially true for datasets containing mixtures of *B. anthracis*,

140 *Y. pestis*, *E. coli*, and *S. enteria* as they had similar sized genomes (Table [2](#_bookmark4)). Three contaminants were

141 detected when spiked in at contaminant proportions of 10*−*8, *B. anthracis* in *E. coli* as well *S. enteria*

142 and *E. coli* in *Y. pestis*. Interestingly the proportion of assigned reads did not decrease with decreasing

143 contaminant proportions after 10*−*4.

144 The lowest detectable simulated contaminant level varied by both contaminant and taget

145 genome. All organisms had comparable minimum contamination levels for which reads were assigned

146 to the contaminant genome. Two notable exceptions are *Escherichia* and *Yersinia*, where *Bacillus*, and

147 *Salmonella* and *Escherichia* were detected at the lowest contaminant levels respectively. As the results

148 are from simulated data and based on proportions of simulated reads, these values do not indicate a limit

149 of detection for the method.

150 **CONCLUSIONS**

151 • Proof of concept study additional work required to validate use in assessing the purity of a test

152 material.

153 • Use of other taxonomic classification methods are likely to have different sensitivity and specificity

154 results.

155 • Need to evaluate the suitability of the reference database for used the genome and contaminant of

156 interest.

157 • Work to further expand the taxonomic database to include genomes from uncultured organism us-

158 ing either metagenome datasets for single cell datasets along with efforts to address issues related

159 to taxonomic ambiguities will help to improve the method applicability.

1e+00

1e−02

1e−04

1e−06

1e+00

1e−02

1e−04

1e−06

Proportion of Matched Reads

1e+00

1e−02

1e−04

1e−06

1e+00

1e−02

1e−04

1e−06

1e+00

1e−02

1e−04

1e−06

1e+00

1e−02

1e−04

1e−06

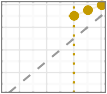
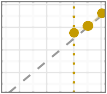
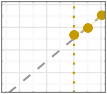
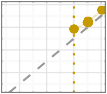
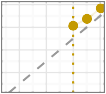
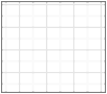
1e+00

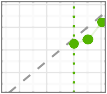
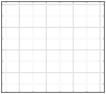
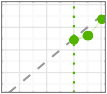
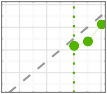
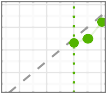
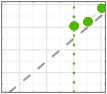
1e−02

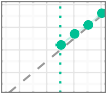
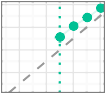
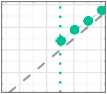
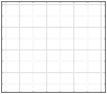
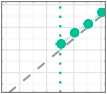
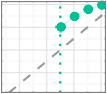
1e−04

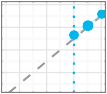
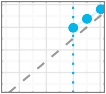
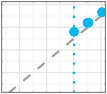
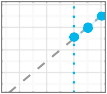
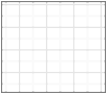
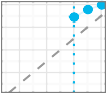
1e−06

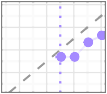
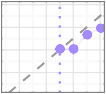
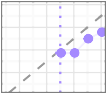
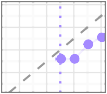
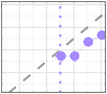
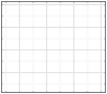
Baci Clos Esch Fran Pseu Salm Stap Yers

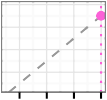
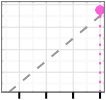
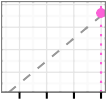
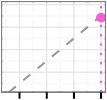
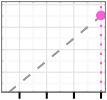
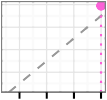


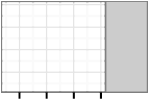
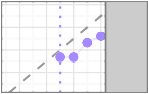
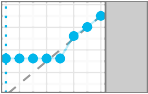
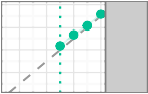
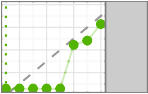
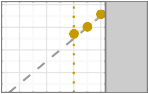
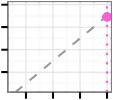
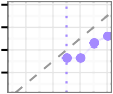
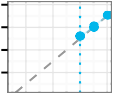
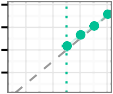
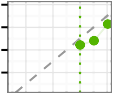
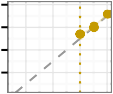
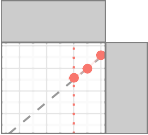
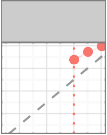
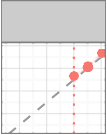
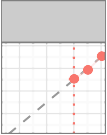
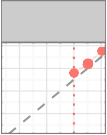
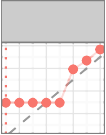
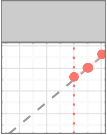
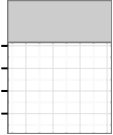
     

Contaminant Proportion

Baci

Clos

Esch

Pseu

Salm

Stap

Yers

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

1e−07

1e−05

1e−03

1e−01

Contaminant

Bacillus anthracis Ames Clostridium botulinum A Hall Escherichia coli O157 H7 EC4115 Pseudomonas aeruginosa

Salmonella enterica serovar Typhimurium Staphylococcus aureus ED133

Yersinia pestis CO92

**Figure 4.** Relationship between the proportion of contaminant reads simulated per dataset and the proportion of reads matched to the contaminant genus.

160 **ACKNOWLEDGMENTS**

161 The authors would like to thanks Dr. Steven Lund for his assistance in developing the study. The Depart-

162 ment of Homeland Security (DHS) Science and Technology Directorate supported this work under the

163 Interagency Agreement HSHQPM-12-X-00078 with the National Institute of Standards and Technology

164 (NIST). Opinions expressed in this paper are the authors and do not necessarily reflect the policies and

165 views of DHS, NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are

166 identified in this paper in order to specify the experimental procedure adequately. Such identification

167 is not intended to imply recommendations or endorsement by NIST, nor is it intended to imply that the

168 materials or equipment identified are necessarily the best available for the purpose. Official contribution

169 of NIST; not subject to copyrights in USA.

170 **REFERENCES**

171 Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond.

172 *Nature methods*, 12(1):59–60.

173 Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., Foster, Z., and

174 O’Donnell, J. (2016). *taxize: Taxonomic information from around the web*. R package version 0.7.4.

175 CLSI (2010). Characterization and Qualification of Commutable Reference Materials for Laboratory

176 Medicine ; Approved Guideline. Technical Report 22.

177 Coates, S. G., Brunelle, S. L., and Davenport, M. G. (2011). Development of standard method perfor-

178 mance requirements for biological threat agent detection methods. *Journal of AOAC International*,

179 94(4):1328–37.

180 EPA (2004). Quality Assurance/Quality Control Guidance for Laboratories Performing PCR Analyses

181 on Environmental Samples October 2004. *October*, (October).

182 Feldsine, P., Abeyta, C., and Andrews, W. (2002). AOAC international methods committee guidelines for

183 validation of qualitative and quantiative food microbiological official methods of analysis. Technical

184 Report May.

185 Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q.,

186 Schaalje, G. B., Clement, M. J., Crandall, K. a., and Johnson, W. E. (2013). Pathoscope: Species

187 identification and strain attribution with unassembled sequencing data. *Genome research*.

188 Guide, E. (1998). The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method

189 Validation and Related Topics. Technical report.

190 Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read

191 simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4.

192 Ieven, M., Finch, R., and van Belkum, a. (2013). European quality clearance of new microbiological

193 diagnostics. *Clinical microbiology and infection : the official publication of the European Society of*

194 *Clinical Microbiology and Infectious Diseases*, 19(1):29–38.

195 ISO/TS (2010). Microbiology of food and animal feeding stuffs - specific requirements and quidance

196 for proficiency testing by interlaboratory comparison. Technical report.

197 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).

198 Bacillus infantis sp. nov. and bacillus idriensis sp. nov., isolated from a patient with neonatal sepsis.

199 *International journal of systematic and evolutionary microbiology*, 56(11):2541–2544.

200 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*,

201 9(4):357–9.

202 Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of

203 taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12(1):1.

204 Marron, A. O., Akam, M., and Walker, G. (2013). A Duplex PCR-Based Assay for Measuring the

205 Amount of Bacterial Contamination in a Nucleic Acid Extract from a Culture of Free-Living Protists.

206 *PloS one*, 8(4):e61732.

207 Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metage-

208 nomics with kaiju. *Nature communications*, 7.

209 Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). Clark: fast and accurate classification of

210 metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16(1):1.

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

212 Statistical Computing, Vienna, Austria.

213 Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from

214 genomic and metagenomic datasets. *PloS one*, 6(3):e17288.

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

216 *F1000Research*.

217 Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metage-

218 nomic microbial community profiling using unique clade-specific marker genes. *Nature methods*,

219 9(8):811–814.

220 Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When Is a Microbial Culture Pure

221 ? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing.

222 Tanner, M. A., Goebel, B. M., Dojka, M. A., and Pace, N. R. (1998). Specific Ribosomal DNA Se-

223 quences from Diverse Environmental Settings Correlate with Experimental Contaminants. *Appl. Envir.*

224 *Microbiol.*, 64(8):3110–3113.

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

226 package version 0.6.

227 Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using

228 exact alignments. *Genome biology*, 15(3):1.

229 Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: Fast and Holistic Quality Control

230 Method for Next-Generation Sequencing Data. *PLoS ONE*, 8(4):e60234.