University of Waterloo
Faculty of Science

BIOL 499: Senior Honours Project

Supervisor: Dr. Andrew Doxey (acdoxey@uwaterloo.ca)
Second Reader: Dr. Gabriel Moreno-Hagelsieb (gmoreno@wlu.ca)

Project Final Report

Determining Novel Natural Reservoirs and Strains of *Yersinia pestis* through SRA Database Mining

Richard (Zhifei) Dong (20775623)

December 16th, 2022

The report is submitted as a requirement of BIOL 499, Senior Honours Project.

Acknowledgment

I would like to express my sincere thanks for Dr. Andrew Doxey, my supervisor for this project. He not only brought me into the fantastic world of bioinformatics and genomics through BIOL 266: Introduction to Computational Biology and BIOL 469: Genomics, but also provided me with directions and ideas throughout the project. I would have not become interested in bioinformatics nor been completing this project without him. Therefore, it is fair to say that his teaching and guidance enabled me to find the academic field that I truly enjoyed, and I cannot be more thankful for that.

In addition, I would like to thank Dr. Gabriel Moreno-Hagelsieb from the Wilfrid Laurier University for being my second reader as well as providing constructive comments for during my presentation. I look forward to testing out some of the ideas he provided in the upcoming months as a continuation for this project. I would also like to thank all the lab members at Doxey Lab for providing feedbacks for my practise run of my presentation. Amongst all those great people, I would like to especially thank Mr. Harold Hodgins and Dr. Briallen Lobb for all the helps in troubleshooting for Compute Canada as well as the various applications I was using.

Lastly, I would also like to thank my friends Ms. Yolanda Tu and Mr. Gary Hou for listening to my presentations and providing comments, as well as Mr. Tony Pan for helping me in getting started for Compute Canada. Completing this project is one of the best things I have done in my final year at the University of Waterloo, and I would love to explore this fantastic field of computational biology, bioinformatics, and genomics further in my future academic studies.

Abstract

Yersinia pestis is a deadly bacterium that is responsible for multiple pandemics throughout the history and still causes hundreds of deaths yearly in the less developed part of the world. To better monitor the evolution of this bacterium, a database search from the NCBI Sequence Read Archive (SRA) was conducted as an attempt to discover novel reservoirs and strains of *Y. pestis*. By using various bioinformatic tools to reconstruct the genome and phylogenetically classify the samples, it was found that *Pseudogymnoascus destructans*, commonly known as white-nose fungus from bats and is a previously undocumented reservoir for *Y. pestis*, carried four distinct strains of the bacterium, with one of them being a highly divergent and potentially novel strain from existing *Y. pestis* references. The results of this study illustrated the importance of genetic surveillance, and further investigations into the pathogenicity of those strains should be conducted in order to avoid future pandemics caused by *Y. pestis*.

Keywords: Genomic Surveillance; Sequence Read Archive; *Yersinia pestis*; *Pseudogymnoascus destructans*

List of Tables

List of Figures

Table of Contents

Introduction

From the earliest recorded pandemic in 430 BC during the Peloponnesian War to the COVID-19 pandemic that started since the late 2019, the human population has been periodically fighting pandemics since the very beginning of civilization [1]. Amongst all those pandemics, the plague arguably has the most profound impact on human societies and civilizations due to its high lethality, killing millions of people across many countries and throughout history [2].

The plague is caused by the bacterial infection of *Yersinia pestis*, which is carried by over 200 species of wild rodents [3] and transmitted to humans through flea bites (primarily causing bubonic or septicaemic plague) or through inhalation of droplets (primarily causing pneumonic plague) [4]. According to a review by Zhou et al., *Y. pestis* was found to be able to infect 87 mammalian species, where 13 of them (all rodents) are determined as primarily reservoirs. In addition, 41 species of fleas were found to be infected with *Y. pestis* in nature, where 14 of them are determined as primary vectors [5]. Three different strains of *Y. pestis* are responsible for the three major plague outbreaks throughout history: the Biovar Antiqua caused the Justinian's plague (AD 541 to 767), the Biovar Mediaevalis caused the Black Death (1346 to early 19th century), and the Biovar Orientalis causes the modern plague (since 1894) [6, 7]. The difference in the strains primarily lies in the ability of the bacterium to reduce nitrate and utilize glycerol [7]. Yet, for all the strains, the case-fatality ratio can range anywhere between 30% to as high as 100% if left untreated [8].

Y. pestis belongs to the *Yersinia genus* from the Enterobacteriaceae family. Out of the current 19 species from the Yersinia genus, only three are disease-causing to humans: *Y. pestis*, *Yersinia pseudotuberculosis*, and *Yersinia enterocolitica* [9, 10]. However, *Y. pseudotuberculosis* and *Y. enterocolitica* are primarily spread via the ingestion of contaminated food and not spread by fleas [11]. Furthermore, *Y. entercolitica* can be found in hosts such as *Miniopterus schreibersii* (common bent-wing bat) species [12], yet so far, no study has reported bats to be the potential

hosts or reservoirs for *Y. pestis*. Additionally, *Y. pseudotuberculosis* and *Y. enterocolitica* primarily cause acute gastroenteritis and mesenteric lymphadenitis, and their fatality rates are much lower than that of *Y. pestis* [13, 14]. However, despite those large differences, *Y. pestis* is very closely related to *Y. pseudotuberculosis* and shares many genes with *Y. pseudotuberculosis* [15, 16].

With improved sanitation and the development of antibiotics to treat infections [17], the plague has been viewed by many in developed nations as a problem of the past. Despite this, the modern plague remains a major public health issue in many less-developed parts of the world. There are over 3,000 cases and 500 deaths of plagues being reported from 2010 to 2015, and in the recent years there are new cases being reported in Libya and Algeria [18]. Furthermore, there is also the possibility of a multi-drug resistant strain emerging and the consequential utilization of plague as a bioweapon for terrorism attacks [9]. As a result, it is important for the world to monitor the presence of *Y. pestis* in a variety of different locations to ensure that potential new reservoirs for the bacterium are quickly identified and the evolution of the species is closely monitored so that the world can be more prepared if there is an emergence of a potential new strain.

Thanks to the advancement in bioinformatic databases, such tasks can now be completed by monitoring the genetic material of the targeted species in databases. The Sequence Read Archive (SRA) is an online database containing short reads from high-throughput sequencing [19]. A portion (about 16%) of the database consists of metagenome sequence reads from various environment samples [20], which makes this database a great starting point for the surveillance of novel *Y. pestis* genetic materials in novel environments. As a result, the purpose of this study is to determine the novel natural reservoirs and strains of *Y. pestis* through SRA database mining. It is hoped that the results of this study can help scientists around the globe to have a better understanding of *Y. pestis* in order to be prepared to contain future plague outbreaks as soon as possible.

Materials and Methods

This study was divided into three major phases: obtaining data from SRA, genome assembly, and genomic analysis (**Figure 1**).

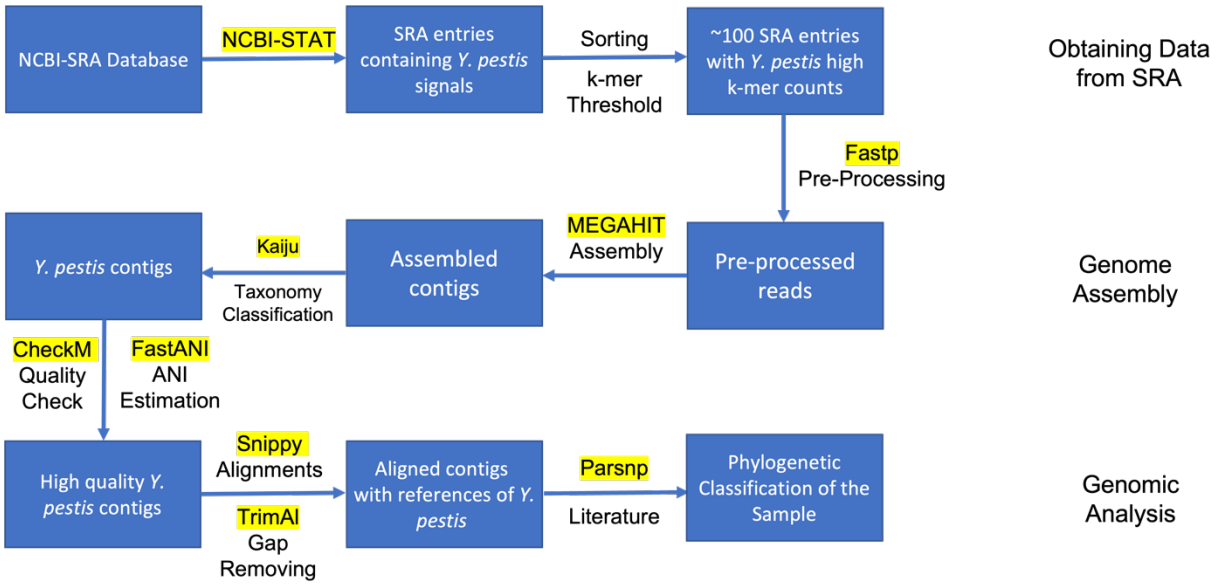

Figure 1: Overview of workflow in the study. The study was divided into three phases, and the bioinformatic applications used for each step is highlighted in yellow.

Phase 1: Obtaining SRA Data

As a starting point, the NCBI SRA database was searched by using the query of *Y. pestis* genome in order to determine the samples containing the genetic material of *Y. pestis*. This was accomplished using the NCBI Sequence Taxonomic Analysis Tool (STAT), which is a method developed by Katz et al. to effectively screen the taxonomic diversity on the submissions in the database based on 32 bp k-mer counts [21]. The returned result of the search would be a .txt file containing the samples with positive *Y. pestis* signals but with different k-mer counts. This file was then sorted based on the k-mer counts. The top hits are expected to be *Y. pestis* sequencing projects, which can be used as a positive control for the search. The bottom hits would be false

positive signals with very low k-mer counts, which needs to be excluded from the analysis. To narrow down the samples for further analysis, a threshold of k-mer counts was determined to be 10,000 to obtain at least about 10% coverage, assuming that the k-mers are non-overlapping. This produced a final list of about 100 samples, and the origins as well as the species of those 100 samples were examined as an initial assessment.

Phase 2: Genomic Assembly

Those samples of interest were then assembled into a full genome for future analysis. Those sequence reads were first filtered using the Fastp program, which is a preprocessor for FASTQ files. Some tasks that can be accomplished by this program include adapter trimming, quality control, per-read quality pruning, and quality filtering [22]. After the filtering process, the processed reads were passed onto the MEGAHIT assembler for genome assembly. This program, developed by Li et al., is capable of efficiently assembling large and complex metagenomics data, which makes this program very suitable for this study [23]. The output of this program was a collection of contigs, but the majority of them will be the genetic material of the environment samples. As a result, a taxonomic classification tool, Kaiju, was applied to the contig reads in order to identify the potential *Y. pestis* contigs [24]. Those selected *Y. pestis* contigs was then used as the draft genomes of interest, which will then be passed onto the CheckM program for assessing the assembly quality and contamination of genomes as well as estimating their completeness [25]. The genomes that passed the quality assessment by CheckM was used as the final genomes of interest to be analyzed in the final phase.

Phase 3: Genomic Analysis

To phylogenetically classify the environment samples obtained, existing 549 reference genomes for *Y. pestis* were downloaded from NCBI RefSeq database [26]. FastANI, a program developed by Jain et al., was used to compute average nucleotide identities (ANI) for the *Y. pestis* contigs against all reference genomes [27]. Then, the contigs were aligned with those reference genomes using Snippy, a program for multiple genome alignment as well as single nucleotide polymorphisms (SNPs) and insertion / deletion (Indels) identifications [28]. The result of alignments were passed to TrimAl, an automated program for alignment trimming [29], to remove any gap regions in order to improve the resolution on the phylogenetic trees to be construed soon. After trimming, the alignments were sent to Parsnp, an automated tool from the Harvest suite for fast phylogenetic tree construction and genome alignment visualization [30], to construct a phylogenetic tree for each individual alignment. The resulted phylogenetic trees were finally compared with literatures on the phylogenetic evolution of *Y. pestis* in order to determine the closest relatives to each sample as well as its potential evolutionary path.

Results

Phase 1: Obtaining SRA Data

The returned SRA-STAT results reveals that the majority of the entries with high k-mer counts are past *Y. pestis* sampling projects as expected. However, out of those high k-mer counts entries, four entries from sequencing projects for *Pseudogymnoascus destructans*, commonly known as white-nose fungus found in bats [31], stand out. **Table 1** outlines the important information from the SRA-STAT query results [21]. Since *P. destructans* is not a documented vector or reservoir for *Y. pestis*, those four samples were viewed as samples of interest and were taken for future analysis.

Table 1: SRA-STAT query result. This table outlines important information for each significant SRA entry of *P. destructans* entries returned, sorted by total k-mer count.

Accession ID	Size of Genome (in mega bases)	Location of Collection	Total k-mer Count	Collection Date
SRR6011472	5773	Canada	305846	2010-03-11
SRR6011475	7792	USA	23977	2009-01-27
SRR6011496	7300	Ukraine	11856	2011-02-03
SRR6011490	6215	USA	10188	2011-02-16

Phase 2: Genomic Assembly

The results of the MEGAHIT-assembled and Kaiju-filtered contigs are passed to CheckM for quality control. CheckM reports two important metrics of the contigs: genome completeness, calculated as the percentage of marker genes found within the assembly, as well as contamination, estimated using the number of times that single-copy marker genes appears [25]. The result from CheckM is reported in **Table 2**.

Table 2: CheckM results for the four samples. This table reports the completeness and contamination for each significant SRA entry of *P. destructans*, sorted by the reversed CheckM estimated percent completeness.

Accession ID	Completeness (in %)	Contamination (in %)
SRR6011490	0.00	0.00
SRR6011475	0.86	0.00
SRR6011496	18.97	0.00
SRR6011472	46.55	0.00

Phase 3: Genomic Analysis

The result for each step of the genomic reconstruction for all four samples are summarized in **Table 3** below.

Table 3: Summary for the genomic reconstruction results of each sample at each stage. It outlines the number of contigs obtained after Kaiju classification and the total length of those contigs, as well as the new length after Snippy and TrimAl for each of the SRA samples, sorted by reversed genome completeness estimated by CheckM.

Accession ID	Number of Contigs after Kaiju	Total Length of Contigs after Kaiju	Number of Base in Parsnp after TrimAl
SRR6011490	162	85,173	6,788
SRR6011475	597	391,718	39,897
SRR6011496	892	1,956,910	415,798
SRR6011472	152	2,478,008	602,859

Parsnp was then used to construct one phylogenetic tree for each of the sample through multiple genome alignment with *Y. pestis* references on NCBI RefSeq Database. The zoomed-in version of the phylogenetic trees for SRR6011490 (**Figure 2a**), SRR6011475 (**Figure 2b**), SRR6011496 (**Figure 2c**), and SRR6011472 (**Figure 2d**) can be found below to illustrate the closest relatives of the samples of interest. The full phylogenetic trees for all four samples are available in **Supporting Information**.

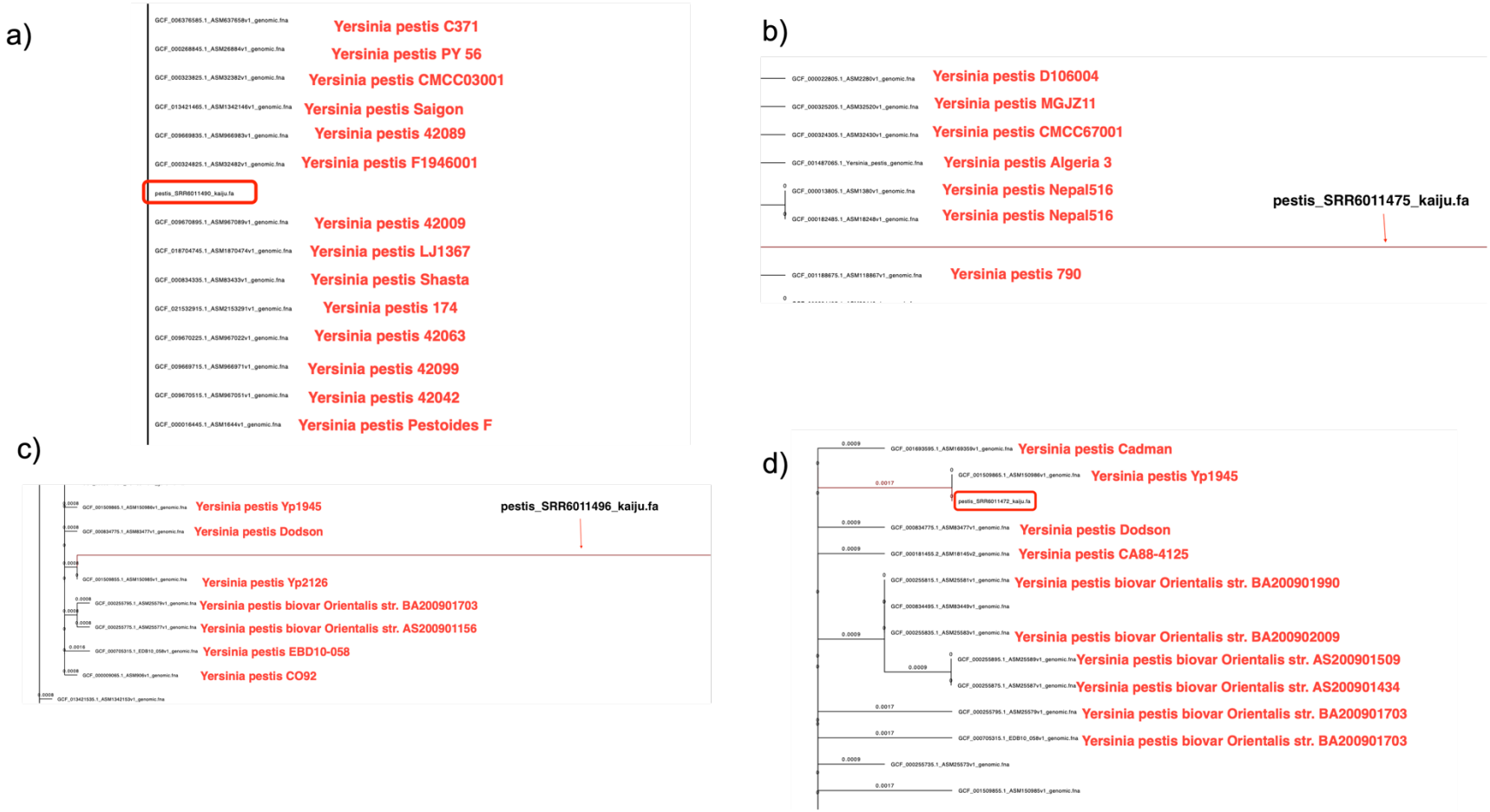

Figure 2: The zoomed-in partial phylogenetic trees. Results are constructed from the multiple genome alignment of the *Y. pestis* contigs of a) SRR6011490, b) SRR6011475, c) SRR6011496, and d) SRR6011472 against the references genomes of *Y. pestis*. Trees are zoomed in to show only the closest relatives of the SRA sample of interest. The trees are displayed with the original assembly IDs from NCBI, and are labelled with the name of the strain of *Y. pestis* in red.

Based on the trees from **Figure 2**, the closest relative from the references of *Y. pestis* to each sample is determined, and the percent genome identity is calculated from the Parsnp alignment by using $1 - \frac{\#\ of\ substitutions}{\#\ of\ aligned\ sites}$, where the number of aligned sites can be found in **Table 3**. The results for these genomic comparisons are summarized in **Table 4**.

Table 4: Genomic comparisons for each sample to its closest relative. The closest relative is determined by using the phylogenetics trees in **Figure 2**. The genomic identity is obtained by analyzing the multiple genomic alignment of the sample of interest and the closest relative from Parsnp alignment, and the ANI score is computed using FastANI. The samples are ordered by their reversed CheckM completeness estimations.

Accession ID	**Closest *Y. pestis* Reference Assembly (Strain Name)**	**Number of Substitutions**	**Calculated Percent Genomic Identity**	**ANI Score**
SRR6011490	GCF_000324825.1 (*Yersinia pestis* F1946001)	0	100.00	94.79
SRR6011475	GCF_001188675.1 (*Yersinia pestis* 790)	7	99.98	99.66
SRR6011496	GCF_001509855.1 (*Yersinia pestis* Yp2126)	456	99.89	99.87
SRR6011472	GCF_001509865.1 (*Yersinia pestis* Yp1945)	0	100.00	99.9958

Discussion

Figure 2a shows the zoomed-in partial phylogenetic tree of SRR6011490, the least completed genome estimated by CheckM. In fact, **Table 2** shows that CheckM estimated this genome to have a completeness of 0% as it does not contain any of the marker genes that CheckM is looking for. This illustrates that despite it having a relatively high k-mer count (**Table 1**) that passes the threshold, it is likely a false positive signal and this *P. destructans* sample probably does not contain *Y. pestis*. This result is also confirmed from the phylogenetic tree, as its closest relative, *Yersinia pestis* strain F1946001 (**Table 4**), has another close relative, *Yersinia pestis* strain Pestoides F. Studies have found that *Y. pestis* strain F1946001 belongs to Biovar Orientalis [32] while *Y. pestis* strain Pestoides F is a very unique strain that is derived from the most ancient lineage of *Y. pestis* [33]. Those two strains are on two different branches of the *Y. pestis* reference phylogenetic tree [34], and therefore they are classifying to be close together on the reconstructed phylogenetic tree likely means that the tree is inaccurate. Last but not the least, its ANI score is also relatively low, meaning that the calculated nucleotide identity of 100% is highly unlikely. This is potentially due to the signal being false positive and therefore after all the processing, the final alignment result (**Table 3**), lacks enough phylogenetic informative sites. In the future, a higher k-mer count threshold can be set in order to avoid including false positive signals in the analysis.

The next sample of interest is SRR6011475, which is the second least completed genome out of the four samples (**Table 2**). This explains the great branch length evident from **Figure 2b** of the zoom-in version of the phylogenetic tree, as similar to the case with the previous sample (SRR6011490), the lack of enough phylogenetically informative sites (**Table 3**) hinders the ability for Parsnp to construct a tree with great resolution. Yet despite the relatively poor resolution of the tree, the phylogenetic classification is indeed accurate as it does cluster phylogenetically close reference strains together. All the relatives of SRR6011475, including its closest relative *Yersinia pestis* 790, belong to Biovar Antiqua and are responsible for the most ancient outbreak of plague

by *Y. pestis* [35]. Although this biovar of *Y. pestis* no longer causes modern plagues, evidence suggests that the ancient versions bacteria are still circulating in the environment [36]. Therefore, it is likely that the few substitutions of this sample seen (**Table 4**) are a result of the long-time evolution from Biovar Antiqua *Y. pestis* in the environment. However, admittedly the resolution of the tree is also not great, and therefore in the future, other methods (such as Mash for whole-genome homology maps [37]) should be tried in order to see if a better homology can be identified for the sample so that a better evolutionary path may be inferred.

The third sample of interest is SRR6011496, a sample obtained from white nose fungus population in Ukraine in 2011 (**Table 1**). When compared to the previous two samples, this sample has a relatively higher completeness estimation (**Table 2**), therefore having a fairly good phylogenetic tree with great resolution when being computed by Parsnp (**Figure 2c**). From the phylogenetic tree, it is evident that the closest relative to this sample is *Y. pestis* strain Yp 2126, which was obtained as a part of a sequencing project for *Y. pestis* from Gunnison's prairie dog colonies (*Cynomys gunnisoni*) in Arizona, USA in 2011 [38]. However, despite that those two species are close relatives, the assembled *Y. pestis* genome from SRR6011496 is very distant from *Y. pestis* strain Yp 2126. In fact, when viewing the full phylogenetic tree, the assembled genome appears to be the most divergent genome of *Y. pestis* ever recorded. Literature has phylogenetically classified Yp 2126 as a strain from Biovar Orientalis, the biovar responsible for the modern plague [16], and therefore it is possible that the sample of interest has undergone significant mutations from this strain. This is possible as *Y. pestis* is found to accumulate mutations much more rapidly (at a rate up to 100 times faster) in modern strains when compared to their ancient counterparts [39]. Yet, the fact that 456 mutations can give rise to a sample this divergent from the rest is alarming and may suggest that the sample may not be *Y. pestis* but its closest relative, *Y. pseudotuberculosis*, or even *Y. entercolitica*, which is found in bat species [12]. Therefore, despite the high calculated genome identity and ANI (**Table 4**) that are greatly above the genome identity between *Y. pestis* and *Y. pseudotuberculosis* at about 97% [40], in the future,

more research should still be conducted to verify the signal from this sample in order to better phylogenetically identify it.

The last sample of interest is SRR6011472, which is a sample obtained from Canada in 2010 (**Table 1**). This sample also has the most k-mer count (**Table 1**) and is the most complete sample (**Table 2**) with the greatest number of bases aligned (**Table 3**) out of the four samples that were studied. Based on its partial phylogenetic tree in **Figure 2d**, its closest relative is *Y. pestis* strain Yp1945, a species sample from black-tailed prairie dog (*C. ludovicianus*) colonies in Texas, USA in 2009 [38]. Amongst all the sites aligned, this sample is exactly identical to strain Yp1945 with no substitutions, yet ANI does reveal that there is a slight difference in nucleotide composition (**Table 4**). Regardless, it is likely that one strain is derived from another due to this high genomic similarity. Several hypotheses have been formulated about the relationship between those two samples: firstly, fleas, the primary vectors for *Y. pestis*, are known to live in moist and dark environments [41, 42], which include the caves that bats live in [43]. As a result, since prairie dogs are known reservoirs for *Y. pestis* [44], a cross-colonization event probably has occurred between fleas on bats infected with white-nose fungus and fleas on prairies dogs, which resulted in *Y. pestis* samples being detected in *P. destructans* samples. Another hypothesis is centred on the idea that there exists a synergistic relationship between prairie dogs and bats. In particular, studies have found that bats may make use of abandoned burrows from prairie dogs as their habitats [45]. Therefore, a cross-colonization event could also potentially occur as a result of this synergistic relationship. Yet regardless of how the transmission could have occurred, this finding demonstrated that *Y. pestis* has the potential to spread far, and therefore monitoring the presence of this species is critical for future pandemic control.

Conclusion and Future Perspectives

In this study, an SRA database search pipeline was developed to identify the presence of *Y. pestis* on unexpected species in order to monitor the emergence of novel strains and reservoirs. Using the pipeline, four distinct SRA samples of *P. destructans* were found to pass the significant threshold for potentially containing *Y. pestis*. Out of those four, one sample (SRR6011490) was discarded due to its low genome completeness and poor phylogenetic resolution. For the three samples considered, one sample (SRR6011475) is believed to have been evolved from an ancient strain of *Y. pestis*, another sample (SRR6011496) is believed to be highly divergent from existing *Y. pestis* references, and the last sample (SRR6011472) was identical to one of the *Y. pestis* reference genome and is believed to emerge as a result of cross-colonization between prairie dogs and bats.

Despite that the findings were exciting, the potential of this study is limited by resolution of the phylogenetic tree, which is not very great for the samples with low genome completeness estimation, such as SRR6011490 and SRR6011475. As a result, a follow-up study should be conducted in the future to utilize better bioinformatic software for the phylogenetic classification of those low completeness samples. In addition, due to the high divergence seen on the phylogenetic tree for the highly divergent SRA sample (SRR6011496), additional verification is required to determine if it is indeed a *Y. pestis* sample. This could be done through a wide homology comparison of this sample on various members of the *Yersinia* genus. When phylogenetically accurately classified, better evolution insights on this sample can be obtained, and additional gene set comparisons or SNPs analysis may be performed to determine if this strain has the potential to be spread onto *P. destructans* to use it as a new reservoir. The findings of this study illustrated the importance of genomic surveillance through the use of various bioinformatic databases, and therefore in the future, similar studies should be constantly conducted for other pathogens as well in order to better prevent the outbreak of diseases.

Supporting Information

This supporting information contains the full phylogenetic trees of the four species in **Figure 2**. Those trees can be found at the following Google Drive link:

https://drive.google.com/drive/folders/1Ds3Ms6y348RbIXNqAVCSPSv6bKWsyLWR?usp=share_link

The sample of interest in each tree is labelled as *pestis_SRRID_kaiju.fa*, where *SRRID* is the SRR ID of the sample.

References

1. Schwartz RA, Kapila R: **Pandemics throughout the centuries**. *Clinics in Dermatology* 2021, **39**(1):5-8.
2. Aberth J: **From the brink of the apocalypse : confronting famine, war, plague, and death in the later middle ages**. New York: Routledge; 2001.
3. Anisimov AP, Amoako KK: **Treatment of plague: promising alternatives to antibiotics**. *Journal of Medical Microbiology* 2006, **55**(11):1461-1475.
4. Davis KM: **All Yersinia Are Not Created Equal: Phenotypic Adaptation to Distinct Niches Within Mammalian Tissues**. *Frontiers in Cellular and Infection Microbiology* 2018, **8**.
5. Zhou D, Han Y, Song Y, Huang P, Yang R: **Comparative and evolutionary genomics of Yersinia pestis**. *Microbes and Infection* 2004, **6**(13):1226-1234.
6. Chain PSG, Hu P, Malfatti SA, Radnedge L, Larimer F, Vergez LM, Worsham P, Chu MC, Andersen GL: **Complete Genome Sequence of Yersinia pestis Strains Antiqua and Nepal516: Evidence of Gene Reduction in an Emerging Pathogen**. *Journal of Bacteriology* 2006, **188**(12):4453-4463.
7. Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H *et al*: **Complete Genome Sequence of Yersinia pestis Strain 91001, an Isolate Avirulent to Humans**. *DNA Research* 2004, **11**(3):179-197.
8. Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E, Carniel E, Gage KL, Leirs H, Rahalison L: **Plague: Past, Present, and Future**. *PLOS Medicine* 2008, **5**(1):e3.
9. Tan SY, Dutta A, Jakubovics NS, Ang MY, Siow CC, Mutha NVR, Heydari H, Wee WY, Wong GJ, Choo SW: **YersiniaBase: a genomic resource and analysis platform for comparative analysis of Yersinia**. *BMC Bioinformatics* 2015, **16**(1):9.
10. Savin C, Criscuolo A, Guglielmini J, Le Guern A-S, Carniel E, Pizarro-Cerdá J, Brisse S: **Genus-wide Yersinia core-genome multilocus sequence typing for species identification and strain characterization**. *Microbial Genomics* 2019, **5**(10).
11. Galindo CL, Rosenzweig JA, Kirtley ML, Chopra AK: **Pathogenesis of Y. enterocolitica and Y. pseudotuberculosis in Human Yersiniosis**. *Journal of Pathogens* 2011, **2011**:182051.
12. Imnadze T, Natradze I, Zhgenti E, Malania L, Abazashvili N, Sidamonidze K, Khmaladze E, Zakalashvili M, Imnadze P, Arner RJ *et al*: **Identification of a Novel Yersinia enterocolitica Strain from Bats in Association with a Bat Die-Off That Occurred in Georgia (Caucasus)**. *Microorganisms* 2020, **8**(7):1000.
13. Long C, Jones TF, Vugia DJ, Scheftel J, Strockbine N, Ryan P, Shiferaw B, Tauxe RV, Gould LH: **Yersinia pseudotuberculosis and Y. enterocolitica infections, FoodNet, 1996-2007**. *Emerg Infect Dis* 2010, **16**(3):566-567.
14. Marks MI, Pai CH, Lafleur L, Lackman L, Hammerberg O: **Yersinia enterocolitica gastroenteritis: A prospective study of clinical, bacteriologic, and epidemiologic features**. *The Journal of Pediatrics* 1980, **96**(1):26-31.
15. Chain PSG, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL *et al*: **Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis**. *Proceedings of the National Academy of Sciences* 2004, **101**(38):13826-13831.

16. Demeure CE, Dussurget O, Mas Fiol G, Le Guern A-S, Savin C, Pizarro-Cerdá J: **Yersinia pestis and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination, and diagnostics**. *Genes & Immunity* 2019, **20**(5):357-370.
17. Riedel S: **Plague: From Natural Disease to Bioterrorism**. *Baylor University Medical Center Proceedings* 2005, **18**(2):116-124.
18. Bragazzi NL, Mahroum N: **Google Trends Predicts Present and Future Plague Cases During the Plague Outbreak in Madagascar: Infodemiological Study**. *JMIR Public Health Surveill* 2019, **5**(1):e13142.
19. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Research* 2008, **36**(suppl_1):D13-D21.
20. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C: **The Sequence Read Archive**. *Nucleic Acids Research* 2011, **39**(suppl_1):D19-D21.
21. Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C: **STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions**. *Genome Biology* 2021, **22**(1):270.
22. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.
23. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**. *Bioinformatics* 2015, **31**(10):1674-1676.
24. Menzel P, Ng KL, Krogh A: **Fast and sensitive taxonomic classification for metagenomics with Kaiju**. *Nature Communications* 2016, **7**(1):11257.
25. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome Res* 2015, **25**(7):1043-1055.
26. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Research* 2015, **44**(D1):D733-D745.
27. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**. *Nature Communications* 2018, **9**(1):5114.
28. Olawoye IB, Frost SDW, Happi CT: **The Bacteria Genome Pipeline (BAGEP): an automated, scalable workflow for bacteria genomes with Snakemake**. *PeerJ* 2020, **8**:e10121.
29. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009, **25**(15):1972-1973.
30. Treangen TJ, Ondov BD, Koren S, Phillippy AM: **The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes**. *Genome Biology* 2014, **15**(11):524.

31. Hoyt JR, Cheng TL, Langwig KE, Hee MM, Frick WF, Kilpatrick AM: **Bacteria Isolated from Bats Inhibit the Growth of Pseudogymnoascus destructans, the Causative Agent of White-Nose Syndrome**. *PLOS ONE* 2015, **10**(4):e0121329.
32. Li Y, Cui Y, Cui B, Yan Y, Yang X, Wang H, Qi Z, Zhang Q, Xiao X, Guo Z *et al*: **Features of Variable Number of Tandem Repeats in Yersinia pestis and the Development of a Hierarchical Genotyping Scheme**. *PLoS One* 2013, **8**(6):e66567.
33. Garcia E, Worsham P, Bearden S, Malfatti S, Lang D, Larimer F, Lindler L, Chain P: **Pestoides F, an atypical Yersinia pestis strain from the former Soviet Union**. *Adv Exp Med Biol* 2007, **603**:17-22.
34. Rasmussen S, Allentoft Morten E, Nielsen K, Orlando L, Sikora M, Sjögren K-G, Pedersen Anders G, Schubert M, Van Dam A, Kapel Christian Moliin O *et al*: **Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago**. *Cell* 2015, **163**(3):571-582.
35. Kutyrev VV, Eroshenko GA, Motin VL, Nosov NY, Krasnov JM, Kukleva LM, Nikiforov KA, Al'khova ZV, Oglodin EG, Guseva NP: **Phylogeny and Classification of Yersinia pestis Through the Lens of Strains From the Plague Foci of Commonwealth of Independent States**. *Frontiers in Microbiology* 2018, **9**.
36. Eroshenko GA, Popov NV, Al'khova ZV, Kukleva LM, Balykova AN, Chervyakova NS, Naryshkina EA, Kutyrev VV: **Evolution and circulation of Yersinia pestis in the Northern Caspian and Northern Aral Sea regions in the 20th-21st centuries**. *PLOS ONE* 2021, **16**(2):e0244615.
37. Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S: **A fast adaptive algorithm for computing whole-genome homology maps**. *Bioinformatics* 2018, **34**(17):i748-i756.
38. Leiser OP, Merkley ED, Clowers BH, Deatherage Kaiser BL, Lin A, Hutchison JR, Melville AM, Wagner DM, Keim PS, Foster JT *et al*: **Investigation of Yersinia pestis Laboratory Adaptation through a Combined Genomics and Proteomics Approach**. *PLOS ONE* 2015, **10**(11):e0142997.
39. Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, Weinert LA, Wang Z, Guo Z, Xu L *et al*: **Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis**. *Proceedings of the National Academy of Sciences* 2013, **110**(2):577-582.
40. Vadyvaloo V, Martínez-Chavarría L: **Yersinia pestis and Yersinia pseudotuberculosis infection: a regulatory RNA perspective**. *Frontiers in Microbiology* 2015, **6**.
41. Whitehead MD, Burton HR, Bell PJ, Arnould JPY, Rounsevell DE: **A further contribution on the biology of the Antarctic flea, Glaciopsyllus antarcticus (Siphonaptera: Ceratophyllidae)**. *Polar Biology* 1991, **11**(6):379-383.
42. Nicoll W: **ON THE LENGTH OF LIFE OF THE RAT-FLEA APART FROM ITS HOST**. *Br Med J* 1912, **2**(2702):926-928.
43. Hastriter MW, Miller KB, Svenson GJ, Martin GJ, Whiting MF: **New record of a phoretic flea associated with earwigs (Dermaptera, Arixeniidae) and a redescription of the bat flea Lagaropsylla signata (Siphonaptera, Ischnopsyllidae)**. *Zookeys* 2017(657):67-79.
44. Biggins DE, Eads DA: **Prairie Dogs, Persistent Plague, Flocking Fleas, and Pernicious Positive Feedback**. *Frontiers in Veterinary Science* 2019, **6**.
45. Adams RA: **Do black-tailed prairie dog (Cynomys ludovicianus) colonies attract foraging bats?** *Journal of Zoology* 2021, **315**(2):156-163.