PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Removing contaminants from databases of draft genomes

Jennifer Lu[1,2]*, **Steven L. Salzberg**[1,2,3]

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States of America, **2** Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States of America, **3** Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD, United States of America

* jlu26@jhmi.edu, jennifer.lu717@gmail.com

Abstract

Metagenomic sequencing of patient samples is a very promising method for the diagnosis of human infections. Sequencing has the ability to capture all the DNA or RNA from pathogenic organisms in a human sample. However, complete and accurate characterization of the sequence, including identification of any pathogens, depends on the availability and quality of genomes for comparison. Thousands of genomes are now available, and as these numbers grow, the power of metagenomic sequencing for diagnosis should increase. However, recent studies have exposed the presence of contamination in published genomes, which when used for diagnosis increases the risk of falsely identifying the wrong pathogen. To address this problem, we have developed a bioinformatics system for eliminating contamination as well as low-complexity genomic sequences in the draft genomes of eukaryotic pathogens. We applied this software to identify and remove human, bacterial, archaeal, and viral sequences present in a comprehensive database of all sequenced eukaryotic pathogen genomes. We also removed low-complexity genomic sequences, another source of false positives. Using this pipeline, we have produced a database of "clean" eukaryotic pathogen genomes for use with bioinformatics classification and analysis tools. We demonstrate that when attempting to find eukaryotic pathogens in metagenomic samples, the new database provides better sensitivity than one using the original genomes while offering a dramatic reduction in false positives.

Author summary

Infectious diseases afflict a majority of the human population around the world, from the common cold to the devastating malaria parasite. As technology has evolved, DNA sequencing emerged as a revolutionary and rapid method for diagnosing human infections. As part of our efforts to boost the ability of scientists to identify the source of an infection by sequencing, we present here a computational method for removing erroneous or misleading sequences from existing DNA databases. When we applied this method to a database of more than 200 eukaryotic pathogens, we were able to successfully and accurately identify the true pathogens infecting real human samples.

OPEN ACCESS

Citation: Lu J, Salzberg SL (2018) Removing contaminants from databases of draft genomes. PLoS Comput Biol 14(6): e1006277. https://doi.org/10.1371/journal.pcbi.1006277

Editor: Fengzhu Sun, University of Southern California, UNITED STATES

Received: February 6, 2018

Accepted: June 6, 2018

Published: June 25, 2018

Copyright: © 2018 Lu, Salzberg. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data is publicly available at http://eupathdb.org/eupathdb/ with the resulting data provided at ccb.jhu.edu/data/eupathDB/.

Funding: This study was funded by the National Institutes of Health https://www.nih.gov/, R01 GM083873. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Next-generation sequencing in pathogen discovery/diagnosis

Next-generation sequencing (NGS) over the last few years has emerged as a valuable tool for human pathogen discovery and diagnosis. In the case of human pathogen detection, traditional histological, immunological, or molecular techniques are limited and often yield an incorrect or incomplete diagnosis [1]. As sequencing has grown faster and cheaper, clinicians have begun to explore the possibility of replacing older methods with NGS, which provides a fast, specific, and relatively unbiased method of capturing the full spectrum of macro- and microorganisms in any sample.

A growing number of case studies illustrate the potential for NGS in diagnosis. For example, in 2013 Loman et al. conducted a retrospective investigation into the 2011 German outbreak of Shiga-toxigenic *Escherichia coli* (STEC) [2]. In this study, sequencing led to rapid and accurate identification of the bacterial infection in fecal specimens of the infected patients. In 2014, Hasman et al. analyzed 35 urine samples from patients with suspected urinary tract infections, confirming cultured bacterial infections using sequencing of isolated and cultured bacteria [3]. They also successfully identified polymicrobial bacterial infections by directly sequencing the urine samples. Later in 2014, Wilson et al. used next-generation sequencing of cerebrospinal fluid (CSF) to identify and treat a bacterial *Leptospira* infection in a 14-year old patient [4]. In 2016, Salzberg et al. tested the possibilities of detecting pathogens by sequencing brain or spinal cord biopsies from 10 patients presenting with neurologic symptoms with previously unidentified infections [5]. In that study, NGS identified both bacterial and viral infections in selected patients, diagnoses that were confirmed by traditional immunologic techniques.

Pathogen discovery bioinformatics pipelines. A critical step in using NGS for diagnosis is in the bioinformatics analysis of the millions (or billions) of genomic reads that result from a sequencing experiment. The identification of the sequenced DNA provides the information about the potential pathogenic organisms causing the infection. Because the source of the sample is human tissue, all the studies mentioned above first filtered out human DNA, which is uninformative for pathogen discovery [2–5]. Following this step, the remaining sequencing reads are compared to reference genomic databases, such as RefSeq or the NCBI nt database, using a variety of alignment and classification tools, including BLAST, Bowtie2, MetaPhlAn, and Kraken [6–9].

Challenges in relying on reference databases

Although databases of sequenced pathogens have grown dramatically larger over the past decade, the dependence on reference databases still presents challenges when used for diagnosis, for at least two reasons: (1) no database contains the full spectrum of all potential human pathogens, and (2) existing reference databases have been found to contain contamination.

Over the past two decades, microbial genome projects have predominantly focused on bacteria and viruses. The GenBank repository [10] contains the majority of genome sequence data submitted by laboratories around the world. As of January 2018, GenBank contained genome entries representing over 54,000 bacterial organisms but only 1,791 fungi and 389 protozoa. The NCBI RefSeq project analyzes and filters the Genbank genome sequences to create a more curated database, which is also widely used [11]. This database also reflects the focus on bacterial and viral genomes, with more than 37,000 bacterial organisms and more than 7,500 viral

Table 1. Organisms in Genbank and RefSeq as of January 2018. Total genome counts are based on summaries found at ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/.

	Draft and Complete genomes		Complete genomes	
	Genbank	**RefSeq**	**Genbank**	**RefSeq**
Bacterial	54,153	37,399	5,372	5,121
Viral	10,412	7,509	10,339	7,484
Archaea	1,861	533	272	235
Fungi	1,791	251	26	8
Protozoa	389	82	3	2
Vertebrates	376	238	71[a]	55[a]
Plants	320	102	3	3

[a] No complete vertebrate genome exists. The number shown here is the number of organisms with chromosome-level assemblies.

https://doi.org/10.1371/journal.pcbi.1006277.t001

organisms represented. In contrast, RefSeq contains genomes for only 251 fungi and 82 protozoa (**Table 1**).

The composition of the reference databases is not representative of the species composition of the natural world, but rather reflects a focus on human pathogens, other species of interest to humans, and the challenges of isolating and sequencing DNA from various species [12]. In many cases, microorganisms are difficult to isolate from their surrounding environments, living among thousands of other species in complex ecosystems [13, 14]. Some microorganisms live in extreme conditions and have gone undiscovered until recently [15]. Other microorganisms are difficult to grow in culture to provide sufficient DNA from which to derive a reference genome. As a result of these constraints, most early research into microorganisms focused on a few easily culturable bacteria [16]. However, studies over the last two decades suggest that culturable bacteria represent only a small fraction of the microorganisms on earth [12, 16–18].

Eukaryotic pathogens comprise an underrepresented group of microorganisms in genomic databases, although they are critically important for the diagnosis of human infections. This group includes a diverse group of species that infect multiple areas in the body; e.g., apicomplexans such as *Plasmodium falciparum*, which causes most cases of human malaria [19], and *Toxoplasma gondii* [20], which may cause neurological defects. Other examples include multiple fungal species belonging to the *Fusarium*, *Aspergillus*, *Curvularia*, and *Candida* genera, and amoebae species belonging to the *Acanthamoeba* genus, the latter of which causes a majority of human corneal infections [21, 22]. These are only a small sample of the hundreds of known eukaryotic pathogens of humans.

EuPathDB is a database representing more than 250 eukaryotic microorganisms [23], including both known pathogens and other closely related non-infectious eukaryotic species. Because no eukaryotic pathogen has yet been completely sequenced, this resource comprises primarily draft genomes at varying degrees of completeness, some of which have had little curation since their initial sequencing. However, EuPathDB is more comprehensive than the RefSeq database, containing more than 150 genomes that are absent from the RefSeq protozoa and fungi databases (see **Table 2**).

In recent years, multiple studies revealed contamination in the public genome sequences of many organisms, particularly for draft genomes. In 2011, Longo et al. identified 492 non-primate public databases from NCBI, UCSC, and Ensembl containing human genome sequences [24]. A 2014 study found that portions of the complete genome for *Neisseria gonorrhoeae* TCDC-NG08107 belonged to the cow and sheep genomes [25]. Another study in 2015 identified over 18,000 microbial isolate genome sequences that were contaminated with PhiX174, a bacteriophage used as a control in Illumina sequencing runs [26]. 10% of those 18,000 genomes

Table 2. EuPathDB genome representation in RefSeq. This table shows the number of genomes from the eukaryotic pathogen database that also exist in the Genbank and/or RefSeq databases along with the breakdown of their assembly status within those databases.

Assembly Status	RefSeq
Complete Genome	3
Chromosome	41
Scaffold	46
Contig	5
Not Represented	*150*
Total	**245**

https://doi.org/10.1371/journal.pcbi.1006277.t002

were published in the literature. In 2016, Kryukov et al. identified 154 non-human genome assemblies containing human sequence fragments that were at least 100bp long [27]. As one example, they discovered that more than 330,000 bp in the reference genome of *Plasmodium gaboni*, a relative of *Plasmodium falciparum*, appears to be contaminating human sequence.

Contamination and incompleteness in reference databases causes bioinformatics analysis of sequencing reads to yield both false positive and false negative results, thereby decreasing the overall reliability of NGS in pathogen diagnostics. False positives, where the wrong pathogen is identified, might in turn lead to inaccurate treatments, with the potential to harm rather than help patients.

In this study, we present a new method for eliminating genomic contamination that can be used on both complete and draft reference genomes. We test our method on a large set of eukaryotic pathogen genomes, yielding a cleaned and filtered eukaryotic pathogen database ready for use in bioinformatics pipelines, including those intended for NGS diagnostics, with decreased false positive and false negative rates.

Methods

The eukaryotic pathogen genomes underwent a multi-step cleaning process to remove both contaminating and non-informative sequences (see Fig 1). Each genome was first split into 100bp overlapping pseudo-reads, with each pseudo-read beginning every 50bp along the genome. The pseudo-reads were then compared to three unique databases, using the Kraken [7] and Bowtie2 [8] classification and alignment programs.

Kraken labels reads only if they contain an exact 31 base-pair (31-mer) match to any 31-mer in the database sequences [7]. For this process, pseudo-reads were classified with Kraken against two unique Kraken databases. The first Kraken database contains 15,000 genomic sequences from the human, human CHM1, mouse, bacteria, archaea, viral, and plant RefSeq databases as of November 30th, 2017. We also included contaminating sequences such as the UniVec database, EmVec database, and phiX174 vector in the first Kraken database. The second Kraken database contains all complete and chromosomal-level assemblies of non-human and non-mouse vertebrate sequences (representing 56 vertebrate species). Kraken requires that the selected database is first loaded into RAM prior to classification. We used two databases in order to reduce RAM usage at a single time, allowing sequential classification of the pseudo-reads to each database. **S1 Table** lists the accession numbers, taxonomy IDs, and organisms for all genomic sequences included in both Kraken databases used for masking.

Bowtie2 aligns sequencing reads against any reference sequence, allowing for gaps or mismatches [8]. We created a bowtie2 index of GRCh38.p11 and aligned the pseudo-reads against it. Note that even though we include GRCh38.p11 in the Kraken database, which enables Kraken to find human reads, Bowtie2's more sensitive alignment algorithm can align some sequences that Kraken will miss.

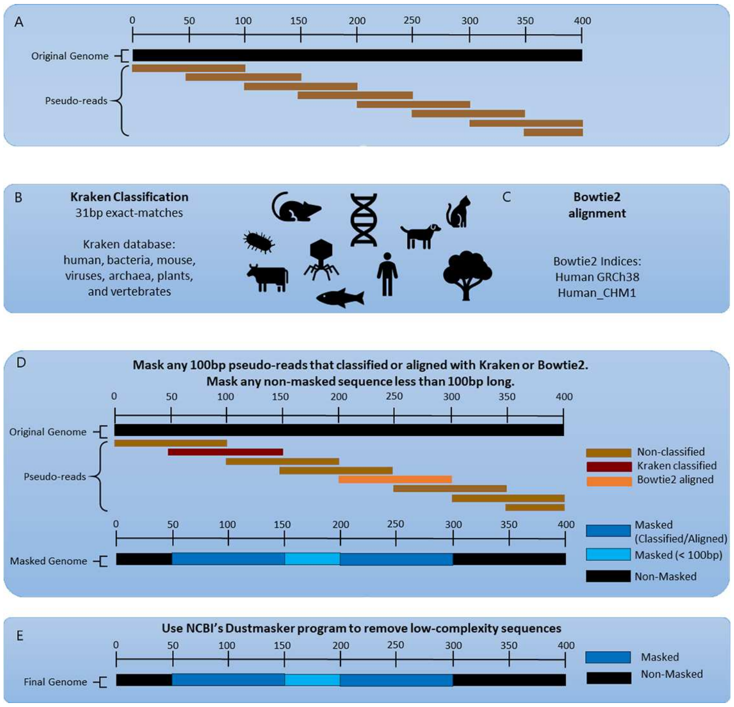

Fig 1. Masking procedure. A) The original genome is split into 100bp overlapping pseudo-reads. B) The pseudo-reads are then classified using Kraken first against the common contaminating vector sequences and the plant, viral, bacterial, archaeal, human, and mouse RefSeq database. The pseudo-reads are also classified using Kraken against non-human and non-mouse vertebrate RefSeq genomes. C) Bowtie2 is then used to align all pseudo-reads against the human genome. D) All pseudo-reads that were classified in the previous steps are masked out of the original genomes. Any remaining non-masked sequence with less than 100p is also masked. E) Finally, Dustmasker is used to mask additional low-complexity sequences.

https://doi.org/10.1371/journal.pcbi.1006277.g001

Any pseudo-read that was classified in these steps represents either a contaminating sequence in the pathogen genome or a low-complexity sequence that matches a distant species only by chance. In either case, these sequences could lead to false positive identifications if they are used for metagenomics analysis. Therefore, we masked any portion of a database genome that corresponded to pseudo-read that was classified or aligned in the previous steps. (Masking can be done in a variety of ways; we simply replaced the sequence with Ns to keep the overall genome length the same.) If, after this initial masking step, we created non-masked sequences that were <100 bp in length, we masked those sequences as well. We then used Dustmasker [28] to mask additional low-complexity sequences (Fig 1).

Results and discussion

We tested our method for eliminating contamination on the draft genomes contained in EuPathDB release 28 [23], which contains 245 genomes categorized into the following sub-databases: AmoebaDB (29 genomes), CryptoDB (11), FungiDB (87), GiardiaDB (6), Microsporidia DB (25), PiroplasmaDB (8), PlasmoDB (9), ToxoDB (30), TrichDB (1), and TriTrypDB (39). **S2 Table** lists all genomes included in EuPathDB, detailing each genome's filename, sub-database category, genus, species, and full scientific name. **Fig 2** shows how much of each of the 245 genomes was masked in each step of the cleaning procedure and the final lengths of the cleaned pathogen genomes. Full details of the amount of masked sequence for all genomes are listed in **S2 Table**.

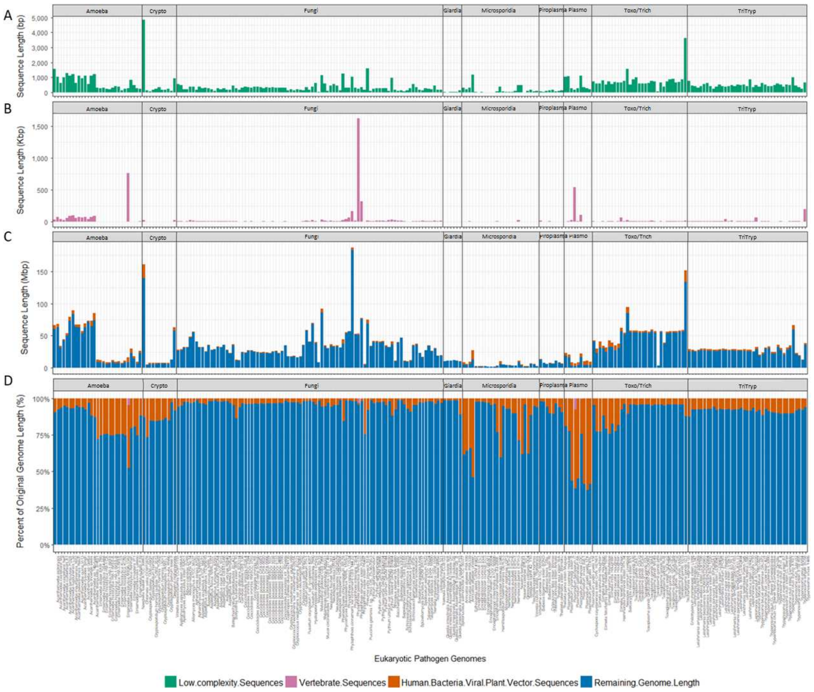

Fig 2. Masking results. Fig 2C provides an overview of sequence lengths for each eukaryotic pathogen genome masked in each step and the sequence lengths of the final cleaned genomes. As low-complexity sequences and vertebrate masked sequences are much smaller compared to the final genome length or human/bacterial/viral/plant/vector sequences, these were additionally plotted in Fig 2A and 2B for each eukaryotic pathogen genome. Low-complexity sequences were masked as a final step as well. Masked sequence lengths are also presented as percentages of the original genome length to show the percent of each genome remaining and the percent masked in each step (Fig 2D). Exact numbers are listed in **S2 Table**.

https://doi.org/10.1371/journal.pcbi.1006277.g002

Genome lengths in EuPathDB ranged from 2Mbp to 186Mbp prior to our cleaning procedure. Post-cleaning genome lengths ranged from 1.7Mbp to 182Mbp, with an average of 11% of each genome identified as contaminating or low-complexity sequences. As Fig 2 illustrates, a few genomes were outliers with over 50% of the genome being masked, but most genomes lost <10% of their length through this process.

In the first masking step, pseudo-reads across all EuPathDB genomes are classified against two Kraken databases containing bacterial, archaeal, viral, human, mouse, vertebrate, and contaminating vector genomes (Fig 1). These classification counts are listed in **S3 Table.** Reads classified as vertebrates are further broken down into sub-classifications such as fish or bird species. **Fig 3** shows the breakdown of these classifications for the 20 pathogen genomes with the largest numbers of classified pseudo-reads. **Fig 4** shows a similar breakdown focusing specifically on the 20 genomes with the most pseudo-reads labelled as mouse or human.

Most genome masking occurred after the first Kraken screen against the database of bacterial, archaeal, viral, human, mouse, and vector genomes. As a result of this step, we masked on average ~10% of each of the EuPathDB genomes. After classifying the remaining pseudo-reads against the vertebrate database, we masked a much smaller amount of sequence, with only 0.1% of each genome matching vertebrate sequences in this step.

The most contaminated eukaryotic pathogen genomes are the three *Plasmodium yoelii* genomes (strains 17XNL, YM, and 17X), with approximately 60% of the genomes identified as human/bacterial/viral/archaeal (Figs 3 and 4). The primary sources of contamination in these three genomes were *Methylococcus capsulatus* (16,000 pseudo-reads) and the mouse genome

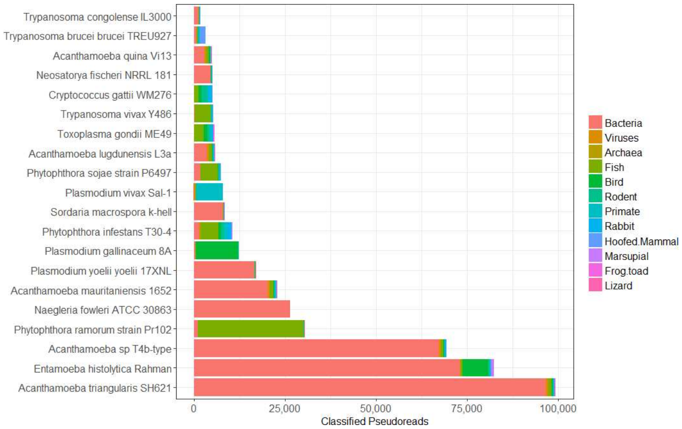

Fig 3. Pseudo-read Kraken classifications. The above plot shows the 20 eukaryotic pathogen genomes with the greatest numbers of pseudo-reads that Kraken identified as matching foreign species when searching against database containing bacteria, viruses, archaea, and a limited set of vertebrate genomes. Vertebrate classifications are grouped by common categories, such as fish, birds, rodents, or primates. Primate and rodent numbers do not include human and mouse, which are counted and shown separately. S3 Table contains pseudo-read classifications for all eukaryotic pathogen genomes.

https://doi.org/10.1371/journal.pcbi.1006277.g003

(12,000 pseudo-reads). The genome for *Plasmodium vivax Sal-1*, which causes malaria in humans, contained the greatest amount of human contamination, with more than 4,000 pseudo-reads classified as *Homo sapiens*. *Entamoeba histolytica Rahman*, a human intestinal parasite, is also notably contaminated, with nearly 50% of its genome identified as either human or bacteria (Figs 3 and 4)

Other eukaryotic pathogens that underwent significant masking due to contamination include *Plasmodium gallinaceum 8A* (62% masked), *Plasmodium falciparum IT* (57% masked), *Plasmodium reichenowi CDC* (55% masked). Each of these pathogens contained significant contamination likely due to host DNA, as the masked pseudo-reads were identified as matching their original host. For example, *Plasmodium gallinaceum* causes malaria in poultry and 11,700 pseudo-reads were identified as chicken DNA (see **S3 Table**) [29]. Although

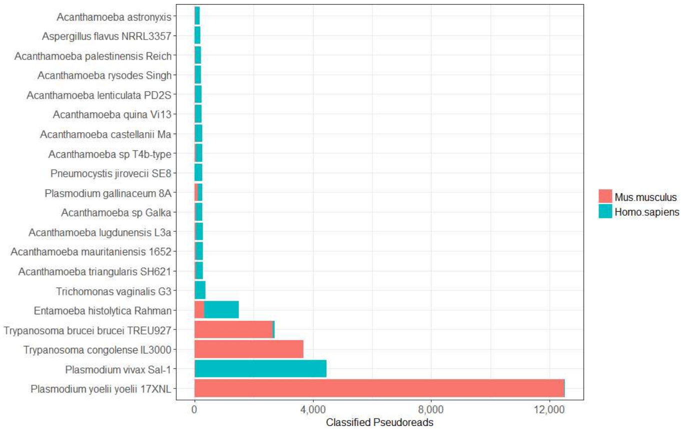

Fig 4. Human/Mouse classified pseudo-reads. This plot shows the 20 genomes with the most number of pseudo-reads classified as either human or mouse. Perhaps not surprisingly, the mouse strain of malaria, *P. yoelii*, contains a substantial number of contaminant reads from mouse. S3 Table contains pseudo-read human and mouse classifications for all eukaryotic pathogen genomes.

https://doi.org/10.1371/journal.pcbi.1006277.g004

Plasmodium falciparum is a human malarial parasite, it originated from the gorilla malarial parasite [30]. More than 450 pseudo-reads for *Plasmodium falciparum* were identified as gorilla. Similarly, *Plasmodium reichenowi* is a malarial parasite in chimpanzees and was one of only two Plasmodium genomes to have chimpanzee pseudo-reads [30]. Interestingly, *Edhazardia aedis* had 55% of its genome length masked, but had very few classified pseudo-reads. Instead, the majority of its non-masked sequences to begin with were stretches of DNA less than 100bp. Over 358,000 individual sequences were very small contigs, shorter than 100bp which are masked due to length.

Testing the pathogen database against a set of human cornea samples

To measure the effectiveness of our database cleaning method for NGS diagnosis of human infections, we evaluated a set of 20 human cornea samples recently described by Li et. al 2018 [31] against our EuPathDB-clean. The 20 corneal samples include 4 bacterial infections, 9 eukaryotic pathogen infections, 3 herpes virus infections, and 4 controls. Details about these samples and the true positive pathogens in each sample are listed in **Table 3, with additional clinical information listed in S4 Table.**

For testing, we used 4 Kraken databases: the original EuPathDB, EuPathDB-clean, RefSeq EuPathDB, and a general Microbe Database. The RefSeq EuPathDB contains all protozoal and fungal genomes from the RefSeq database as of December 2017. The Microbe database contains all RefSeq complete bacterial, archaeal, and viral genomes as of December 2017, and it also includes EuPathDB-clean. Genomes contained in each of the above databases are listed in **S5 Table.**

We first used Bowtie2 to align all corneal sample reads against the human genome reference, GRCh38.p7, and extracted any unaligned reads for each sample (Table 3). The non-

Table 3. Cornea sample true positives. This table summarizes the pathogens present in each of the corneal samples. Metagenomic shotgun sequencing was performed on all samples as described in [31] generating from 20–46 million pairs of 75-bp reads per sample. Sequencing was done in two batches of 10 samples each, where the 10 samples were multiplexed.

Case #	True Positives	Total 75-bp Paired Reads	Non-Human Aligned Reads
Case 1	*Staphyloccoccus aureus*	35,947,243	8,166,922
Case 2	*Streptococcus agalactiae*	42,281,022	2,354,821
Case 3	*Mycobacterium*	32,321,057	1,440,343
Case 4	*Mycobacterium chelonae*	31,259,428	2,927,088
Case 5	*Candida parapsilosis*	22,572,576	3,615,840
Case 6	*Fusarium solani*	43,187,311	3,048,256
Case 7	*Candida albicans/dubliensis*	45,410,366	1,993,853
Case 8	*Curvularia*	42,359,755	3,181,901
Case 9	*Aspergillus flavus*	46,033,752	2,875,199
Case 10	*Anncaliia algerae*	20,060,037	2,756,229
Case 11	*Acanthamoeba*	43,742,352	2,880,293
Case 12	*Acanthamoeba*	46,648,496	3,602,638
Case 13	*Acanthamoeba*	44,554,101	3,472,961
Case 14	*Herpes simplex type 1*	22,460,961	1,470,059
Case 15	*Herpes simplex type 1*	25,512,845	1,411,580
Case 16	*Herpes simplex type 1*	23,749,398	3,874,558
Case 17	None	43,643,461	2,637,693
Case 18	None	45,824,224	2,341,716
Case 19	None	25,623,975	1,071,939
Case 20	None	25,202,226	1,823,615

https://doi.org/10.1371/journal.pcbi.1006277.t003 human reads from each sample were then classified against each database using Kraken. S6 Table lists the read counts for each species and genus identified in the corneal samples when classified against each database. S7 Table details the reads per megabase for the same species classifications identified in the corneal samples for each database.

Fig 5 summarizes the results when using each of the four databases to identify the pathogens in these samples. The classifications differed greatly depending on the database used, demonstrating the importance of database selection prior to the computational analysis of any NGS sample. However, in the case of diagnostics, the contamination in the raw (unprocessed) genome databases creates false positive signals that overwhelm the true pathogen of the samples. Classification with the RefSeq EuPathDB yields a similar distribution of microbes for every corneal sample (Fig 5B). The resulting read counts suggest that each cornea has a significant presence of *Magnaporthe oryzae*, a pathogen that infects rice plants, and *Toxoplasma gondii* [32]. Similarly, classification against the original EuPathDB presents *Toxoplasma gondii* as one of the primary infections in all but one of the corneal samples (Fig 5A). None of the cornea samples had infections by either *Magnaporthe oryzae* or *Toxoplasma gondii* [31],thus both of these classifications are false positives.

The contamination removal process masked on average 5% of each *Toxoplasma gondii* genome. For example, the initial *Toxoplasma gondii ME49* genome is ~60 Mb long and the final masked genome is 57 Mb. Fortunately, removing this relatively small proportion of the genome produced a cleaned database with a far better classification profile for the corneal samples. As shown in Fig 5C, the correct eukaryotic infections for Cases 7, 9, 10, 11, and 12 are immediately evident with the new database. Instead of thousands of reads per megabase identified as *Toxoplasma gondii*, the new database shows very high (and correct) reads per megabase counts for *Anncaliia algerae* in Case 10, *Candida albicans* in Case 7, *Aspergillus* in Case 9, and *Acanthamoeba* in Cases 11 and 12, all true positive infections. With EuPathDB-clean, the maximum number of reads per megabase labeled as *Toxoplasma gondii* in any single sample was 0.35.

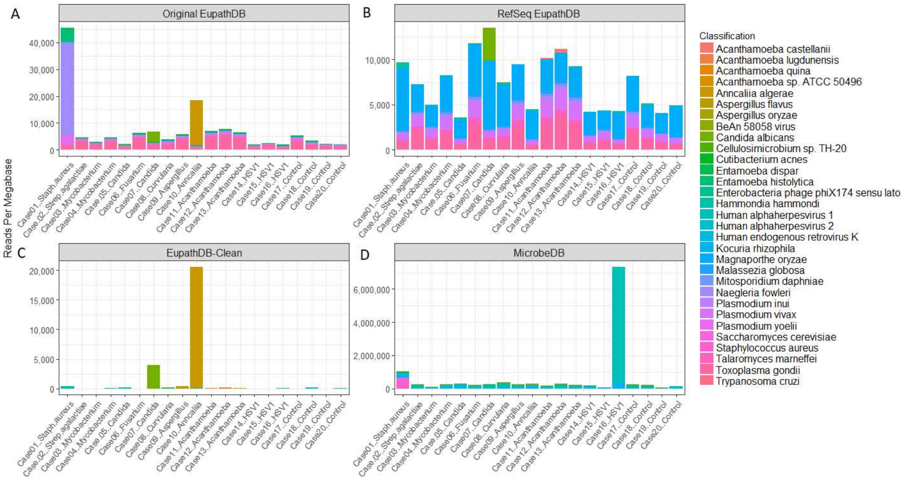

Fig 5. Top 10 species identified in corneal samples per database. The non-human reads from the 20 corneal samples were classified against four different Kraken databases: the original EuPathDB (A), EuPathDB-clean (B), RefSeq EuPathDB (C), and the final MicrobeDB (D). The plot above shows the 10 species with the most classified reads per megabase in a single corneal sample.

https://doi.org/10.1371/journal.pcbi.1006277.g005

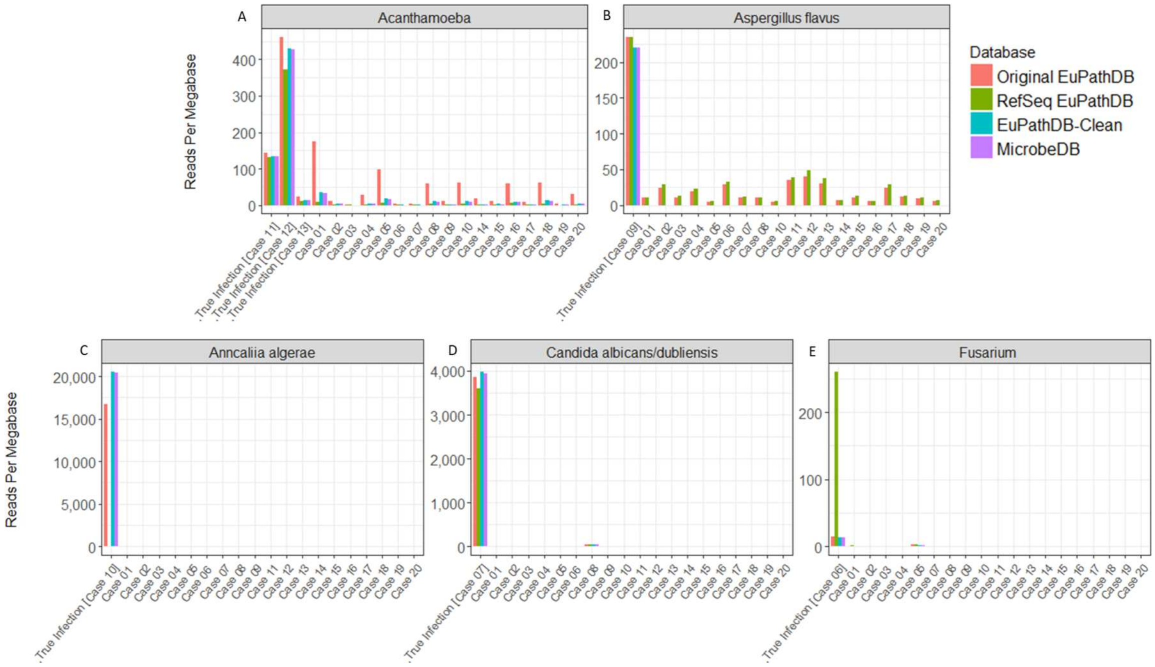

Fig 6. Number of classified reads per megabase for five true species/genera compared among four databases across all corneal samples. The above plot compares the reads per megabase for the true pathogens in the infected samples and also shows the reads per megabase from those pathogens in the remaining corneal samples. The five true species/genera are *Acanthamoeba* (A), *Aspergillus flavus* (B), *Anncaliia algerae* (C), *Candida albicans/dubliensis* (D), and *Fusarium* (E) S7 Table lists classified reads per megabase for each species for each database.

https://doi.org/10.1371/journal.pcbi.1006277.g006

After combining EuPathDB-clean with the RefSeq prokaryotes to create MicrobeDB (Fig 5D), we still found a strong signal for the eukaryotic pathogens in their corresponding true positive samples; e.g., the signal from *Anncaliia algerae* in Case 10 in **Fig 5D.** We note that other microbial contamination appears evident when using this database: in particular, *Kocuria rhizophila* appears in every sample, often at high levels. This does not appear to be a database error, as the *K*. *rhizophila* genome shows no sign of contamination. Instead, the reads from *K*. *rhizophila* are likely a consequence of physical contamination of the samples at some point in the process.

Another way to look at the data is to examine the reads per megabase counts for the true positive species only, as shown in Fig 6. Here we show the number of reads per megabase in each sample that were assigned to the 5 eukaryotic pathogens known to be present in at least one of the samples. With the original EuPathDB, the non-infected samples, alongside the truly infected samples, each appear to have numerous reads classified as both Acanthamoeba (Fig 6A) and *Aspergillus flavus* (Fig 6B) The RefSeq EuPathDB performed much better than the original EuPathDB, identifying the correct pathogen in the infected cases for Fusarium, Candida, and Acanthamoeba. However, RefSeq EuPathDB missed the *Anncaliia algerae* infection because that genome is missing from that database. Although RefSeq EuPathDB and EuPathDB both had a strong signal for *Aspergillus flavus* in the infected Case 9, the databases also identified hundreds to thousands of *Aspergillus flavus* reads in all of the non-infected samples. By comparison, EuPathDB-clean identified less than 10 *Aspergillus flavus* reads in all non-*Aspergillus*-infected samples while maintaining a strong signal for *Aspergillus flavus* in Case 9. MicrobeDB had near identical results to EuPathDB-clean for the true positive species/genera, identifying the infections in the infected samples.

Conclusion

In principle, next-generation sequencing can identify all microbial organisms within any sample, making it a potentially a revolutionary method for the diagnosis of human infections. However, this method relies heavily on the computational analysis that compares sequencing reads against reference databases, such as RefSeq and GenBank. Although new genomes are being sequenced daily, the reference databases remain incomplete and, because most new genomes are in draft form, inaccurate. Recent studies have identified contamination in many published genomes, hindering our ability to use them for accurate diagnosis.

We therefore developed a comprehensive contamination removal process, identifying human, vertebrate, bacterial, viral, archaeal, and vector contamination in 245 eukaryotic pathogen draft genomes. By removing contamination and low-complexity sequences, we have created a much cleaner database that minimizes false positives and provides better identification of true positive pathogens in NGS diagnostic samples.

Supporting information

S1 Table. Masking database composition. This table lists all genomes used to filter the Eukaryotic Pathogen genomes, including genome accessions, taxonomy IDs, species names, and strain specific names.
(XLSX)

S2 Table. Eukaryotic pathogen database genomes. Each genome is listed along with their filename, sub-categories, genus, species and genome lengths ***pre-/post-cleaning.***
(XLSX)

S3 Table. Pseudo-read counts. This table lists the number of pseudo-reads for each eukaryotic pathogen that were mistakenly classified as Bacteria, Archaea, Viral, Human, Mouse, or a number of vertebrate categories (i.e. fish, bird, mammal, etc).
(XLSX)

S4 Table. Cornea samples test case. Each cornea sample is listed alongside the clinical diagnosis, microbiology test result, and the expected true positive pathogen.
(XLSX)

S5 Table. Testing database composition. This table lists all genomes included in the RefSeq Eukaryotic Pathogen database and the MicrobeDB database.
(XLSX)

S6 Table. Cornea sample reads. These tables contain the read counts for all genus and species classifications assigned by Kraken for the 20 corneal samples when using the original EuPathDB, EuPathDB-clean, the RefSeq EuPathDB, and the MicrobeDB databases.
(XLSX)

S7 Table. Cornea sample reads per megabase. These tables contain the reads per megabase for all species classifications assigned by Kraken for the 20 corneal samples when using the original EuPathDB, EuPathDB-clean, the RefSeq EuPathDB, and the MicrobeDB databases.
(XLSX)

Author Contributions

Conceptualization: Jennifer Lu, Steven L. Salzberg.

Formal analysis: Jennifer Lu.

Funding acquisition: Steven L. Salzberg.

Investigation: Steven L. Salzberg.

Methodology: Jennifer Lu, Steven L. Salzberg.

Project administration: Steven L. Salzberg.

Resources: Steven L. Salzberg.

Software: Jennifer Lu.

Supervision: Steven L. Salzberg.

Validation: Jennifer Lu.

Visualization: Jennifer Lu.

Writing – original draft: Jennifer Lu.

Writing – review & editing: Jennifer Lu, Steven L. Salzberg.

References

1. Glaser CA, Gilliam S, Schnurr D, Forghani B, Honarmand S, Khetsuriani N, et al. In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998–2000. Clin Infect Dis. 2003; 36(6):731–42. https://doi.org/10.1086/367841 PMID: 12627357
2. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZM, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA. 2013; 309(14):1502–10. https://doi.org/10.1001/jama.2013.3231 PMID: 23571589
3. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, et al. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 2014; 52(1):139–46. https://doi.org/10.1128/JCM.02452-13 PMID: 24172157
4. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. 2014; 370(25):2408–17. https://doi.org/10.1056/NEJMoa1401268 PMID: 24896819
5. Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, et al. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. Neurol Neuroimmunol Neuroinflamm. 2016; 3(4):e251. https://doi.org/10.1212/NXI.0000000000000251 PMID: 27340685
6. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012; 9(8):811–4. https://doi.org/10.1038/nmeth.2066 PMID: 22688413
7. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807; PubMed Central PMCID: PMC4053813.
8. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
10. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2015; 43(Database issue):D30–5. https://doi.org/10.1093/nar/gku1216 PMID: 25414350; PubMed Central PMCID: PMCPMC4383990.
11. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804
12. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995; 59(1):143–69. PMID: 7535888
13. Daniel R. The metagenomics of soil. Nat Rev Microbiol. 2005; 3(6):470–8. https://doi.org/10.1038/nrmicro1160 PMID: 15931165

14. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon Cenarchaeum symbiosum. J Bacteriol. 1998; 180(19):5003–9. PMID: 9748430

15. Butinar L, Spencer-Martins I, Gunde-Cimerman N. Yeasts in high Arctic glaciers: the discovery of a new habitat for eukaryotic microorganisms. Antonie Van Leeuwenhoek. 2007; 91(3):277–89. https://doi.org/10.1007/s10482-006-9117-3 PMID: 17072534

16. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. Genome Biol. 2002; 3(2): REVIEWS0003. https://doi.org/10.1186/gb-2002-3-2-reviews0003

17. Karl DM. Hidden in a sea of microbes. Nature. 2002; 415(6872):590–1. https://doi.org/10.1038/415590b PMID: 11832923

18. Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol. 2007; 5(3):e82. https://doi.org/10.1371/journal.pbio.0050082 PMID: 17355177

19. Haldar K, Kamoun S, Hiller NL, Bhattacharje S, van Ooij C. Common infection strategies of pathogenic eukaryotes. Nat Rev Microbiol. 2006; 4(12):922–31. https://doi.org/10.1038/nrmicro1549 PMID: 17088934

20. Jones JL, Kruszon-Moran D, Wilson M, McQuillan G, Navin T, McAuley JB. Toxoplasma gondii infection in the United States: seroprevalence and risk factors. Am J Epidemiol. 2001; 154(4):357–65. https://doi.org/10.1093/aje/154.4.357 PMID: 11495859

21. Thomas PA. Fungal infections of the cornea. Eye. 2003; 17(8):852–62. https://doi.org/10.1038/sj.eye.6700557 PMID: 14631389

22. Niederkorn JY, Alizadeh H, Leher H, McCulley JP. The pathogenesis of Acanthamoeba keratitis. Microbes Infect. 1999; 1(6):437–43. https://doi.org/10.1016/S1286-4579(99)80047-1 PMID: 10602676

23. Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. 2017; 45(D1):D581–D91. https://doi.org/10.1093/nar/gkw1105 PMID: 27903906

24. Longo MS, O'Neill MJ, O'Neill RJ. Abundant human DNA contamination identified in non-primate genome databases. PLoS One. 2011; 6(2):e16410. https://doi.org/10.1371/journal.pone.0016410 PMID: 21358816

25. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014; 2:e675. https://doi.org/10.7717/peerj.675 PMID: 25426337

26. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci. 2015; 10:18. https://doi.org/10.1186/1944-3277-10-18 PMID: 26203331

27. Kryukov K, Imanishi T. Human Contamination in Public Genome Assemblies. PLoS One. 2016; 11(9): e0162424. https://doi.org/10.1371/journal.pone.0162424 PMID: 27611326

28. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006; 13(5):1028–40. https://doi.org/10.1089/cmb.2006.13.1028 PMID: 16796549

29. Böhme U, Otto TD, A Cotton J, Steinbiss S, Sanders M, Oyola SO, et al. Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals. Genome Res. 2018. https://doi.org/10.1101/gr.218123.116 PMID: 29500236

30. Prugnolle F, Durand P, Neel C, Ollomo B, Ayala FJ, Arnathau C, et al. African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. Proc Natl Acad Sci U S A. 2010; 107(4):1458–63. https://doi.org/10.1073/pnas.0914440107 PMID: 20133889

31. Li Z, Breitwieser FP, Lu J, Jun AS, Asnaghi L, Salzberg SL, et al. Identifying Corneal Infections in Formalin-Fixed Specimens Using Next Generation Sequencing. Invest Ophthalmol Vis Sci. 2018; 59 (1):280–8. https://doi.org/10.1167/iovs.17-21617 PMID: 29340642

32. Liu J, Wang X, Mitchell T, Hu Y, Liu X, Dai L, et al. Recent progress and understanding of the molecular mechanisms of the rice-Magnaporthe oryzae interaction. Mol Plant Pathol. 2010; 11(3):419–27. https://doi.org/10.1111/j.1364-3703.2009.00607.x PMID: 20447289