University of Waterloo

BIOL 469: Genomics

Prof. A. Doxey

Group 16 Final Project Report

An Investigation into the Pathogenicity of *Yersinia pestis*

Submitted by:

Ore Banjoko

Simonida Brankovic

Richard (Zhi Fei) Dong

Caitlin Anne Furgal

Srinija Palaparty

December 13th, 2021

1. Introduction

The human population has been periodically fighting pandemics since the very beginning of civilization, with the earliest record tracing as far back as 430 BC, during the Peloponnesian War (Schwartz and Kapila, 2021). Amongst all those pandemics, the plague arguably has the most profound impact on human societies and civilizations, killing millions of people across many countries and throughout history (Aberth, 2001).

The plague is caused by the bacterial infection of *Yersinia pestis*, which is carried by over 200 species of wild rodents (Anisimov and Amoako, 2006) and transmitted to humans through flea bites (primarily causing bubonic or septicaemic plague) or through inhalation of droplets (primarily causing pneumonic plague) (Davis, 2018). Three different strains of *Y. pestis* are responsible for the three major plague outbreaks throughout history: the strain *Antiqua* caused the Justinian's plague (AD 541 to 767), the strain *Mediaevalis* caused the Black Death (1346 to early 19th century), and the Strain *Orientalis* causes the modern plague (since 1894) (Chain et al., 2006; Song et al., 2004). The difference in the strains primarily lies in the ability of the bacterium to reduce nitrate and utilize glycerol (Song et al., 2004). Yet, for all the strains, the case-fatality ratio can range anywhere between 30% to as high as 100% if left untreated (Stenseth et al., 2008).

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* (Tan et al., 2015; Savin et al., 2019). However, *Y. pseudotuberculosis* and *Y. enterocolitica* are primarily spread via the ingestion of contaminated food and not spread by fleas (Galindo et al., 2011). 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* (Long et al., 2010; Marks et al., 1980). Differences between these species and *Y. pestis* may be crucial in uncovering the contributing factors of the increased pathogenicity seen in *Y. pestis*.

2. Objective

With improved sanitation and the development of antibiotics to treat infections (Riedel, 2017), 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. The number of countries reporting incidences of this disease is increasing and the plague has been attributed to thousands of deaths within the previous decade (Stenseth et al., 2008; Keeling and Gilligan, 2000). 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 (Tan et al., 2015). In response to these issues, the objective of this report is to determine the source of pathogenicity for the most recent modern plague-causing strain of *Y. pestis CO92* (*Orientalis*) at the genetic level. This will be accomplished through the comparison of the genome for *Y. pestis CO92* and other human pathogens (*Y. pseudotuberculosis* and *Y. enterocolitica*), as well as human non-pathogens (*Yersinia kristensenii* and *Yersinia ruckeri*) from the *Yersinia* genus. The results of this analysis can be used to understand the underlying mechanisms of *Y. pestis* pathogenicity so that more effective treatment methods and vaccines can be developed.

3. Methods

All the command lines and the Python program used for this section can be found in **Appendix B**.

3.1 Downloading the Genomes

The complete genome for *Y. pestis* (Strain CO92) was downloaded from GenBank (Clark et al., 2016) to be used as the reference genome. Then, complete genomes for *Y. pseudotuberculosis*, *Y. enterocolitica*, *Y. kristensenii*, and *Y. ruckeri* were also downloaded from GenBank for comparisons. The use of human non-pathogenic genomes from the *Yersinia* genus allowed for the elimination of genes that are only for housekeeping and non-disease-causing purposes. The use of other human pathogenic genomes from the *Yersinia* genus allowed for the determination of the genes that cause the ultra-high pathogenicity in *Y. pestis*. **Appendix A** lists the GenBank links for all genome files used.

3.2 Synteny and Evolutionary Relationship Analysis

The *Y. pestis* genome file and four other *Yersinia* species genome files (*Y. pseudotuberculosis*, *Y. enterocolitica*, *Y. kristensenii*, and *Y. ruckeri*) were input into Mauve software (Darling et al., 2004). Whole-genome alignments and synteny visualizations were performed in Mauve by using the “Align with progressiveMauve” option.

3.3 Gene Ontology (GO) Terms Analysis

The genome files used in 3.2 were input onto the online Google Cloud Linux server. Prokka (Seemann, 2014) was then used to annotate all genomes. GO terms were assigned to the annotation by utilizing the information stored in the General Feature Format (.gff) files produced by Prokka and UniProtKB (The UniProt Consortium, 2008). All five files containing the information about the GO terms and their frequencies for each genome were then run through a Python program to generate an Excel list that combined all the information. Key GO terms were searched on AmiGO to identify the functions (Ashburner et al., 2000).

3.4 Gene Set Comparisons

A gene list that removed the duplications for each genome was generated from the Prokka output table (.tbl) files from 3.3. Then, the list for *Y. pestis* was compared to each of the gene lists of other genomes to obtain the unique gene lists that contain the genes specific to *Y. pestis* when compared. Lastly, all computed unique gene lists were compared iteratively to produce a final list that contains the unique genes in *Y. pestis* when compared to all other species.

4. Results and Analysis

4.1 Synteny

Synteny mappings for *Y. pestis* against *Y. pseudotuberculosis* (**Figure 1a**), *Y. enterocolitica* (**Figure 1b**), *Y. ruckeri* (**Figure 1c**), and *Y. kristensenii* (**Figure 1d**) were compiled in Mauve. Since the genomes of *Yersinia* bacteria are circular (Eppinger et al., 2007), realigning the starting positions of the genome to be at the same place was essential in determining the true synteny. It was evident that *Y. pestis* shared the most synteny with *Y. pseudotuberculosis*. In addition, despite both being pathogenic to humans, after realigning the starting positions, significant chromosomal structural changes were evident in *Y. enterocolitica* compared to *Y. pestis*. In fact, the chromosomal structure for *Y. enterocolitica* was much more similar to that of *Y. kristensenii* (**Figure 2**). Amongst the species compared, *Y. pestis* shared the least synteny with *Y. ruckeri*.

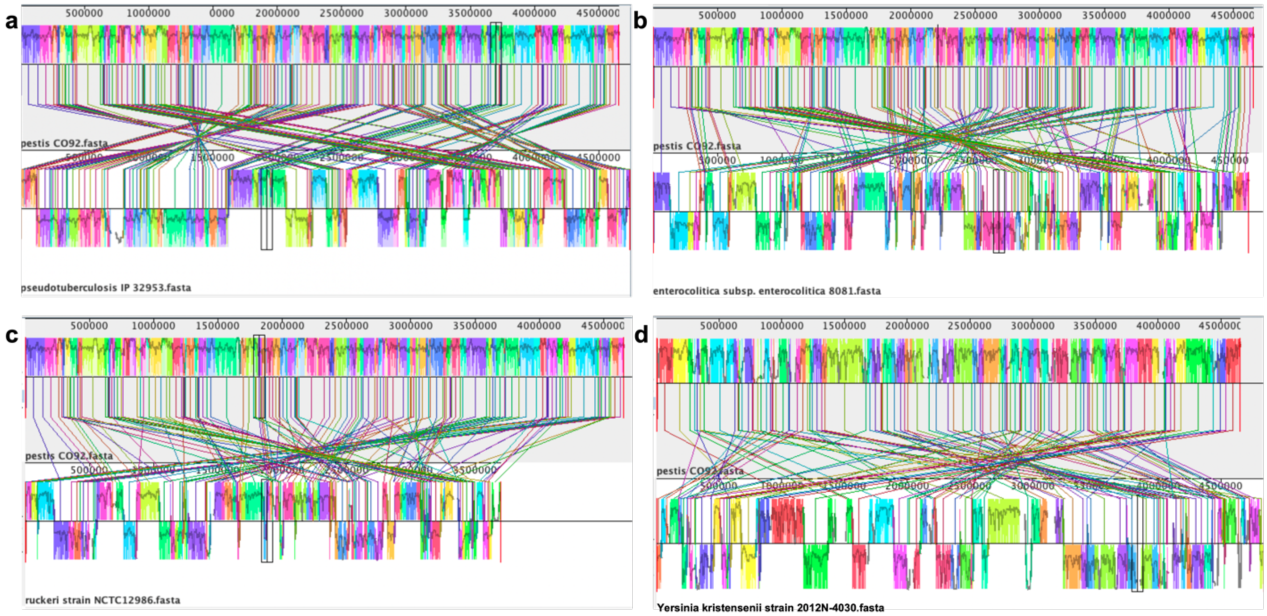

Figure 1: Synteny analysis of *Y. pestis* against a) *Y. pseudotuberculosis*, b) *Y. enterocolitica*, c) *Y. ruckeri*, and d) *Y. kristensenii* using Mauve

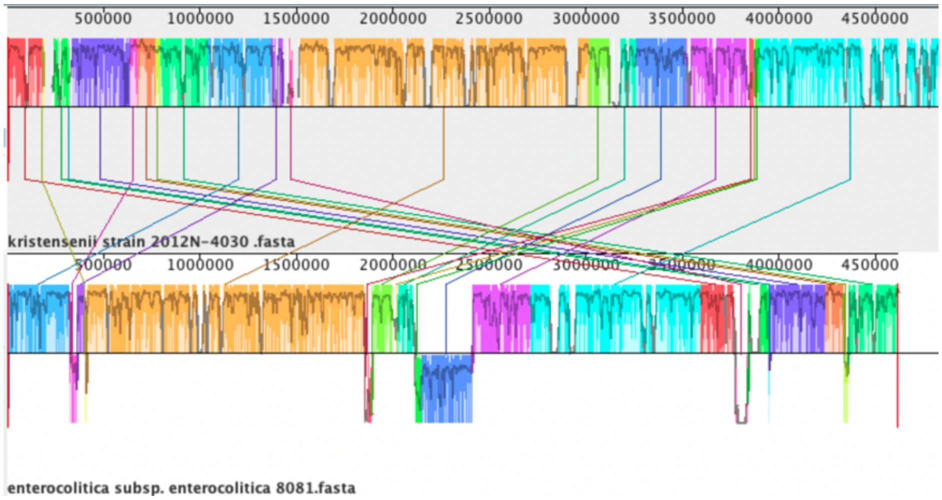

Figure 2: Synteny analysis of *Y. kristensenii* against *Y. enterocolitica* using Mauve

4.2 GO Terms

A GO terms analysis for all five species revealed that the top 20 GO terms for all species were fairly similar in function as well as in the number of genes associated with the function. Most of them were associated with essential house-keeping functions such as the formation of cell components, utilization of energy, and general gene translation and transcription (**Appendix C Table C1**). Therefore, a deeper analysis of the less frequent GO terms was conducted. Based on the synteny analysis in 4.1, close attention was paid to the unique GO profiles associated with *Y. pestis* but not with its closest relative, *Y. pseudotuberculosis*. It was found that most GO terms unique to *Y. pestis* from *Y. pseudotuberculosis* were also unique to all other *Yersinia* species compared, and some of those terms are highlighted in **Table 1**. A link to the complete Excel file containing all GO terms and their frequencies for all five species can be found in **Appendix D**.

Table 1: Some unique GO terms associated with *Y. pestis* but not with *Y. pseudotuberculosis*

GO Term	Function	*pestis*	*pseudotuberculosis*	*enterocolitica*	*kristensenii*	*ruckeri*
GO:1990216	positive regulation by symbiont of host transcription	1	0	0	0	0
GO:1902603	carnitine transmembrane transport	1	0	0	0	0
GO:1900751	4-(trimethylammonium)butanoate transport	1	0	0	0	0
GO:0085034	suppression by symbiont of host I-kappaB kinase/NF-kappaB cascade	1	0	0	0	0
GO:0052036	suppression by symbiont of host inflammatory response	1	0	0	0	0
GO:0051865	protein autoubiquitination	1	0	0	0	0
GO:0061630	ubiquitin-protein ligase activity	1	0	0	0	1
GO:0043424	protein histidine kinase binding	1	0	0	0	0
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	1	0	0	0	0
GO:0032238	adenosine transport	1	0	0	0	0
GO:0015879	carnitine transport	1	0	0	0	0
GO:0015864	pyrimidine nucleoside transport	1	0	0	0	0
GO:0015226	carnitine transmembrane transporter activity	1	0	0	0	0
GO:0015214	pyrimidine nucleoside transmembrane transporter activity	1	0	0	0	0
GO:0004549	tRNA-specific ribonuclease activity	1	0	0	0	0
GO:0050114	myo-inosose-2 dehydratase activity	1	0	1	1	0
GO:0044314	protein K27-linked ubiquitination	1	0	0	0	0
GO:0043424	protein histidine kinase binding	1	0	0	0	0
GO:0043214	ABC-type bacteriocin transporter activity	1	0	1	2	0
GO:0042930	enterobactin transport	1	0	1	2	1
GO:0042914	colicin transport	1	0	0	0	0
GO:0020002	host cell plasma membrane	1	0	0	0	1
GO:0015860	purine nucleoside transmembrane transport	1	0	0	1	0
GO:0001907	killing by symbiont of host cells	1	0	0	0	0

4.3 Gene Set Comparisons

Gene set comparisons were performed to identify the unique genes associated with *Y. pestis* when compared to *Y. pseudotuberculosis* as well as other *Yersinia* species. Similar to the result of GO terms, it was found that most of the unique genes of *Y. pestis* compared to *Y. pseudotuberculosis* were also unique when compared to other species. Overall, there were 39 genes unique to *Y. pestis* when compared to *Y. pseudotuberculosis* and 21 unique genes when compared to all four other *Yersinia* species. Those 21 genes unique to *Y. pestis* when compared to all species are *angR*, *caiT, cdiA2*, *ehaG*, *fepA*, *fliU*, *idhA*, *levD*, *mcbR*, *nupG*, *rhsB*, *tar*, *tibA*, *toxA*, *upaG*, *xylP*, *yagU*, *yenA2*, *yhfK*, *yihN,* and *yopM*, with the following additional genes being unique only when compared to *Y. pseudotuberculosis*: *bin3*, *bvgS*, *caf1*, *dinJ*, *higB2*, *ilvN*, *intQ, lagD*, *noc*, *ompD*, *parA*, *pld*, *repB*, *rop*, *tfaE*, *virB*, *xni*, and *yhdJ*.

5. Discussion

5.1 Synteny Analysis

The goal of the synteny analysis was to determine the evolutionary relationship in order to identify the comparisons that require particular attention. From the synteny mapping, it was evident that *Y. pestis* is closely related to *Y. pseudotuberculosis* due to a large amount of shared synteny. *Y. enterocolitica* and *Y. kristensenii* should be closely related as well due to similar reasons, despite *Y. enterocolitica* being pathogenic to humans while *Y. kristensenii* is not. This predicted phylogenetic relationship is confirmed by multiple pieces of literature, where the phylogenetic tree of the *Yersinia* family (**Figure 3**) indicates that *Y. enterocolitica* is evolved independently from *Y. pseudotuberculosis*, which later undergoes speciation to produce *Y. pestis* (Savin et al., 2019; Tan et al., 2015). Therefore, this synteny reveals that the pathogenicity of *Y. pestis* is primarily associated with the change of gene functions in *Y. pseudotuberculosis* rather than *Y. enterocolitica*.

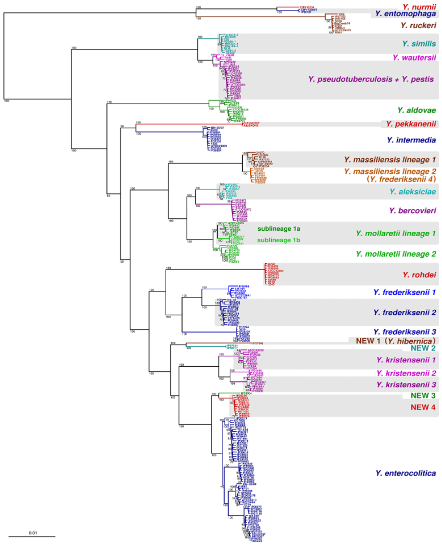

Figure 3: A Maximum-Likelihood tree of the genus *Yersinia*. Reproduced from Savin et al. with permission from Microbiology Society

5.2 GO Terms Analysis

The goal of GO terms analysis was to look for unique gene functions that were present in *Y. pestis*, which might infer pathogenicity. Since the majority of the unique GO terms for *Y. pestis* compared to *Y. pseudotuberculosis* were also present in other *Yersinia* species, this suggests that most of those unique GO terms are likely to be associated with the increase in pathogenicity in *Y. pestis* since they are less likely to be responsible for general house-keeping functions. Due to the length limitation of the report, only one type of the unique GO terms will be investigated in-depth: processes involving ubiquitination, which have four unique occurrences, three of which are associated only with *Y. pestis*. Ubiquitination occurs when ubiquitin (Ub) attaches to proteins, which triggers a variety of changes in protein functions (Amemiya et al., 2010; Li et al., 2016). In *Y. pestis*, the protein responsible for this is the *Yersinia* outer protein (Yop) M E3 Ub ligase. It is found that YopM can associate with NLRP3, a component of the host's innate immune system, and ubiquitinate it to induce cell death and necrosis in the later stage of the infection, which increases inflammation and can lead to sepsis (Wei et al., 2016).

5.3 Gene Set Comparisons

Despite GO terms comparisons being a very powerful technique in identifying novel functions, when it comes to the new genes associated with existing functions (such as gene transcription), it is difficult to identify these genes through the sole use of GO terms. As a result, gene set comparisons were computed to uncover unique genes that might have been veiled under pure GO terms analysis. Again, due to the length limitation, only one particular gene will be discussed in detail: the *caf1* gene, which is additionally present only in *Y. ruckeri*. *caf1* is associated with the GO term of cell adhesion (GO:0007155). This GO term was found in all five genomes with numerous quantities. When expressed, *caf1* produces the capsular antigen F1, a dimer that attaches to the IL-1 receptors of human epithelial cells and macrophages (Abramove et al., 2002). This attachment inhibits phagocytosis and allows *Y. pestis* to evade the immune response, which contributes to the increased pathogenicity when compared to *Y. pseudotuberculosis* and other *Yersinia* species (Al-Jawdah et al., 2019).

6. Conclusion and Future Perspectives

The pathogenicity of the plague causing bacterium *Y. pestis* at the genetic level was investigated through synteny analysis, GO terms analysis, and gene set comparisons. It was found that *Y. pestis* is closely related to *Y. pseudotuberculosis*, and the unique GO term associated with ubiquitination and gene associated with host cell interaction can contribute to the cause of the high pathogenicity of *Y. pestis*. It is hoped that this information can be beneficial in creating novel treatments and potential vaccines for *Y. pestis* infections to address the public health issue of the modern plague. There is so much more analysis that can be done based on the results obtained in this study. Therefore, in the future, with more time and paragraph spaces to work with, a full analysis should be conducted to investigate all the unique GO terms and genes of *Y. pestis* when compared to *Y. pseudotuberculosis* and other *Yersinia* relatives in order to obtain the full picture of the pathogenicity of this species at the genetic level.

References

Aberth, John (2001). From the Brink of the Apocalypse: Confronting Famine, War, Plague and Death in the Later Middle Ages (second ed.). Routledge. ISBN 978-1134724802

Abramov, V. M., Vasiliev, A. M., Khlebnikov, V. S., Vasilenko, R. N., Kulikova, N. L., Kosarev, I. V., Ishchenko, A. T., Gillespie, J. R., Millett, I. S., Fink, A. L., & Uversky, V. N. (2002). Structural and functional properties of Yersinia pestis Caf1 capsular antigen and their possible role in fulminant development of primary pneumonic plague. J*ournal of proteome research*, 1(4), 307–315. https://doi.org/10.1021/pr025511u

Al-Jawdah, A. D., Ivanova, I. G., Waller, H., Perkins, N. D., Lakey, J. H., & Peters, D. T. (2019). Induction of the immunoprotective coat of yersinia pestis at body temperature is mediated by the CAF1R transcription factor. *BMC Microbiology*, 19(1). https://doi.org/10.1186/s12866-019-1444-4

Amemiya, Y., Azmi, P., & Seth, A. (2008). Autoubiquitination of BCA2 RING E3 ligase regulates its own stability and affects cell migration. M*olecular cancer research: MCR*, 6(9), 1385–1396. https://doi.org/10.1158/1541-7786.MCR-08-0094

Anisimov, A. P., & Amoako, K. K. (2006). Treatment of plague: Promising alternatives to antibiotics. Journal of Medical Microbi*ology*, 55(11), 1461–1475. https://doi.org/10.1099/jmm.0.46697-0

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), 25–29. https://doi.org/10.1038/75556

Chain, P. S., Hu, P., Malfatti, S. A., Radnedge, L., Larimer, F., Vergez, L. M., Worsham, P., Chu, M. C., & Andersen, G. L. (2006). Complete genome sequence of *Yersinia pestis* strains *Antiqua* and Nepal516: evidence of gene reduction in an emerging pathogen. *Journal of bacteriology*, 188(12), 4453–4463. https://doi.org/10.1128/JB.00124-06

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic acids research*, 44(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7), 1394–1403. https://doi.org/10.1101/gr.2289704

Davis, K.M. (2018). All Yersinia Are Not Created Equal: Phenotypic Adaptation to Distinct Niches Within Mammalian Tissues. *Front. Cell. Infect. Microbiol*. 8(261). https://doi.org/10.3389/fcimb.2018.00261

Eppinger, M., Rosovitz, M. J., Fricke, W. F., Rasko, D. A., Kokorina, G., Fayolle, C., Lindler, L. E., Carniel, E., & Ravel, J. (2007). The complete genome sequence of *Yersinia pseudotuberculosis IP31758*, the causative agent of Far East scarlet-like fever. *PLoS genetics*, 3(8), e142. https://doi.org/10.1371/journal.pgen.0030142

Galindo, C. L., Rosenzweig, J. A., Kirtley, M. L., & Chopra, A. K. (2011). Pathogenesis of Y. enterocolitica and Y. pseudotuberculosis in Human Yersiniosis. *Journal of pathogens*, 2011, 182051. https://doi.org/10.4061/2011/182051

Keeling M. J. & Gilligan C. A. (2000). Bubonic plague: a metapopulation model of a zoonosis. *Proc. R. Soc. Lond*., B(267), 2219–2230. http://doi.org/10.1098/rspb.2000.1272

Li, J., Chai, Q. Y., & Liu, C. H. (2016). The ubiquitin system: a critical regulator of innate immunity and pathogen-host interactions. *Cellular & molecular immunology*, 13(5), 560–576. https://doi.org/10.1038/cmi.2016.40

Long, C., Jones, T. F., Vugia, D. J., Scheftel, J., Strockbine, N., Ryan, P., Shiferaw, B., Tauxe, R. V., & Gould, L. H. (2010). Yersinia pseudotuberculosis and Y. enterocolitica infections, FoodNet, 1996-2007. *Emerging infectious diseases*, 16(3), 566–567. https://doi.org/10.3201/eid1603.091106

Marks, M. I., Pai, C. H., Lafleur, L., Lackman, L., & Hammerberg, O. (1980). *Yersinia enterocolitica* gastroenteritis: A prospective study of clinical, bacteriologic, and epidemiologic features. *The Journal of Pediatrics*, 96(1), 26–31. https://doi.org/10.1016/s0022-3476(80)80318-0

Riedel, S. (2005). Plague: From Natural Disease to Bioterrorism. *Baylor University Medical Center Proceedings*, 18(2), 116-124. https://doi.org/10.1080/08998280.2005.11928049

Savin, C., Criscuolo, A., Guglielmini, J., Le Guern, A. S., Carniel, E., Pizarro-Cerdá, J., & Brisse, S. (2019). Genus-wide *Yersinia* core-genome multilocus sequence typing for species identification and strain characterization. *Microbial genomics*, 5(10), e000301. https://doi.org/10.1099/mgen.0.000301

Schwartz, R. A., & Kapila, R. (2021). Pandemics throughout the centuries. *Clinics in Dermatology*, 39(1), 5–8. https://doi.org/10.1016/j.clindermatol.2020.12.006

Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Song, Y., Tong, Z., Wang, J., Wang, L., Guo, Z., Han, Y., Zhang, J., Pei, D., Zhou, D., Qin, H., Pang, X., Han, Y., Zhai, J., Li, M., Cui, B., Qi, Z., Jin, L., Dai, R., Chen, F., Li, S., … Yang, R. (2004). Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA research: an international journal for rapid publication of reports on genes and genomes*, 11(3), 179–197. https://doi.org/10.1093/dnares/11.3.179

Stenseth, N. C., Atshabar, B. B., Begon, M., Belmain, S. R., Bertherat, E., Carniel, E., Gage, K. L., Leirs, H., & Rahalison, L. (2008). Plague: past, present, and future. *PLoS medicine*, 5(1), e3. https://doi.org/10.1371/journal.pmed.0050003

Tan, S. Y., Dutta, A., Jakubovics, N. S., Ang, M. Y., Siow, C. C., Mutha, N. V. R., Heydari, H., Wee, W. Y., Wong, G. J., & Choo, S. W. (2015). Yersiniabase: A genomic resource and analysis platform for comparative analysis of yersinia. *BMC Bioinformatics*, 16(1). https://doi.org/10.1186/s12859-014-0422-y

UniProt Consortium (2008). The universal protein resource (UniProt). *Nucleic acids research*, 36(Database issue), D190–D195. https://doi.org/10.1093/nar/gkm895

Wei, C., Wang, Y., Du, Z., Guan, K., Cao, Y., Yang, H., Zhou, P., Wu, F., Chen, J., Wang, P., Zheng, Z., Zhang, P., Zhang, Y., Ma, S., Yang, R., Zhong, H., & He, X. (2016). The *Yersinia* Type III secretion effector YopM Is an E3 ubiquitin ligase that induced necrotic cell death by targeting NLRP3. *Cell death & disease*, 7(12), e2519. https://doi.org/10.1038/cddis.2016.413

Appendix A: List of all genome files used

***Yersinia pestis* CO92 chromosome, complete genome:**

https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP009973.1

***Yersinia pseudotuberculosis* IP 32953 strain IP32953 chromosome, complete genome:**

https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP009712.1

***Yersinia enterocolitica* subsp. enterocolitica 8081, complete sequence:**

https://www.ncbi.nlm.nih.gov/nuccore/NC_008800.1

***Yersinia ruckeri* strain NCTC12986, whole genome shotgun sequence:**

https://www.ncbi.nlm.nih.gov/nuccore/NZ_UHJF01000001.1

***Yersinia kristensenii* strain 2012N-4030 chromosome, complete genome:**

https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP054049.1

Appendix B: Command codes and the python program used

3.3 GO Terms Analysis

Annotation using Prokka

```
prokka genome.fasta
```

where *genome.fasta* is the name of the genome file.

GO terms assignment

```
cat PROKKAout.gff | grep -o "UniProtKB.*;" | awk -F'[:;=]' '{print $4" "$2}'>uniProts.txt
uniprot2go.py -i uniProts.txt -d /data/uniprot2go/uniprot-vs-go-db.sl3 >go.annotations
cat go.annotations | awk '{print $3}' | tr "," "\n" | sort | uniq -c | sort -n -r >GOterms.txt
```

where *PROKKAout.gff* is the .gff output of Prokka. *GOterms.txt* is the final file containing all the GO terms and the corresponding frequencies of each genome.

Combining GO terms for all five genomes

Please refer to the python program written and used for this part at:

https://drive.google.com/file/d/1s96HRW2YM9LcZUgRbPugoKND0QIzvOOF/view?usp=sharing

3.4 Gene Set comparisons

Obtaining unduplicated gene lists

```
cat Prokka_xx.tbl | awk '{if ($1 == "gene") {print $2}}' | awk -F'_' '{print $1}' | sort >list.txt
uniq list.txt > uniq_list_xx.txt
```

where *uniq_list_xx.txt* is the list of unique genes for the genome of xx.

Obtaining unique gene lists

```
comm uniq_list_pestis.txt uniqe_list_xx.txt > gene_comp_xx.txt
cat gene_comp_xx.txt | awk -F'\t' '{print $1}' | grep -v -e '^$' >gene_comp_xx_final.txt
```

where *gene_comp_xx_final.txt* is the unique gene list when comparing the gene list of *Y. pestis* to that of xx.

Obtaining the final gene list

```
comm gene_comp_xx_final.txt gene_comp_yy_final.txt > gene_comp_xxyy.txt
cat gene_comp_xxyy.txt |awk -F'\t' '{print $3}'| grep -v -e '^$' >gene_comp_xxyy_final.txt
```

where *gene_comp_xxyy_final.txt* is the unique gene list when the gene list of *Y. pestis* is compared to both xx and yy. This process is repeated four times to obtain the final gene list.

Appendix C Table C1: Top 20 GO terms for the five *Yersinia* species

GO Term	Function	***pestis***	***pseudotuberculosis***	***enterocolitica***	***kristensenii***	***ruckeri***
GO:0016020	membrane	933	923	977	1024	790
GO:0005886	plasma membrane	798	798	859	897	689
GO:0005515	protein binding	778	772	814	818	739
GO:0005829	cytosol	735	740	781	779	722
GO:0005737	cytoplasm	735	733	744	742	680
GO:0016021	integral component of membrane	654	646	709	737	569
GO:0046872	metal ion binding	513	516	543	560	477
GO:0000166	nucleotide binding	486	484	496	504	422
GO:0016740	transferase activity	457	455	497	505	421
GO:0005524	ATP binding	414	414	422	429	357
GO:0016787	hydrolase activity	341	340	377	377	319
GO:0005887	integral component of plasma membrane	323	329	359	374	306
GO:0055085	transmembrane transport	303	305	315	336	221
GO:0003824	catalytic activity	299	298	306	312	266
GO:0042802	identical protein binding	282	283	297	305	268
GO:0003677	DNA binding	276	283	321	331	249
GO:0016491	oxidoreductase activity	248	251	256	267	203
GO:0006355	regulation of transcription, DNA-templated	168	170	202	213	146
GO:0022857	transmembrane transporter activity	161	165	181	195	113
GO:0000287	magnesium ion binding	147	146	156	155	137

Appendix D: Complete list for all GO terms associated with the five species

Please refer to the following Google Drive link for the complete Excel file:

https://docs.google.com/spreadsheets/d/e/2PACX-1vTgAhGItXummyiAKp_IsTeMWDcmW7vXaNabThoGGtAQDVaYE3x-iH8JVivkSFp3Lw/pub?output=xlsx