Pathoscope: Species identification and strain attribution with unassembled sequencing data

Owen E. Francis, Matthew Bendall, Solaiappan Manimaran, et al.

Genome Res. published online July 10, 2013
Access the most recent version at doi:10.1101/gr.150151.112

Pathoscope: Species identification and strain attribution with unassembled sequencing data

Owen E. Francis[a], Matthew Bendall[b], Solaiappan Manimaran[c], Changjin Hong[c], Nathan L. Clement[d], Eduardo Castro-Nallar[e], Quinn Snell[d], G. Bruce Schaalje[a], Mark J. Clement[d], Keith A. Crandall[e,*], and W. Evan Johnson[c,*]

[a]*Department of Statistics, Brigham Young University, 223 TMCB, Provo, UT 84602;*

[b]*Department of Biology, Brigham Young University, 680 WIDB, Provo, UT 84602;*

[c]*Division of Computational Biomedicine, Boston University School of Medicine, 72 E. Concord St., Boston, MA 02118*

[d]*Department of Computer Science, Brigham Young University, 3200 TMCB, Provo, UT 84602;*

[e]*Computational Biology Institute, George Washington University, Ashburn, VA 20147*

* To whom correspondence should be addressed: wej@bu.edu, kcrandall@gwu.edu

Keywords: metagenomics, bioforensics, biosurveillance, infectious disease diagnosis, pathology, next-generation sequencing.

Abstract

Emerging next-generation sequencing technologies have revolutionized the collection of genomic data for applications in bioforensics, biosurveillance, and for use in clinical settings. However, to make the most of these new data, new methodology needs to be developed that can accommodate large volumes of genetic data in a computationally efficient manner. We present a statistical framework to analyze raw next-generation sequence reads from purified or mixed environmental or targeted infected tissue samples for rapid species identification and strain attribution against a robust database of known biological agents. Our method, *Pathoscope*, capitalizes on a Bayesian statistical framework that accommodates information on sequence quality, mapping quality and provides posterior probabilities of matches to a known database of target genomes. Importantly, our approach also incorporates the possibility that multiple species can be present in the sample and considers cases when the sample species/strain is not in the reference database. Furthermore, our approach can accurately discriminate between very closely related strains of the same species with very little coverage of the genome and without the need for multiple alignment steps, extensive homology searches, or genome assembly—which are time consuming and labor intensive steps. We demonstrate the utility of our approach on genomic data from purified and *in silico* 'environmental' samples from known bacterial agents impacting human health for accuracy assessment and comparison with other approaches. Software is available at: https://sourceforge.net/projects/pathoscope/.

Introduction

The accurate and rapid identification of species and strains of pathogens is an essential component of biosurveillance from both human health and biodefense perspectives (Vaidyanathan 2011). For example, misidentification was among the issues that resulted in a three week delay in accurate diagnosis of the recent outbreak of hemorrhagic *Escherichia coli*

being due to strain O104:H4 resulting in over 3,800 infections across 13 countries in Europe with 54 deaths (Frank et al. 2011). The most accurate diagnostic information, necessary for species identification and strain attribution, comes from the most refined level of biological data—genomic DNA sequences (Eppinger et al. 2011). Advances in DNA sequencing technologies allows for the rapid collection of extraordinary amounts of genomic data, yet robust approaches to analyze this volume of data are just developing, from both statistical and algorithmic perspectives.

Next-generation sequencing approaches have revolutionized the way we collect DNA sequence data, including for applications in pathology, bioforensics, and biosurveillance. Given a particular clinical or metagenomic sample, our goal is to identify the specific species, strains or substrains present in the sample, as well as accurately estimate the proportions of DNA originating from each source genome in the sample. Current approaches for next-gen sequencing usually have read lengths between 25-1000 base pairs; however, these sequencing technologies include error rates that vary by approach and by samples. Such variation is typically less important for species identification given the relatively larger genetic divergences among species than among individuals within species. But for strain attribution, sequencing error has the potential to swamp out discriminatory signal in a data set, necessitating highly sensitive and refined computational models and a robust database for both species identification and strain attribution.

Current methods for classifying metagenomic samples rely on one or more of three general approaches: composition or pattern matching (McHardy et al. 2007; Brady and Salzberg 2009; Segata et al. 2012), taxonomic mapping (Huson et al. 2007; Meyer et al. 2008; Monzoorul Haque et al. 2009; Gerlach and Stoye 2011; Patil et al. 2012; Segata et al. 2012), and whole genome assembly (Kostic et al. 2011; Bhaduri et al. 2012). Composition and pattern matching

algorithms use pre-determined patterns in the data, such as taxonomic clade markers (Segata et al. 2012), *k*-mer frequency, or GC content, often coupled with sophisticated classification algorithms such as support vector machines (McHardy et al. 2007; Patil et al. 2012) or interpolated Markov Models (Brady and Salzberg 2009) to classify reads to the species of interest. These approaches require intensive preprocessing of the genomic database before application. In addition, the classification rule and results can often change dramatically depending on the size and composition of the genome database.

Taxonomy based approaches typically rely on a “lowest common ancestor” approach (Huson et al. 2007), meaning that they identify the most specific taxonomic group for each read. If a read originates from a genomic region that shares homology with other organisms in the database, the read is assigned to the lowest taxonomic group that contains all the genomes that share the homologous region. These methods are typically highly accurate for higher-level taxonomic levels (e.g. phylum and family) but experience reduced accuracy at lower levels (e.g. species and strain) (Gerlach and Stoye 2011). Furthermore, these approaches are not informative when the reads originate from one or more species or strains that are closely related to each other or different organisms in the database. In these cases, all the reads can be reassigned to higher-level taxonomies, thus failing to identify the specific species or strains contained in the sample.

Assembly-based algorithms can often lead to the most accurate strain identification. However, these methods also require the assembly of a whole genome from a sample, which is a computationally difficult and time-consuming process that requires large numbers of reads to achieve an adequate accuracy—often on the order of 50-100× coverage of the target genome (Schatz et al. 2010). Given current sequencing depths, obtaining this level of coverage is usually possible for purified samples, but coverage levels may not be sufficient for mixed samples or in multiplexed sequencing runs. Assembly approaches are further complicated by

the fact that data collection at a crime scene or hospital might include additional environmental components in the biological sample (host genome or naturally occurring bacterial and viral species), thus requiring multiple filtering and alignment steps in order to obtain reads specific to the pathogen of interest.

Here we describe an accurate and efficient approach to analyze next-generation sequence data for species identification and strain attribution that capitalizes on a Bayesian statistical framework, implemented in the new software package *Pathoscope v1.0*. Our approach accommodates information on sequence quality, mapping quality, and provides posterior probabilities of matches to a known database of reference genomes. Importantly, our approach incorporates the possibility that multiple species can be present in the sample or that the target strain is not even contained within the reference database. It also accurately discriminates between very closely related strains of the same species with much less than 1× coverage of the genome and without the need for sequence assembly or complex preprocessing of the database or taxonomy. No other method in the literature can identify species or substrains in such a direct and automatic manner and without the need for large numbers of reads. We demonstrate our approach through application to next-generation DNA sequence data from a recent outbreak of the hemorrhagic *E. coli* (O104:H4) strain in Europe (Frank et al. 2011; Rohde et al. 2011; Turner 2011) and on purified and *in silico* mixed samples from several other known bacterial agents impacting human health. Software and data examples for our approach are freely available for download at: https://sourceforge.net/projects/pathoscope/.

Results

Overview of the Identification Approach

For the purposes of this demonstration, we constructed a reference database bacterial genomes obtained from GenBank chosen based on their phylogenetic affinity to eight bacterial agents from the "CDC Category A and B lists of bioterrorism agents/diseases" (http://www.bt.cdc.gov/agent/agentlist-category.asp). The query next-gen sequencing reads were independently aligned to the reference genomes using three different aligners, BLAST, GNUMAP, and Bowtie 2 (exact parameters used are given in the Methods section below). Reads with a single or unique alignment to only one organism in the database were denoted as uniquely mapped reads*, or unique reads* in short. However, since our database contains many closely related species and strains, many of the sequence reads map to multiple genomes in the database. These reads are denoted as *non-unique reads*. Reads that do not match any genome in the database are only utilized to help determine whether the source species is present in the database. From data examples presented in the sections below, we observed that between 6.4% and 99.9% of the reads map to multiple organisms depending on the number of closely related strains in the database (see Figure 1).

When reads align to multiple genomes due to their sequence similarity, the reads are less likely to be assigned to the correct source genome. For example, in the *E. coli* K12 MG1655 example described below, more than 99.9% of the reads aligned to multiple genomes due to the presence of multiple related substrains in the database. In this case the correct genome received the same proportions of the reads as a closely related, but incorrect, substrain due to non-uniqueness. This leads to the inability to conclusively identify the correct substrain—especially for methods based only on the alignment, context matching, homology searching, or genome assembly. However, by reassigning the ambiguous reads, we show below that it is possible to remove reads assigned to genomes that are less likely to be the source of the reads and reassign them to the source template of the reads.

Through an iterative process, our novel Bayesian read reassignment method is capable of identifying the genomes that are the most likely source of the reads. However, even though a set of reads could have originated from the DNA from multiple organisms, each individual read was derived from one template DNA strand that came from a single organism. To correctly and precisely identify the species present in the sample, the non-unique read probabilities must be reassigned to the correct template genome of origin. To address this need, we have formulated a Bayesian *missing data* mixture modeling approach (where the template genome of origin is the 'missing data') that integrates information contained within the read (mapping probability) with information obtained by *borrowing strength* across all reads from the sample (e.g. proportions of unique reads or imbalances in non-unique probabilities across all reads). This approach is superior to a naïve mapping approach that assigns reads based on information contained solely in the reads. Using this additional information helps to overcome mistakes in mapping caused by sequencing errors or low quality bases.

Application to the European E. coli Outbreak of 2011

The recent outbreak of *Escherichia coli* (*E. coli*) O104:H4 in Europe resulted in a number of deaths that may have been prevented by an early identification of the affecting pathogen. We obtained 92,370 sequencing reads from an O104:H4 sequencing run generated at the Beijing Genome Institute (Shanghai, China), using the Ion Torrent sequencing technology (Guilford, CT) (ftp://climb.genomics.cn/pub/10.5524/100001_101000/100001/run1.fastq.gz). Most of the reads in the data set (94.1%) ranged in length from 80bp to 120bp. We used BLAST (Altschul et al. 1997), Bowtie 2 (Langmead and Salzberg 2012), and GNUMAP (Clement et al. 2010) to independently align these query reads to our reference database,, which included the genomes of 30 strains of *E. coli*—many of which were closely related to the O104:H4 strain. The *Pathoscope* results from the BLAST, Bowtie 2, and GNUMAP alignments were nearly identical (<1% different), so we only report the results from the BLAST alignment below.

In addition to *Pathoscope*, we compared several other approaches for inferring the genomic source of sequencing reads. These included a naïve mapping strategy, where we aligned reads to the database and generated a posterior probability of alignment based on the read's alignment score for each genome. The read probabilities are then summed for each genome, resulting in the total (probabilistic) portion of the reads mapped to each specific genome. We also compared with PhymmBL (Brady and Salzberg 2009), MEGAN4 (Huson et al. 2007), PhyloPhythiaS (Patil et al. 2012), and MetaPlAn (Segata et al. 2012). Finally, we applied an alignment approach using the Trinity assembler (Grabherr et al. 2011) to assemble high quality contiguous sequences (contigs) from the reads followed by the probabilistic alignment of the contigs to the database (see Methods for specific parameter settings for each algorithm).

For this example, we used the full dataset of 92,370 reads, representing 1.3× coverage of the reference O104:H4 genome, as well as reduced datasets using 1,000 random subsamples of reads for each of the following sample sizes: 9,237 (0.13×), 924 (0.01×), and 92 (0.001×). For the smaller subsets (92, 924, 9,237), we compared the average accuracy and range across samples for each method. These smaller sets were designed to evaluate algorithmic performance when the reads are generated using multiplexed sequencing runs or when they originated from contaminated samples that may be dominated by other genomic sources. However, we note that for MEGAN (graphical user interface), PhyloPhythiaS (manual webserver), and the assembly approach, we did not use 1,000 random datasets; rather we used a single random sample of each dataset size, as they would either require thousands of manual submissions or an excessive amount of computation time. Table 1 contains the average accuracy and range across samples for each algorithm.

Naïve alignment, PhymmBL, and MetaPhlAn: The naïve algorithm consistently assigned around 12.9% of the read probability to the O104:H4 strain independent of the number of reads used. However, on average between 7.4% and 9.4% of the read probability was assigned to the *E. coli* 55989, which is the closest fully sequenced genome to the O104:H4 strain (Rohde et al. 2011; Turner 2011). Several other *E. coli* strains received 1-3% of the reads, and several species in the *Shigella* genus also received 1-2% of the reads. In all, roughly 93% of the read probabilities were assigned to an *E. coli* strain. The PhymmBL algorithm assigned 14.7% on average to O104:H4 strain and exhibited similar profiles of false mapping to other strains and species. Overall the performance of PhymmBL was only slightly better than the naïve approach. The MetaPhlAn algorithm aligns reads to taxonomic clade-specific markers, which in its current implementation can only identify DNA templates at the species level—and therefore cannot distinguish between strains or substrains of the same species. In addition, because it only uses short clade markers, merely 815 (0.9%) of the reads were assigned by MetaPhlAn. Of these reads only 90.0% were aligned to *E. coli*, whereas 9.6% were incorrectly assigned to *S. dysenteriae*. The method gave inconsistent results for the subsamples of 9,237 and most of the time failed to assign any reads to *E. coli* for the subsamples of 92 and 924. From these approaches, it is clear that an *E. coli* strain is present in the sample and the naïve and PhymmBL approaches point to O104:H4 as the most likely source, but all results are ambiguous as to whether there are multiple *E. coli* strains or other species present in the sample.

Genome assembly approach: For the assembly approach, no contigs were generated from the 92 and 924 read datasets. For the dataset with 9,237 reads, only 5 contigs were generated ranging in length from 221 bases to 442 bases in length (N50=409; N90=221). Although these five contigs best matched to the O104:H4 strain, they also aligned to several other (incorrect) genomes in the database. Finally, on the complete sequencing run representing 1.3× coverage of the genome, the assembler constructed 3,637 short contigs (N50=292; N90=216) with only

21.5% of the contig mapping probability being assigned to the correct strain. Therefore, although this approach is a slight improvement over the naïve approach or context mapping, it is clear that a single sequencing run for a purified (single source) sample is not sufficient for strain attribution using an assembly-based approach.

Pathoscope reassignment: In contrast as shown in Table 1, *Pathoscope* reassigned on average 99.4% of the read probability directly to the O104:H4 strain for the datasets with 92 reads, and averaged 99.6% of the reads correctly for the larger datasets. These results imply that *Pathoscope* is a substantial improvement over naïve mapping, context mapping, and assembly-based methods for species identification and strain attribution.

Identification of the nearest genome

The results from the MEGAN and PhyloPythiaS analyses were not included in the previous section because the annotation tables used by these approaches do not contain the O104:H4 strain (and cannot be manually added by the user). For this reason, we removed the O104:H4 strain from our reference database and re-analyzed the query reads using the naïve mapping and *Pathoscope* reassignment. In addition, we note that the PhyloPythiaS web server only allowed for a maximum of 10,000 reads for each submission, so the results presented here were based on random sets of 92, 924, and 9,237 only (and not the full dataset).

For the naïve mapping with O104:H4 removed, most of the aligned reads (99.8%) mapped to at least one strain of *E. coli*, thus rapidly and clearly identifying the species of origin. However, 96.1% of these reads aligned ambiguously to multiple *E. coli* strains. The 55989 strain received the largest proportion of the aligned reads (9.5%), followed by the O103:H2 strain (3.2%), the B7A, O26:H11, E24377A, and the E22 strains (3.1%), then the SE11 and IAI1 strains (3.0%). Therefore, although the correct species was easily identified using a naïve mapping strategy,

the identification of the correct strain within the species proves to be more difficult and a simple mapping strategy leaves much uncertainty in the process of identifying the strain most similar to the origin strain. This uncertainty can prove to be important for *E. coli*—which contains both benign and harmful strains—as the misclassification of the origin or nearest strain might lead to negative economic and human health consequences.

In contrast, the lowest common ancestor approach utilized by MEGAN assigned 80.2% of the reads to the family taxonomic level or higher. The remaining reads were assigned at the genus level; 19.7% of the total reads were assigned to the *Escherichia* genus and 0.2% of the reads were incorrectly assigned to the *Shigella* genus. MEGAN did not assign any reads at the species or strain level for any of the datasets. PhyloPhythiaS also performed poorly on this example: overall, more than 84% of the reads were assigned to the family level or above, and less than 50% of all the reads were correctly assigned *E. coli* taxonomy levels. Furthermore, 32 incorrect genera received more reads than *Escherichia*, and 5 incorrect species received more reads than *E. coli*.

After application of the *Pathoscope* reassignment, 89.5% of the reads were reassigned to the 55989 strain. The genomes with the next highest read proportions were the O157:H7 strain (3.2%), and the O103:H2 strain (1.1%). Therefore, even though our approach did not completely converge on one genome (as it shouldn't because in this analysis the origin strain was not present in the database), it is clear that *Pathoscope* can clearly and definitively identify the closest fully sequenced neighboring strain with high confidence.

To evaluate whether the lack of sensitivity for MEGAN and PhyloPythiaS is due to the missing O104:H4 annotation, we applied MEGAN and PhyloPhythiaS to our analysis of reads from the *E. coli* K-12 MG1655 substrain (described in detail below), which is contained in the annotation.

For MEGAN, the result was similar in that all the reads were assigned at the genus level or higher. For PhyloPythiaS, 98.5% of the reads were assigned to the genus level or above, and 34.7% of the reads were assigned to incorrect taxonomies. The *E. coli* species only received 1.4% of the reads, and no reads were assigned at the strain or substrain level. Therefore these methods can fail to identify substrains even when they are present in the annotation.

Computational Time: MetaPhlAn was by far the fastest algorithm (Table 1), requiring only one minute to complete because it aligns the reads to a set of small clade markers, however, the approach assigned less than 1% of the reads in this example. The naïve approach required 38 minutes for a BLAST alignment, 21 minutes for GNUMAP, and 3 minutes for Bowtie 2. *Pathoscope* and MEGAN used the naïve alignments and required an additional 7 minutes and 3 minutes, respectively. PylopythiaS required a total of 7 minutes to assign 9,237 reads. PhymmBL required ~36hrs of database preprocessing, and then approximately 2 hours to assign the reads. Finally, the assembly approach required 30 minutes to complete.

NCBI Sequence Read Archive Datasets

To further evaluate the effectiveness of our method in different scenarios, we obtained sets of reads from twelve different bacterial species/strains from the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra), all of which were sequenced using the 454 platform (Roche, Branford, CT). The reads from each sample were aligned to our full database of genomes and identified as though the true source of the reads were unknown. These datasets consisted of between 28,221 and 1,504,985 reads with read lengths typically ranging from 77 to 277 base pairs. Overall these datasets amounted to only 1.2× to 31.2× coverage of the target genomes. For more than half of these purified sample datasets, the read coverage is not sufficient to fully assemble a genome (Schatz et al. 2010).

Our *Pathoscope* strain attribution method worked extremely well on all of these samples (Table 2). Before reassignment, the read probability assigned to the correct genome ranged between 4.8% and 98.1%. To further evaluate this phenomenon, we plotted the naïve alignment probabilities (along with the *Pathoscope* reassignments) versus the number of closely related strains contained in the database (Figure 1). Clearly, the accuracy of the naïve approach relies heavily on the number of similar genomes in the database, and to distinguish between closely related strains and substrains, a *Pathoscope* reassignment is absolutely necessary for proper identification. After reassignment using *Pathoscope*, the read probability for the correct genome ranged between 92.7% and 99.9%, showing very strong evidence for the correct genome of interest. In nine cases *Pathoscope* reassigned more than 99% of the reads to the correct genome. In the three sets where reassignment led to less than 99%, all had special circumstances and are discussed below. These examples clearly show the benefit of our pathogen detection approach and its ability to reliably identify the correct genome under a diverse set of conditions, not only to species, but also to strain level.

Closely related strains in the database

There were 30 different strains and sub-strains of *E. coli* present in the genome database, three of which were sub-strains of the K-12 strain. Notably, the K-12 MG1655 and the K-12 W3110 sub-strains have greater than 99.9% sequence similarity between the genomes; in fact, a recent study identified only 23 sites with point mutations to differentiate between these genomes (Hayashi et al. 2006). This created difficulty for strain attribution for the naïve mapping strategy: for example, when we attempted to assign reads from the K-12 MG1655 substrain, we observed that only 10.0% of the read probability mapped to K-12 MG1655, 10.0% to K-12 W3110 and 9.4% mapped to K-12 DH10B. Clearly this shows the failure of a naïve mapping strategy and points to the need for a highly sensitive mapping strategy with greater differentiation among substrains. Our reassignment method, *Pathoscope,* was able to confidently reassign the reads

to the correct genome. Our method reassigned an impressive 99.6% of the reads to the *E. coli* K-12 MG1655 genome. The ability to differentiate at the substrain level will become increasingly important as databases of bacterial genomes are rapidly growing.

Unassembled genome

The *Yersinia pestis* KIM D27 dataset provided an interesting scenario that further illustrates the performance of *Pathoscope* in cases where the genome is not fully contained in the reference database. In our database there were 21 different strains or substrains of *Y. pestis*, two of which were substrains of the KIM strain. The correct KIM D27 genome was contained in our database, but was not fully assembled. Specifically, our database contained only 9 contigs of the D27 substrain, whereas the database contained the complete genome of KIM 10 strain. The percentage of read probabilities assigned using a naïve approach mapped only 4.8% of the reads to *Y. pestis* KIM D27, which was closely followed by the KIM 10 substrain (4.4%), and then the Mediaevalis strain (4.4%). After reassignment, 97.3% of the reads were reassigned to the correct KIM D27 unassembled substrain. While impressive, this percentage is smaller than was observed with many of the other genome examples from the SRA read sources, primarily because if the genome database contains closely related species to the target genome, many of the reads from unassembled regions will align to these genomes, resulting in a small but significant read probability for incorrect genomes. The greater probability assigned to the other genomes is an effect of the increased uncertainty due to the incomplete target genome.

Genome not present in the database

As was the case with the European *E. coli* example described previously, to further test our approach in a scenario where the source genome is not present in the database, we focused our attention on the SRA sample from *Francisella tularensis* ATCC 6223 sub-strain. This substrain was not contained in the reference database; however, thirteen strains and substrains

of *F. tularensis*, including five substrains of the *F. t. tularensis* subspecies (type A), were present in the genome database. In this example, only 19.8% of the 67,276 reads mapped to a genome in the database, and 99.4% of these mapped reads aligned to more than one genome. However, after reassigning the reads, 92.7% of the read probability was assigned to the *F. tularensis* WY96-3418 strain, and 4.8% of the read mass was assigned to the *F. tularensis* SCHU S4 strain, both of the *F. t. tularensis* subspecies. It is interesting to note that there are two strong indicators providing evidence that the identified genome is not the true source, but just a closely related substrain. The first indicator is that only a small proportion of the reads (19.8%) mapped to any genome in this example. In addition, after reallocation, the read probabilities assigned is less than what was observed in the eleven cases when the true genome was contained in the reference database. Therefore, these two quantities provide promising metrics for identifying whether the true genome is contained in the reference database.

Combination of multiple SRA datasets

We also generated a mixed read dataset by combining reads originating from *Y. pestis* KIM D27 (SRR033501), *E. coli* K-12 MG1655 (SRR031601) and *F. tularensis* subsp. *holarctica* OSU18 (SRR032505). After alignment to our genome database, 462,996 of the reads aligned to at least one genome in the database with 67.8%, 31.0%, and 1.2% originating from the *Y. pestis*, *E. coli*, and *F. tularensis*, respectively. Using a naïve mapping strategy, only 4.7% of the read probability was assigned to the correct *Y. pestis* strain, 4.4% matched the *E. coli* strain, and the *F. tularensis* strain received only 0.2% of the reads. In fact, the *F. tularensis* strain received fewer reads than 49 (of 131) genomes in the database. This clearly shows the failure of a naïve mapping strategy on mixed samples. Once the reads were reassigned using *Pathoscope*, 67.7%, 31.0% and 1.2%, of the read probability was assigned to the correct *Y. pestis*, *E. coli*, and *F. tularensis* strains, respectively. Thus, *Pathoscope* was able to recover genome

proportions almost identical to the original mixing proportions and the results were substantially better than the naïve approach.

To further evaluate *Pathoscope* on mixed samples, we generated 1,000 mixtures of ~5,770 reads (based on the size of the smaller *F. tularensis* dataset) with random proportions of each species. The naïve approach produced extremely biased results by consistently underestimating the correct read proportions, whereas *Pathoscope* closely estimated the read proportions with average absolute differences of 0.0008 for *Y. pestis*, 0.0092 for *E. coli*, and 0.0038 for *F. tularensis* (Table 3). In addition, the naïve approach consistently ranked genomes in the sample lower than many genomes that were not in the sample. For example, the average rank of *Y. pestis* across the 1,000 simulations was 13.1 for the naïve approach, and for 627 samples *Y. pestis* was not ranked among the top 10 genomes. Alternatively, after *Pathoscope*, *Y. pestis* was ranked among the top 3 (there were 3 genomes in the mixture) in all but 4 of the mixtures and in the top 5 for all of the mixtures. In these simulations, *Pathoscope* did fail to rank the proper *E. coli* substrain in the top 3 for 67 of the samples, in which cases *Pathoscope* either selected a different *E. coli* K12 substrain, or split the reads among the three K12 substrains in the database. For these 67 samples, we observed that the average number of *E. coli* reads was ~700, representing approximately 2.5% (0.025×) coverage of the *E. coli* genome. This points out that 2.5% coverage is not sufficient for *Pathoscope* to distinguish between substrains with 99.9% sequence identity (Hayashi et al. 2006), although *Pathoscope* did perform well in distinguishing between these substrains when coverage percentages ranged from 5% to 20%.

Discussion

Here we present an accurate and sophisticated computational approach for species identification and strain attribution. Our approach relies on the construction of a genome

database containing multiple strains or species that are possible source genomes for the sample and utilizes a probabilistic mapping approach to align the reads to the genome. Reads that map to multiple genomes are then reassigned to the most likely source genome using a Bayesian statistical framework that accommodates information on sequence quality and mapping quality. We attribute the increased accuracy of *Pathoscope* compared to other methods to the fact that *Pathoscope* considers all the reads jointly when reassigning reads to source genomes, whereas most other approaches only look at one read at a time. We show in multiple real data examples that our method is highly accurate in identifying the source genome or genomes for a biological sample. We show that in many cases, we can identify the source species or strain with only a small number of reads that represent only fractional coverage of the genome. In addition, we show that our approach is able to accurately identify the proper origin genome, even when several closely related strains or substrains are present within the database. We also show the failure of other approaches to assign reads and identify source genomes at the species, strain and substrain level.

We demonstrate the performance of *Pathoscope* on purified samples and for 'environmental' samples mixed *in silico*. In theory, this approach can also be applied to a to a variety of other scenarios including host-dominated clinical samples, unpurified environmental samples, and other types of community sequencing data. However, the performance and utility of *Pathoscope* in these contexts are yet to be determined. However, we believe that our approach will play an important role in future applications in pathology, bioforensics, and biosurveillance.

Methods

Genome database construction

Central to our approach is a robust database against which to map the query sequencing reads. For the purposes of this demonstration, we gathered a database of 170 complete bacterial chromosomes obtained from 131 distinct strains (610 Mbp) (See the Supplement for accession numbers for the genomes included in this reference database). The database was intended to aid in the identification of eight bacterial agents of bioterrorism identified by the CDC: *Bacillus anthracis*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Brucella* sp., *Clostridium botulinum*, *Escherichia coli* O157:H7, *Francisella tularensis*, and *Yersinia pestis*.

In order to differentiate closely related strains and species (often non-pathogenic) from target strains of interest, we wanted to include in our reference database genomes from any closely related strains/species. Therefore, closely related species/strains were identified by phylogenetic analysis of the 16S ribosomal RNA genes. 16S sequences for all eight pathogens of interest were obtained from GenBank and used to query the nr database using BLASTN (Altschul et al. 1997) using default parameters (Word Size = 28, Expect Value = 10, Match/Mismatch Scores = 1, -2, Gap Costs = Linear). We identified 3,206 sequences corresponding to 1,050 named species or subspecies with multiple sequences represented within a number of these taxonomic groups using a partial or full match with BLASTN. We then estimated phylogenetic relationships amongst these sequences and our target species. From this phylogeny, we selected 131 completed genome sequences, 332 fully sequenced plasmids, and 207 whole genome shotgun sequencing projects to serve as our reference database (see the Supplementary Materials for details). Although this study uses the entire genome database, any subset of these sequence types could be used for reference material. The genetic distances for Figure 1 were calculated by performing an all-against-all BLAST as implemented previously (Agren et al. 2012). Strains of the same species that were more than 98% similar using this metric were considered ‘closely related’ strains for Figure 1.

Probabilistic Alignment

We used the FastQC pipeline (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the quality of the read datasets. For the datasets used in these examples, the qualities of the datasets were generally acceptable by the standards set for the FASTQC pipeline. Once the quality of the dataset was ascertained, we aligned the reads utilizing a modification of BLAST (Altschul et al. 1997) alignment scores and the GNUMAP probabilistic alignment algorithm (Clement et al. 2010) because of their abilities to compute the likelihood that a read will be mapped to multiple locations in a reference database. In these approaches, segments of size *k* bases (*k*-mers) from the reads are indexed into a genomic hash table that contains all *k*-mers and their location in the reference genomes from the database. Once a set of putative locations is identified using the *k*-mer hash, the reads are then aligned to the full genomic sequences at these locations using a seed extension (Altschul et al. 1997) or a probabilistic Needleman-Wunsch algorithm (Clement et al. 2010). The latter approach incorporates the base quality information provided for each nucleotide, allowing GNUMAP to rely more on high-quality base calls and less on bases which are less certain. This approach improves mapping results from reads with bases that may be low quality or miscalled by the sequencer and reduces the chance that the reads will be misaligned to an incorrect genome. After alignment, the scores for each alignment location for both alignment algorithms (BLAST and GNUMAP) are then converted to posterior probabilities. Given the alignment scores $S_1, S_2, \ldots, S_n$ the posterior probability assigned to the *j*th alignment, P_j, is computed as

$$P_j = \frac{\exp\{S_j\}}{\sum_{k=1}^{n} \exp\{S_k\}}.$$

These posterior probabilities are interpreted as the probability that each alignment is the true source of the read. A probabilistic aligner performs significantly better than other alignment algorithms that discard non-unique reads or randomly assign non-unique reads to a single

genome (Li et al. 2008; Langmead et al. 2009; Li and Durbin 2009; Li et al. 2009). Alignment algorithms that discard non-unique reads lose much of the power to identify the correct genome. Unique reads often represent only a small fraction of the available reads and mapping algorithms that discard reads that occur in multiple locations are not able to discriminate between possible pathogens based on this limited amount of data. Randomly assigning the non-unique reads creates similar problems because it does not allow for read reassignment, leading to many reads being attributed to incorrect genomes.

Bayesian Reassignment Method

The Bayesian mixture model of *Pathoscope* assumes that reads are drawn from a small subset of unknown size from the pathogen genomes in the database. It assumes that each read is drawn from only one of the genomes in the subset. Parameters in the model represent the proportions of reads that originate from each genome as well as the proportion of the non-unique reads that are incorrectly assigned to each genome due to sequence similarity. Our Bayesian missing data mixture model re-weights the read assignment probabilities using the mapping qualities and the parameters of the model. In practice, in the reassignment process the parameters are designed to penalize the value of non-unique reads in the presence of unique reads and re-weight the non-unique reads based on overall mapping proportions when no reads map uniquely.

To formally describe our model, let $i=1,...,R$ be the index of the reads and let $j=1,...,G$ be the index of the genomes in the database. Let $\mathbf{x_i}=(x_{i1},x_{i2},...,x_{iG})=\{x_{ij}\}$ be a set of genome indicators for read i where $x_{ij}=1$ if the read originated from the jth genome and $x_{ij}=0$ if the read did not come from genome j. Note that by assumption one and only one element in the vector $\mathbf{x_i}$ can be equal to 1 (i.e. each read has only one template genome). We assume that $\mathbf{x_i}$ follows a

multinomial distribution, with probability of success $\boldsymbol{\pi}=(\pi_1,\pi_2,\ldots,\pi_G)=\{\pi_j\}$ where π_j is the proportion of the reads that originated from the *j*th genome.

For the *unique reads*, we know the template genome of interest or, in other words, we directly observe the genome indicator $\mathbf{x_i}$ for these reads. In the case of the *non-unique* reads, the genome indicator $\mathbf{x_i}$ is unobserved or *missing data*. For the non-unique reads, the observations are partial mapping qualities for each of the genomes. These mapping probabilities are provided as posterior probabilities, which are scaled mapping qualities or relative likelihood alignment scores obtained from the algorithm. More specifically, for the *i*th read we denote these mapping scores by $\mathbf{q_i}=(q_{i1},q_{i2},\ldots,q_{iG})=\{q_{ij}\}$. For unique reads, the q_{ij} values are equal to the x_{ij} values. For non-unique reads, these represent the uncertainty in mapping and need to be rescaled—or equivalently these reads need to be reassigned to the correct template genome of origin. In order to do this, we define a second set of parameters, $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_G)=\{\theta_j\}$ where θ_j is a reassignment parameter that represents the proportion of the non-unique reads that need to be reassigned to the *j*th genome.

In order to simplify the notation in the likelihood function, we define y_i as the *uniqueness indicator* for read *i*, namely letting y_i=1 if read *i* is unique and y_i=0 otherwise. Under the modeling assumptions above, the complete data likelihood of the parameters ($\boldsymbol{\pi}$, $\boldsymbol{\theta}$) given the observed data (reads, y_i, unique $\mathbf{x_i}$) and the missing data (non-unique $\mathbf{x_i}$) is given by:

$$\mathrm{L}(\boldsymbol{\pi},\boldsymbol{\theta}|\mathbf{x_i},\mathbf{q_i},\mathbf{y}) \propto \prod_{i=1}^{R}\prod_{j=1}^{G}[\pi_j\theta_j^{(1-y_i)}q_{ij}]^{x_{ij}}.$$

Although the reassigned reads (estimated $\mathbf{x_i}$) and reassignment parameters (estimated $\boldsymbol{\theta}$) are very informative, the quantities of interest from the modeling steps are the estimates for the genome read proportions (estimated $\boldsymbol{\pi}$). These probabilities will identify the single or multiple

organisms from the database that are present in the samples, based on the proportion of the reads that are assigned to the genome after the reads are reassigned.

Bayesian Prior Distributions

We assume a priori that both **π** and **θ** follow Dirichlet distributions, the densities of which can be seen in the following equations:

$$f(\boldsymbol{\pi}|\mathbf{a}) \propto \prod_{j=1}^{G} \pi_j^{a_j-1} \text{ and } f(\boldsymbol{\theta}|\mathbf{b}) \propto \prod_{j=1}^{G} \theta_j^{b_j-1}.$$

If $a_j=1$ for all $j=1,...,G,$ this is equivalent to adding one unique read for each of the G genomes, and $a_j=n$ would be the equivalent of adding n unique reads to the jth genome. Similarly, $b_j=n$ is the equivalent of adding n reads of non-unique read probabilities to the jth genome. However, the prior information for **θ** does not behave like true non-unique reads because it is not subject to reassignment. Prior information assigned to each genome will always be associated with that genome, but its effect is diminished as the number of reads increases. This can be seen clearly in the maximization formulas given in the following section. The prior information stabilizes the algorithm by preventing the estimates of **π** and **θ** from converging to the boundaries of 0 and 1. Inclusion of prior information will bias the results, possibly even leading to the identification of the wrong genome if the prior is not selected carefully. However, this would only happen in rare circumstances, and it would require initially favoring some genomes above others. To avoid this, each genome will usually receive the same values for its priors for **a** and **b**. If prior information is included, evidence of a read being present in the sample can be inferred based on whether or not the final read probability is statistically greater than the original prior information inserted. Note that non-informative priors can also be used for the experimental data by assigning zero unique reads and zero non-unique reads to each genome.

Read Reassignment via the EM algorithm

Estimation of the model parameters and reassignment of the reads is accomplished using an Expectation-Maximization (EM) algorithm (Dempster 1977). Each of iterations of the EM algorithm consists of two simple steps. The first, called the expectation step or E-step, reassigns each read to its most likely or *expected* genome based on its mapping quality score and current estimates of the read proportion and non-unique misclassification model parameters. In the second step, called the maximization step or M-step, the model parameters are re-estimated using the new read assignment probabilities from the most recent E-step. These steps are repeated until the read assignments and proportion estimates converge to stable values between iterations. This algorithm is guaranteed to converge to a local maximum (Hastie et al. 2009), and in our data examples presented above, the parameter's posterior distributions appeared to be unimodal and therefore the EM converged to a global maximum without issue.

To implement this algorithm, initial estimates of the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are proposed, usually $\pi_j=\theta_j=1/G$ for all j. In the E-step, the expected value of $\mathbf{x_i}$ is computed for each combination of $i=1,...,R$ and $j=1,...,G$ based estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, as well as the observed data $\mathbf{q_i}$ and $\mathbf{y}$. In the E-step, the expected values of the elements of $\mathbf{x_i}$ are estimated as:

$$\hat{\delta}_{ij} = \mathrm{E}(x_{ij}) = \frac{\pi_j \theta_j^{(1-y_i)} q_{ij}}{\sum_{k=1}^{G} \pi_k \theta_k^{(1-y_i)} q_{ik}}.$$

Next, the M-step calculates the new estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ given $\mathbf{q_i}$, $\mathbf{y}$ and the current expected values $\hat{\delta}_{ij}$. The formulas for estimating $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, provide the Bayesian maximum a posteriori (MAP) estimates; however, if the prior information a_j and b_j are set to 0 for all j genomes, these equations provide the maximum likelihood estimates. Letting $\mathrm{N} = \sum_{k=1}^{G} \sum_{i=1}^{R} \hat{\delta}_{ik}$, these estimates are as follows:

$$\hat{\pi}_j = \frac{\sum_{i=1}^{R} \hat{\delta}_{ij} + a_j}{N + \sum_{k=1}^{G} a_k} \text{ and } \hat{\theta}_j = \frac{\sum_{i=1}^{R} (1-y_i)\hat{\delta}_{ij} + b_j}{\sum_{i=1}^{R} (1-y_i) + \sum_{k=1}^{G} b_k}.$$

The E-step is then repeated using the updated estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ followed again by the M-step. These steps are repeated until the expected value of $\mathbf{x}_i$ and the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ converge to stable values across iterations.

Parameters used in Methods Comparisons

For the O104:H4 example comparisons and SRA datasets, we used the following parameters:

Naïve alignment: A) BLAST version 2.2.27+ at default parameters, B) Bowtie 2 version 2.0.2 at with: "-k 100", and C) GNUMAP version 3.0.2 with: "-m 16 -h 500 -a 0.8 --print_all_sam".

MEGAN: Version 4.70.4 with default parameters and the BLAST output file (described above).

PhymmBL: Version 4.0 at default parameters after manually adding O104:H4 to the database.

MetaPhlAn: Version 1.7.7 at default parameters and the BLAST marker database.

PhyloPhythiaS: Webserver: http://phylopythias.cs.uni-duesseldorf.de; Submission date: 3/2013.

Trinity Assembler: Release 2012-06-08 with "--seqType fa --JM 10G --single".

Acknowledgements

We gratefully acknowledge allocation of supercomputer resources from the Fulton Supercomputing Lab at Brigham Young University and the Linux Clusters for Genetic Analysis (LinGA) at the Boston University Medical Campus. This work was partially supported by funds from the National Institutes of Health (R01HG00569).

Author Contributions

KAC and WEJ conceived the study. WEJ, QS, MJC, and GBS developed the algorithm and provided direction for its implementation, and OEF, SM, CH, NLC, and WEJ implemented the algorithm and developed the software. MB, ECN, and KAC developed the target genome

database and calculated genome similarity. OEF, SM, and CH applied the software to the sequencing datasets. OEF, WEJ, and KAC wrote the paper.

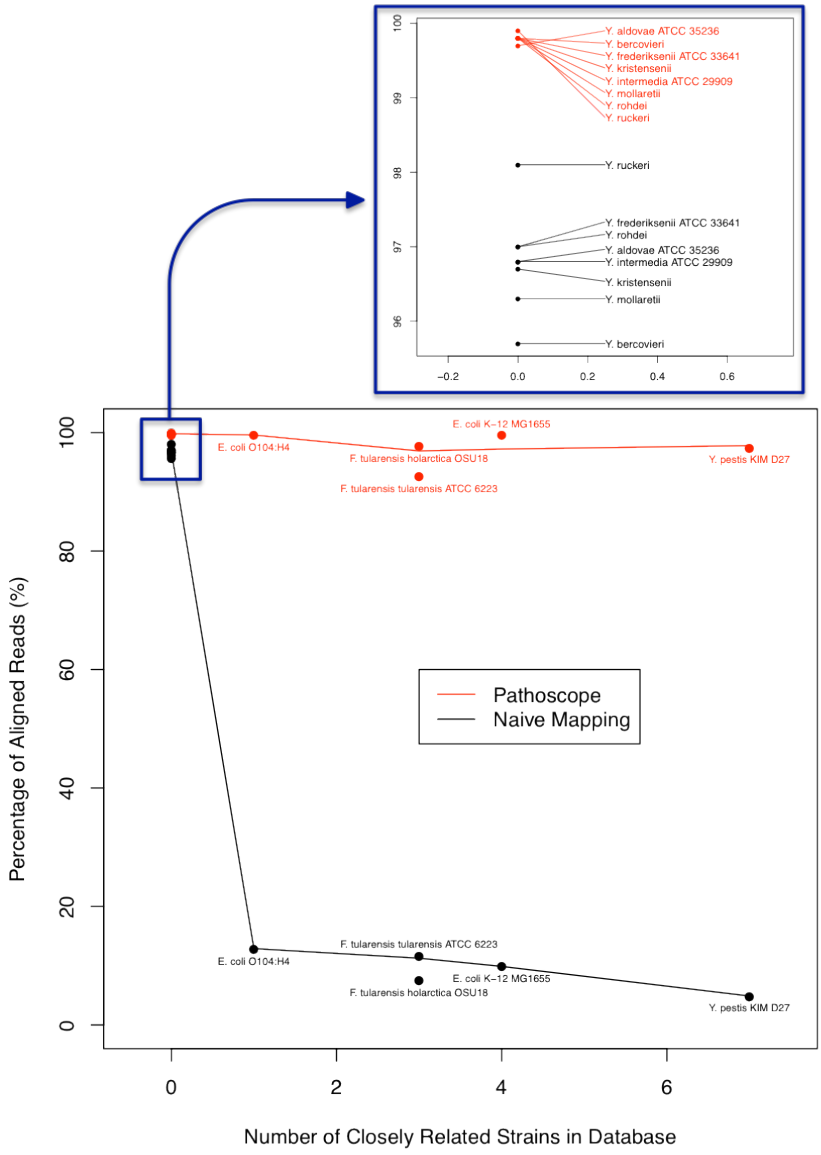
Disclosure Declaration

The authors declare no conflicts of interest.

Figure Legends

Figure 1: Impact of the closely related strains on the read alignment proportions. The genomes in the database were aligned to each other using an all-against-all BLASTN approach (Agren et al. 2012), and strains of the same species that were more than 98% similar using this metric were considered ‘closely related’ strains. As the number of closely related strains increases, the naïve algorithm was not able to definitively identify the origin species. However, *Pathoscope* performed consistently well independent of the number of closely related strains.

Figures



Tables

	% of Reads to Correct Genome (Second Highest) [Range for 1,000 Random Samples]				Time Required (Full Dataset)
Number of Reads (Coverage)	**92 (0.001×)**	**924 (0.01×)**	**9,237 (0.13×)**	**92,370 (1.3×)**	
Naïve mapping	12.9 (6.5) [7.5-20.9]	12.9 (6.1) [10.5-15.5]	12.9 (7.4) [12.2-13.5]	12.9 (7.4)	BLAST: 38min Bowtie 2: 13min
Pathoscope	99.4 (0.5) [95.1-100.0]	99.6 (0.3) [98.0-100.0]	99.6 (0.3) [99.3-99.8]	99.6 (0.3)	Naïve + 7min
PhymmBL	14.7 (7.0) [4.3-26.1]	14.7 (7.0) [11.3-18.5]	14.7 (7.1) [13.6-15.7]	14.7 (7.1)	13hrs**
MetaPhlAn (species only)	--	36.1 (0.0) [0.0-100.0]	96.9 (2.4) [54.1-100.0]	90.0 (9.6)	1 min
Trinity Contigs	--	--	70.8 (22.6)	21.5 (13.4)	30 min
PhyloPythiaS*					7 min**
Family (or above)	47.8 (7.6)	48.4 (2.2)	45.6 (2.8)	--	
Genus	0.0 (2.2)	0.1 (1.6)	0.1 (1.2)	--	
Species	0.0 (0.1)	0.0 (0.2)	0.1 (0.3)	--	
MEGAN*					Naïve + 3min
Family (or above)	84.7 (0.0)	79.5 (0.0)	80.2 (0.0)	80.2 (0.0)	
Genus	16.3 (0.0)	20.5 (0.0)	19.6 (0.2)	19.7 (0.2)	
Species/Strain*	--	--	--	--	

*Source strain was not contained in annotation

** PhymmBL also required 36 hours of database preprocessing, and PhyloPhythiaS was only applied to 9,237 reads and not the whole dataset because of its webserver limitations.

Table 1: Results from the application of several species identification approaches to subsets of the 92,370 sequencing reads from the first O104:H4 Ion Torrent sequencing run. Presented here are the percentages of all reads assigned to the correct genome along with the second highest scoring genome in parenthesis. It is clear that *Pathoscope* is the most effective algorithm for strain identification. The asterisk (*) for MEGAN and PhyloPythiaS is to denote that the O104:H4 annotation is not available, so the nearest strain *E. coli* 55989 was considered the 'correct' strain.

Sequence Read Archive Accession	Source Genome	Total number of reads (percent mapped)	Reads mapped correctly (next best match)	Reassignment (next best match)
SRR031601	*E. coli* K-12 MG1655	143,836 (99.5%)	10.0% (10.0%)	99.6% (0.2%)
SRR032505	*F. tularensis* subsp. *holarctica* OSU18	28,221 (24.4%)	7.6% (6.9%)	97.8% (1.2%)
SRR032501	*Y. pestis* KIM D27	318,332 (15.1%)	4.8% (4.4%)	97.4% (1.7%)
SRR031600	*Y. aldovae* ATCC 35236	91,788 (75.3%)	96.8% (0.8%)	99.7% (0.2%)
SRR029367	*Y. bercovieri*	1,263,275 (73.9%)	95.7% (1.2%)	99.8% (0.1%)
SRR031602	*Y. frederiksenii* ATCC 33641	1,504,985 (76.0%)	97.0% (0.4%)	99.8% (0.1%)
SRR000311	*Y. kristensenii*	1,374,452 (85.7%)	96.7% (0.9%)	99.8% (0.2%)
SRR031268	*Y. intermedia* ATCC 29909	1,341,997 (79.6%)	96.8% (0.4%)	99.8% (0.1%)
SRR031599	*Y. mollaretii*	1,463,985 (77.0%)	96.3% (1.2%)	99.8% (0.1%)
SRR029323	*Y. rohdei*	199,435 (90.5%)	97.0% (0.5%)	99.8% (0.1%)
SRR000904	*Y. ruckeri*	299,829 (90.2%)	98.1% (0.2%)	99.9% (0.0%)
SRR031603	*F. tularensis* subsp. *tularensis* ATCC 6223*	67,276 (19.8%)	11.7% (8.4%)	92.7% (4.8%)

*Source strain was not contained in database

Table 2: Results from the application of our species identification method on 12 datasets from the NCBI Sequence Read Archive (SRA, reference website). These examples consisted of Roche 454 samples ranging from 1× to 31× coverage of the origin genomes. Notice that in many cases the naïve read mapping identifies the correct genome (based on a large majority of aligned reads assigned to the genome). However, for several examples, particularly the cases where there are closely related strains in the database, the correct genome cannot be clearly identified using only the read mapping. However, after application of our Bayesian reassignment algorithm in every case the reads are reallocated to the correct genome with increased and very high confidence.

	Naïve Mapping	*Pathoscope*
Average Absolute Difference		
Y. pestis KIM D27	0.3160	0.0008
E. coli K-12 MG1655	0.3073	0.0092
F. tularensis subsp. *holarctica* OSU18	0.2708	0.0038
Average Ranking (among 131 full genomes)		
Y. pestis KIM D27	13.1	2.0
E. coli K-12 MG1655	7.4	2.2
F. tularensis subsp. *holarctica* OSU18	4.4	2.0
Number of Times Not Ranked in Top 3 (not in Top 10)		
Y. pestis KIM D27	964 (627)	4 (0)
E. coli K-12 MG1655	613 (140)	67 (1)
F. tularensis subsp. *holarctica* OSU18	311 (79)	1 (0)

Table 3: Results from 1,000 random mixtures of approximately 5,770 reads from the *Y. pestis* KIM D27 (SRR033501), *E. coli* K-12 MG1655 (SRR031601) and *F. tularensis* subsp. *holarctica* OSU18 (SRR032505) datasets. The proportion estimates from the naïve approach were extremely biased, typically underestimating the true read proportion, while *Pathoscope* estimated the true proportions with high precision. In addition, the naïve approach consistently ranked genomes in the sample lower than many genomes that were not in the sample. *Pathoscope* did fail identify the *E. coli* substrain in some of the samples—in these cases, *Pathoscope* identified a nearly identical K12 substrain or split the reads between the three K12 substrains in the database.

References

Agren J, Sundstrom A, Hafstrom T, Segerman B. 2012. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one* **7**(6): e39107.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**(17): 3389-3402.

Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. 2012. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **28**(8): 1174-1175.

Brady A, Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods* **6**(9): 673-676.

Clement NL, Snell Q, Clement MJ, Hollenhorst PC, Purwar J, Graves BJ, Cairns BR, Johnson WE. 2010. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* **26**(1): 38-45.

Dempster APL, N.M.; Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* **39**(1): 1-38.

Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. 2011. Genomic anatomy of Escherichia coli O157:H7 outbreaks. *Proceedings of the National Academy of Sciences of the United States of America* **108**(50): 20142-20147.

Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A et al. 2011. Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. *The New England journal of medicine* **365**(19): 1771-1780.

Gerlach W, Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids research* **39**(14): e91.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**(7): 644-652.

Hastie T, Tibshirani R, Friedman JH. 2009. *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York, NY.

Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H et al. 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular systems biology* **2**: 2006 0007.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome research* **17**(3): 377-386.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature biotechnology* **29**(5): 393-396.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**(11): 1851-1858.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**(15): 1966-1967.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods* **4**(1): 63-72.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**: 386.

Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25**(14): 1722-1730.

Patil KR, Roune L, McHardy AC. 2012. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PloS one* **7**(6): e38581.

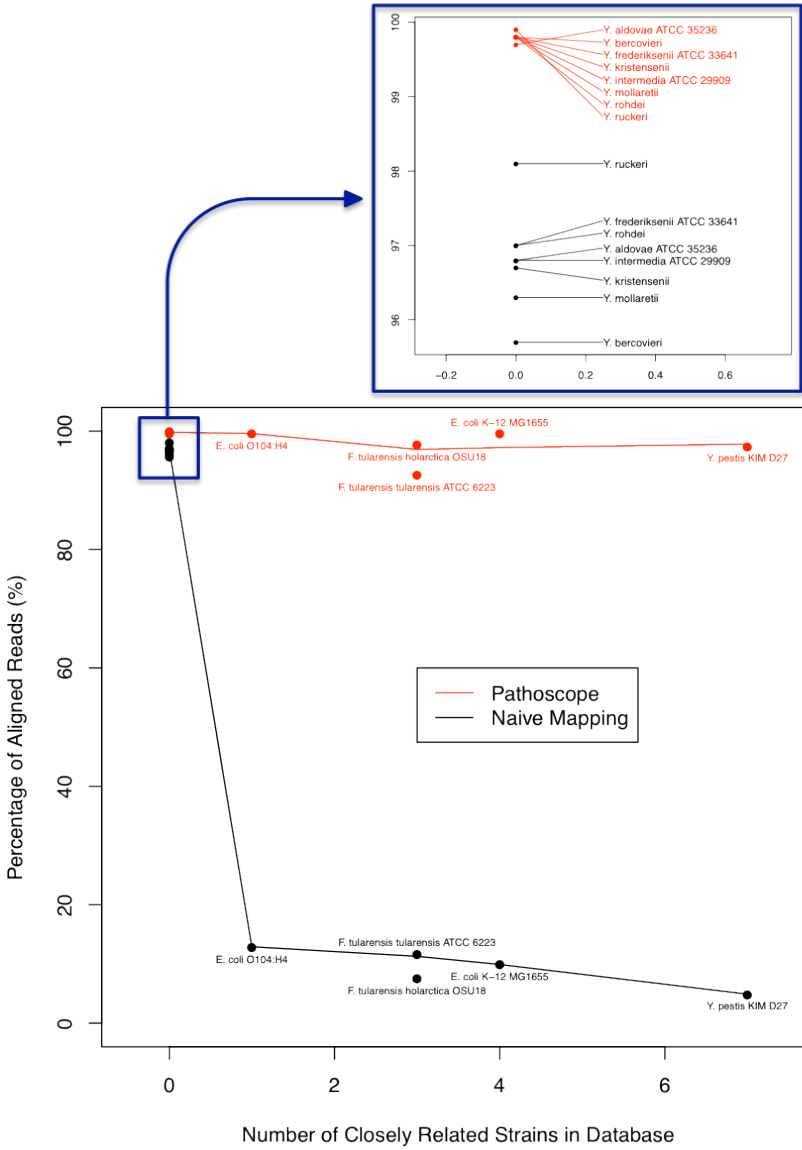
Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J et al. 2011. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *The New England journal of medicine* **365**(8): 718-724.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome research* **20**(9): 1165-1173.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9**(8): 811-814.

Turner M. 2011. Microbe outbreak panics Europe. *Nature* **474**(7350): 137.

Vaidyanathan G. 2011. Better biosurveillance could halt disease spread. *Nature News*.


Y. aldovae ATCC 35236
Y. bercovieri
Y. frederiksenii ATCC 33641
Y. kristensenii
Y. intermedia ATCC 29909
Y. mollaretii
Y. rohdei
Y. ruckeri
Y. ruckeri
Y. frederiksenii ATCC 33641
Y. rohdei
Y. aldovae ATCC 35236
Y. intermedia ATCC 29909
Y. kristensenii
Y. mollaretii
Y. bercovieri
100
99
98
97
96
−0.2
0.0
0.2
0.4
0.6
E. coli K−12 MG1655
E. coli O104:H4
F. tularensis holarctica OSU18
Y. pestis KIM D27
F. tularensis tularensis ATCC 6223
Pathoscope
Naive Mapping
F. tularensis tularensis ATCC 6223
E. coli O104:H4
E. coli K−12 MG1655
F. tularensis holarctica OSU18
Y. pestis KIM D27
Percentage of Aligned Reads (%)
0
20
40
60
80
100
0
2
4
6
Number of Closely Related Strains in Database