**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: DocuSigned by: Brianna Jeanne Bixler E1646792BA344CE...

Brianna Jeanne Bixler
Name

12/11/2023 | 11:37 AM EST
Date

| | |
|---|---|
| **Title** | The nature of co-infection during outbreaks of infectious disease |
| **Author** | Brianna Jeanne Bixler |
| **Degree** | Doctor of Philosophy |
| **Program** | Biological and Biomedical Sciences<br>Genetics and Molecular Biology |

**Approved by the Committee**

DocuSigned by:
Timothy Read
2A4D4A4362BB4E0...

Timothy Read
*Advisor*

DocuSigned by:
Anne Piantadosi
45334A7CE7C442C...

Anne Piantadosi
*Advisor*

DocuSigned by:
Anke Huels
6FEC7C88FBB3492...

Anke Huels
*Committee Member*

DocuSigned by:
Karen Conneely
C72ED0D54A5E4C3...

Karen Conneely
*Committee Member*

DocuSigned by:
83A77F01615D4D8...

David Katz
*Committee Member*

*Committee Member*

**Accepted by the Laney Graduate School:**

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

Date

# The nature of co-infection during outbreaks of infectious disease

By

Brianna Jeanne Bixler,  BS

Advisor: Tim Read, Ph.D. and Anne Piantadosi Ph.D./MD

An abstract of
A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of
Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Genetics and Molecular Biology 2023

# Abstract

In light of recent technological advancements that have decreased the cost of sequencing, we have seen a massive increase in publicly available datasets. This increase in quantity has led to new approaches to be implemented when solving epidemiological studies. With this, I have created computational pipelines to process and leverage publicly available datasets from around the world to understand the prevalence and impact of co-infections in the context of this wealth of data.

In my first research project, I focused on an outbreak of H. influenza (Hi), a pathogen that typically causes respiratory infections but in this case, was causing severe septic arthritis in individuals who were also HIV+. It is not clear whether this unusual clinical presentation was from a change introduction into a vulnerable population of HIV+ individuals, or if there is a genetic feature of these particular strains that causes increased virulence. In this study, we performed a comparative genomic analysis of the clinical isolates originally identified in metropolitan Atlanta in the context of the larger pangenome of over 4,000 *Hi* strains to identify potential features that may suggest enhanced virulence in the cluster strains.

In the second research project, I built a computational pipeline to process over 800 metatranscriptomic samples collected from individuals with COVID-19. After establishing this pipeline, we can ask basic epidemiological questions. With the output of this analysis, we are able to assess if there are any co-infections of viral, bacterial, or fungal pathogens in the nasal cavity at the onset of COVID-19. To understand the effect of co-infection, we looked for a correlation to viral burden and found that none of the pathogens seemed to correlate with an increased COVID-19 viral load.

To conclude, this dissertation will discuss aspects of co-infection in the larger context of thousands of isolates, a scale that is unprecedented for these pathogens. From this scale, we can determine what is unique and what remains common. We generate multiple hypotheses of what co-infections exist and how they may impact public health.

# The nature of co-infection during outbreaks of infectious disease

By

Brianna Jeanne Bixler,  BS

Advisor: Tim Read, Ph.D. and Anne Piantadosi Ph.D./MD

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of
Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Genetics and Molecular Biology 2023

# Acknowledgments

The biggest thank you to my mentors, Tim Read and Anne Piantadosi. You have taught me so much, science and life-skill alike. I am so grateful for your unconditional support, when experiments were going well and when there was a lull in work productivity. Should I ever find myself in a position to mentor people, I hope to emulate you both.

I'd also like to thank my mentors who came before Emory. Taylor Wahlig, you were a huge part of my early growth as a scientist and someone that I am happy to call a friend. Piroska Szabó, thank you so much for taking a chance on me and giving me the support to develop a project under your expert guidance. I will always have a love for developmental biology because of you. Elliott Sober, thank you for teaching me the philosophy of biology and being someone who I could unapologetically try to think about new ideas without the fear of judgment. I am so grateful for your support throughout my career.

I'd like to thank my friends and other students at Emory. You have become a second family to me over the years and I could not imagine going through this process without you. I'd also like to extend the warmest thank you for the support and guidance I have received from Heather Kotler and Meleah Hickman; it takes a village and I am so happy you are part of it.

And finally, I'd like to thank my family. Mom and Dad, you have been instrumental in the development of my values, work ethic, and curiosity. I would not have made it through grad school without these. Thank you to my sisters, Amanda and Claudia, for your unconditional support and love. Thank you to our partners, Marcus, Jake, and Hannah for helping us create our own little family units. And of course to my cats, Olive and Junebug, for ensuring there is never a dull moment. And to my grandparents, Grandpa Al and Grandma Yo. Thank you for always offering so much love, undivided attention when I would ramble about what was new, and unfaltering support through the hard times too. I would give anything to be able to share this day with you both.

I'd also like to recognize that this research was done in offices and laboratories located on the ancestral and unceded traditional territories of the Muscogee/Creek Peoples.

# Table of contents

# Chapter 1: Introduction

## BACKGROUND

I have been fortunate to have two projects that have access to sequencing directly from patients who are affected by ongoing outbreaks in Atlanta. Having access to these patient samples is a privilege and I am grateful to have been able to work with data that is so directly linked to our community. With this data and the relevant metadata, we can answer basic questions of epidemiology which establishes an outbreak and the causative agent. For both of my projects, this epidemiological work has been done before me. From there, we can transition to a genomic perspective of disease, leveraging the massive amount of data that sequencing patient samples provides us.  As a foundation, I will talk about the basics of epidemiology in order to understand how genomic sequencing is revolutionizing the field.

### Classic epidemiology

Epidemiology is the study of the incidence, frequency, pattern, and determinants of diseases and other factors relating to health in a population over time [1]. Epidemiological observations can be traced back to ancient civilizations, typically starting with the father of modern medicine, Hippocrates, in about 400 BC [2]. Historians credit John Graunt, a 17th-century demographer, with the birth of modern epidemiology [3]. Graunt was a largely self-educated person who pioneered epidemiological analyses that placed a lot of value on the accuracy and nuanced quantifications of public health [3]. In the 19th century, notable advancements in epidemiology were made by the work of William Farr and John Snow [44–64]. There is a lore surrounding the underdog of the time, John Snow, and how he discovered that the cholera outbreak of 1849 was fueled by a contaminated water supply [4–6].  This laid the groundwork for modern epidemiological methods and principles.

Scientists such as John Snow, William Farr, and many other scientists have established the process of conducting and managing disease outbreak investigations [7]. This process starts with establishing the existence of an outbreak and verifying the diagnosis of the causative pathogen. Neither of these are trivial. After the outbreak is established, scientists must identify relevant cases to perform descriptive epidemiology. At this stage, they must set criteria for inclusion, collect information on each case, and describe the affected people, place, and time. Once this foundation is laid, scientists can develop a hypothesis of what the source of the pathogen is and the vector responsible. Then we are left with the work that is common to all fields of science, which is to evaluate hypotheses and communicate those results. This workflow has been instrumental in identifying risk factors, understanding disease transmission, and informing public health interventions to prevent and control diseases. These historical and classic techniques have laid the foundation for modern epidemiological research and continue to be used alongside newer methods of molecular and genome-based epidemiology.

Much of the classic epidemiology for Chapter 2 and Chapter 3 of this dissertation was established and has allowed me to focus on genome-based epidemiology questions. The epidemiological foundation for Chapter 2,  was established by a paper published by Collins et al. 2019 “Invasive Nontypeable Haemophilus influenzae Infection Among Adults With HIV in Metropolitan Atlanta, Georgia, 2008-2018” [8]. These researchers set criteria for inclusion and information on each case collected, an outbreak was discovered and the place, people affected, and the timeline of this outbreak was established. They established that there was an outbreak of Nontypeable Haemophilus (NTHi) that disproportionately affected HIV+ men who have sex with men living in metro-Atlanta. With the observation of genetic similarity and the close geospatial location of the infections, a novel mode of transmission was discovered. With the outbreak discovered, novel transmissibility described, and causative pathogen isolated, the genomic information of those NTHi infections was sequenced which marked the beginning of my project. In Chapter 3, I investigate the potential of co-infection in patients infected with COVID-19, which was already established as a pandemic by the samples being collected[9].

## Molecular Epidemiology

Molecular epidemiology is the use of molecular laboratory techniques to answer epidemiological questions [10]. With the inclusion of these techniques, the field of epidemiology expands to include the study of how molecular pathways, metabolites, or novel genes impact the risk of disease development[11]. This is a broad field, encompassing many advancements in technology. For example, in the quest to establish an outbreak and verify diagnoses of the causative pathogen, PCR-based and serology-based diagnostics have proven to be instrumental[10,12]. Another way molecular techniques have broadened epidemiology is in understanding biomarkers of disease risk, such as increased cholesterol as a biomarker for heart disease[11]. The genre of molecular epidemiology that my dissertation focuses on is genome-based epidemiology.

## Genome-based epidemiology

In recent decades, next-generation DNA sequencing has emerged as a tool that has revolutionized the field of infectious disease[13]. It has transformed how infections are discovered, understood, diagnosed, and treated. Before the development of this technology, we were limited by challenging and time-consuming benchwork experiments required to isolate, culture, and identify pathogens collected from affected people[14]. While these methods are still valuable for the advancement of the field, they are not as scalable and precise as next-generation sequencing. The wide application of DNA sequencing in epidemiology has also given us a basis to understand the evolution and transmission patterns of pathogens by following the molecular signature of a population of samples[13].

These methods rely on next-generation sequencing, which largely replaced earlier methods in the 2000s[15]. In 2000 Lynx Therapeutics, later bought by Illumina, rolled out the first next-generation sequencing technology called Massively parallel signature sequencing. By 2004, 454 Life Sciences marketed a pyrosequencing technology that could produce up to 20 million base pairs. With this groundwork and a lot of money, time, and resources in 2008 the first paper studying the human genome using next-generation sequencing was published[16]. Although these were all very exciting and important milestones, next-generation sequencing was not accessible to most scientists because of its prohibitive costs. It was not until 2014 with the new HiSeq X Ten Sequencer available through Illumina did the cost of a genome dropp to $1000. With this milestone, we observed a massive uptick in genomes available. Next-generation sequencing gave way to DNA, RNA, and eventually metagenomic sequencing to become commonplace.

Sequencing pathogens in the context of an outbreak gives researchers the power to answer a number of questions that would not be possible with other identification methods[13]. DNA sequencing gives researchers information far beyond the organism's identity because most, if not all, of the genome is defined by this method. This gives researchers access to all of the genes and regulatory elements in the organism's genome. In the context of more than one genome, researchers can start to discern what makes the pathogen's genome unique and which of these changes could lead to increased virulence, altered transmission, or any other biological factors of interest[17]. As the databases of sequence information from infection increases, the complexity of questions and the significance of our conclusions tend to increase, allowing for an accurate and holistic picture of an outbreak[18]. This is largely the strength of our NTHi study which I will discuss in the second chapter of this dissertation, in the wide comparison to thousands of other samples collected across decades and around the world.

## Applications of DNA sequencing in Molecular Epidemiology

Building on the impact of next-generation DNA sequencing in the field of infectious disease, it is essential to consider the practical applications this technology has made possible. One area where DNA sequencing has significantly contributed to our understanding of infectious diseases is in the realm of diagnostics. Unlike traditional diagnostic methods that often rely on isolating and culturing pathogens, which could be time-consuming and yield inconclusive results, DNA sequencing offers a more efficient alternative.  This technology enables the rapid and accurate identification of infectious agents directly from clinical samples. Notably, it not only expedites the diagnostic process but also empowers us to detect previously unknown or emerging pathogens, a capability of utmost importance in addressing emerging infectious diseases. We observed the power of this technique in the identification of the COVID-19 pathogen as it emerged in 2019.

The advent of metagenomics further exemplifies the transformative potential of DNA sequencing in diagnostics. With metagenomics, we are no longer constrained by the need to sequence isolated cultures. Instead, we can sequence the entire microbial landscape present on the isolating swab, encompassing bacteria, viruses, and fungal organisms. This technological advancement forms the cornerstone of my work in the COVID-19 study, which constitutes the third chapter of this dissertation.

The application of DNA sequencing also extends to parts of the genome that do not necessarily involve the identification of the pathogen. Examples of this could be understanding the state of antibiotic resistance or the development of vaccines in response to the protein structures visible to our immune systems. Understanding the genetic makeup of pathogens at a granular level enables the identification of antigenic targets for vaccine development. This knowledge empowers scientists to design vaccines that are more effective and can be developed more rapidly, a vital capability when responding to outbreaks of infectious diseases. The development of the COVID-19 RNA vaccine is a perfect example of how critical the application of DNA sequencing can be.

## Genomic surveillance

The genomic surveillance that is the foundation of the work done in the NTHi study in chapter 2 of this dissertation was done by the Active Bacterial Core surveillance[8,19]. This group actively surveys for a set of invasive pathogens of interest, setting up the infrastructure to study many epidemiological questions.

Genomic surveillance is an infrastructure to monitor the genetic material of pathogens, tracking the transmission, evolution, and impact[20]. In 2022 the Whole Health Organization (WHO) put out a 10-year plan to implement increased genomic surveillance[21,22]. This involves tracking cases of infection, sequencing select samples that represent a portion of the population, and using this information to answer epidemiological questions. Having this infrastructure in place is invaluable to identifying outbreaks and real-time tracking of pathogen spread, helping to inform public health interventions [20].

The Active Bacterial Core surveillance is a component of the CDC's Emerging Infections Programs which sets up collaborations between clinicians, university researchers, and health departments[19]. They have a set of pathogens that they actively survey for, including *Haemophilus influenzae*, group A *Streptococcus*, group B *Streptococcus*, *Neisseria meningitidis*, and *Streptococcus pneumoniae*. The CDC uses the ABC surveillance group to track disease trends and inform public health policy, which is how it has contributed to the identification of this outbreak and the epidemiological foundation of this project.

## Bacterial Pangenomes

As the number of DNA sequencing experiments in the context of disease outbreaks increases and this information is added to public databases, the power and depth of conclusions we can make in the context of this information grows[18]. One of the approaches to come out of this wealth of information is a pangenome analysis[23–25]. Pangenomes refer to the complete set of genes present in a species, including the core genome, which are genes shared by all strains, and the accessory genome, which are genes present in only some strains[23]. Unlike traditional genomics, which focuses on a single reference genome, pangenome analysis takes into account the genetic diversity within a species by analyzing multiple genomes.  This large-scale comparative genomics allows researchers to identify commonalities and variations among pathogen genomes giving a more comprehensive understanding of how this outbreak compares to other individual infections or even how it compares to other outbreaks in the data available in our public databases. These insights are invaluable for predicting disease trends, understanding the emergence of drug resistance, and designing public health interventions. It also allows for a comprehensive picture of niche adaptations and evolutionary trends as they arise in the population.

Pangenomes offer a technique to understand the components that are common or unique among the population and how this changes over time, which lends itself to answering biological which genes in the accessory genome could be contributing to niche adaptations, such as the rise antibiotic resistance or virulence factors[26,27]. Pangenomics has revolutionized our understanding of pathogen evolution by revealing the genetic variations that contribute to their diversity and adaptation. By comparing the pangenomes of different strains, researchers can identify genes that are gained or lost during evolution, providing insights into the acquisition of virulence factors or drug resistance mechanisms[26].

Pangenomic analysis allows for the identification of virulence factors, which are genetic elements that enhance the ability of pathogens to cause disease[27]. By comparing the accessory genomes of pathogenic and non-pathogenic strains, researchers can pinpoint the specific genes or genetic variations associated with increased virulence*.* This information is invaluable for understanding the mechanisms underlying pathogenesis. In a similar way, pangenomic analysis has also significantly advanced our understanding of drug resistance in infectious diseases. Again by comparing the pangenomes of drug-resistant and drug-susceptible strains, researchers can identify genetic variations associated with resistance[28]. Pangenomic analysis has been particularly instrumental in studying drug resistance in bacteria, such as methicillin-resistant *Staphylococcus aureus* (MRSA)[29,30].

We take advantage of this pangenome approach in our NTHi study described in Chapter 2, which creates a local database of information on nearly 4,000 samples from NTHi infections across the world. In the next section of this introduction, I will expand on the biology of NTHi and what is known to date about this pathogen in order to create a foundation to understand the novelty of the outbreak our research group discovered and I have been fortunate enough to study alongside them.

## Diverse presentation of Hflu

*Haemophilus influenzae* (H. influenzae) gram-negative bacterium that colonizes the upper respiratory tract of humans[31]. In 1892 Richard Pfeiffer isolated *H. influenzae* from patients during an outbreak of influenza, believing it was the causative agent[32]. This is where the bacteria got its name which is a bit of a misnomer, given that it is not a virus. Despite this misstep, Pfeiffer was instrumental in its discovery and early understanding of the pathogen. As *H. influenzae* was established as a bacterial pathogen and further characterized, it was divided into two groups, one defined by the presence of an outer polysaccharide capsule, encapsulate *H. influenzae*, or the absence of it, non-typeable *H. influenzae* (NTHi)[31].

Encapsulated *H. influenzae* depends on its capsule to evade the host immune response and is typically responsible for more severe clinical presentation[33]. Clinical presentations of encapsulated H. influenzae infections can vary depending on the specific strain and the patient's age and overall health. Encapsulated H. influenzae, particularly type b (Hib), is known for causing severe diseases, especially in young children[34]. Common clinical presentations include invasive diseases like meningitis, bacteremia, and pneumonia. Meningitis caused by Hib can lead to symptoms such as high fever, severe headaches, neck stiffness, and altered mental status. Bacteremia, the presence of bacteria in the bloodstream, may manifest with symptoms like fever, chills, and low energy[35]. Hib pneumonia can result in respiratory distress, cough, and fever. In children, these infections can progress rapidly and be life-threatening, underscoring the importance of vaccination against Hib to prevent these serious clinical manifestations. Prior to the introduction of the H. influenzae type b (Hib) vaccine, Hib was the leading cause of bacterial meningitis in children under five years of age[36]. However, the widespread use of the Hib vaccine has significantly reduced the incidence of invasive Hib disease. NTHi, on the other hand, remains a common cause of respiratory tract infections in both children and adults.

Non-typeable *Haemophilus influenzae* (NTHi) infections present a spectrum of clinical manifestations that primarily target the respiratory tract and present in sporadic infections in the very young, elderly, or immunocompromised patients[37]. Unlike encapsulated H. influenzae strains, NTHi lacks a protective polysaccharide capsule, making it less virulent but still capable of causing various illnesses[38]. Clinical presentations of NTHi infections often include non-invasive conditions such as otitis media (middle ear infections), sinusitis, and exacerbations of chronic obstructive pulmonary disease (COPD). Otitis media caused by NTHi can lead to ear pain, hearing loss, and fever, particularly in children. Sinusitis symptoms may include facial pain, nasal congestion, and headache. In adults with underlying respiratory conditions like COPD, NTHi can exacerbate symptoms[39]. While NTHi infections are typically less severe than those caused by encapsulated H. influenzae, they can still significantly impact the quality of life, particularly in vulnerable populations[38].

## *Haemophilus influenzae* and HIV

The interplay between H. influenzae and Human Immunodeficiency Virus (HIV) is an area of ongoing research and will be the focus of chapter 2 of this dissertation. Both pathogens can affect the immune system and have the potential to interact in various ways. It is well documented that people affected by HIV are more susceptible to infections because of the effect HIV has on CD4 T-cells, and there are clinical publications noting a link between HIV and H. influenzae and NTHi infections. Some studies have suggested that individuals with HIV may be more prone to chronic respiratory infections, including chronic obstructive pulmonary disease (COPD), bronchitis, and pneumonia. H. influenzae, especially non-typeable strains, can contribute to these respiratory infections, exacerbating the health challenges faced by people with HIV. Despite some knowledge about the interplay between HIV and H. influenzae, there is still much to learn. Ongoing research aims to better understand the mechanisms through which these pathogens interact, the impact on disease progression, and potential strategies for prevention and treatment.

# CHAPTER 2: NTHi OUTBREAK

In the second chapter, we described the genomic components of a particularly unique outbreak of NTHi that was captured by genomic surveillance.

Still in line with the hypothesis of genomic changes leading to hypervirulence, we could observe that a set of genes is uniquely gained or lost in the C1 and C2 samples compared to a subset of all the published Hflu genomes. The strength of this conclusion depends on the scale of the database that we are working with, and we have conducted this analysis with a uniquely large database of Hflu samples. To explore the hypothesis in this light, we must conduct a pangenome analysis. A pangenome analysis allows us to assess if novel genes were acquired in the C1 and C2 lineages that could contribute to transmission, carriage, or increasing invasiveness of NTHi infection. The pangenome will allow us to define the core, accessory, and rare genomes by the relative percentage of each gene family in the population of samples. For this paper, we will define the core genome as greater than or equal to 95%, accessory as between 95% and 5% prevalence, and rare as less than 5% prevalence. A gene family is a group of alleles in the pangenome that are grouped through high sequence similarity by PIRATE, and will be the unit by which the core, accessory, and rare genomes are defined. Accessory gene families are found in some, but not all genomes, and can lend fitness advantages, virulence factors, antibiotic resistance, and other niche adaptations. Focusing on accessory genes present in the C1/C2 samples, we aim to classify gene families by uniquely gained or lost in C1/C2 samples compared to a representative subset of the 4,842 samples and gene families that can distinguish C1 and C2 from each other. With their presence characterized, we can look into potential niche adaptations the predicted gene function may confer.

## COVID-19 outbreak

From here I will briefly transition to some introductory text that is specialized to only our COVID-19 study. In March 2020, the novel SAR-CoV-2 virus (COVID-19) was declared a pandemic by the World Health Organization (WHO)[40]. As time progressed, it continued to affect more people around the world, resulting in millions of deaths[41]. Although COVID-19 was a novel virus, it is related to other coronaviruses (CoV) that are well-studied and understood[42,43]. These viruses are positive-sense RNA viruses, that affect a range of hosts from mammals and birds[44]. Because they are positive-sense, they are immediately ready to be translated by the host cell and create the viral proteins that are important for their replication cycle. These are an interesting and diverse set of viruses, with large genomes and complex open-reading frames, making them a very interesting virus to study.

## Metagenomics and Metatranscriptomics

Our study in Chapter 3 differs from Chapter 2 because instead of working with cultured bacterial samples, we are working with metatranscriptomic data. Metagenomics is a DNA library that is made from a population of cells, representing the microbiome at that site[45,46]. Metatranscriptomics is a variation of metagenomics, which is an RNA library that represents the transcriptome of all of the cells in the microbiome. For the COVID-19 co-infection study, metatranscriptomic libraries were made since the COVID-19 virus is an RNA virus and we wanted to detect this. With these libraries we not only get the identification of other cells in the samples, whether they be host, bacterial, fungal, or viral, but we also get the transcriptome. For RNA viruses like COVID-19 and many other respiratory viruses, this is their entire genome. For other eukaryotic and prokaryotic cells, this will be the steady-state RNA levels of genes that have been expressed. This is very cool and offers a wealth of data that I have only started to tap into.

## CHAPTER 3: COVID-19 CO-INFECTION

In chapter 3, I have built a computational pipeline that processes a high volume of COVID-19 metatranscriptomic samples in order to establish if there are any co-infection of interest in these patients. At the onset of COVID-19 infection, nasal samples from patients were collected and used as way to diagnose infection. From the same sample, total RNA was isolated and metatranscriptomic libraries were created. From a computational standpoint, these samples had to be pre-processed, filtered for read quality, and classified. Downstream analysis looking at co-infections also needed to be written to parse through a large database of samples. We found some of the viral candidates present in our samples; however, they were only found in samples with 0 reads assigned to COVID-19 by Kraken2. Some fungal and bacterial candidates were also found in samples with ranging levels of COVID-19, but none of these candidates correlated with increased viral load at the onset of infection.

# CONCLUSIONS

As an outline, the second chapter is the NTHi study, the third chapter is the COVID-19 co-infection study, the fourth chapter is the conclusion, followed by an appendix of other computational resources I wrote during my time at Emory. In the second chapter, we will investigate the genetic components of NTHi infections in HIV+ men. We find many candidates for increased virulence, but none that have well-established biological significance. In the third chapter, I describe a computational pipeline to process metagenomic samples that allow us to have a picture of the prevalence of co-infection in COVID-19 patients. We find that there are a few viral co-infection in patients with very low COVID-19 viral burden, and also some bacterial and fungal co-infections that are prevalent and do not correlate with more severe COVID-19 infections. In the appendix, I have included a tutorial for using local computational resources and Emory AWS resources.

# REFERENCES

1. Dicker, R. C., Coronado, F., Koo, D. & Parrish, R. G. Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics. 3rd ed.
2. Kayali, G. The forgotten history of pre-modern epidemiology: contribution of Ibn An-Nafis in the Islamic golden era. *East. Mediterr. Health J.* **23**, 854–857 (2018).
3. Connor, H. John Graunt F.R.S. (1620-74): The founding father of human demography, epidemiology and vital statistics. *J. Med. Biogr.* 9677720221079826 (2022).
4. Bingham, P., Verlander, N. Q. & Cheal, M. J. John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. *Public Health* **118**, 387–394 (2004).
5. Snow, J. On the Mode of Communication of Cholera. *Edinb. Med. J.* **1**, 668–670 (1856).
6. Farr, W. Influence of Elevation on the Fatality of Cholera. *Journal of the Statistical Society of London* **15**, 155–183 (1852).
7. Cruickshank, R. Principles of epidemiology. *BMJ* **1**, 1464–1464 (1958).
8. Collins, L. F. *et al.* Invasive Nontypeable Haemophilus influenzae Infection Among Adults With HIV in Metropolitan Atlanta, Georgia, 2008-2018. *JAMA* **322**, 2399–2410 (2019).
9. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
10. Riley, L. W. & Blanton, R. E. Advances in Molecular Epidemiology of Infectious Diseases: Definitions, Approaches, and Scope of the Field. *Microbiol Spectr* **6**, (2018).
11. Honardoost, M., Rajabpour, A. & Vakili, L. Molecular epidemiology; New but impressive. *Med. J. Islam. Repub. Iran* **32**, 53 (2018).
12. Zheng, X. *et al.* Accuracy of serological tests for COVID-19: A systematic review and meta-analysis. *Front Public Health* **10**, 923525 (2022).
13. Gwinn, M., MacCannell, D. & Armstrong, G. L. Next-Generation Sequencing of Infectious

Pathogens. *JAMA* **321**, 893–894 (2019).

14. Srivastava, S., Singh, P. K., Vatsalya, V. & Karch, R. C. Developments in the Diagnostic Techniques of Infectious Diseases: Rural and Urban Prospective. *Adv Infect Dis* **8**, 121–138 (2018).
15. Mobley, I. A brief history of Next Generation Sequencing (NGS). *Front Line Genomics* https://frontlinegenomics.com/a-brief-history-of-next-generation-sequencing-ngs/ (2021).
16. Wadman, M. James Watson's genome sequenced at high speed. *Nature* **452**, 788 (2008).
17. Armstrong, G. L. *et al.* Pathogen Genomics in Public Health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).
18. Petit, R. A., 3rd & Read, T. D. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. *PeerJ* **6**, e5261 (2018).
19. Active Bacterial Core surveillance system (ABCs). https://www.cdc.gov/abcs/index.html (2021).
20. Inzaule, S. C., Tessema, S. K., Kebede, Y., Ogwell Ouma, A. E. & Nkengasong, J. N. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.* **21**, e281–e289 (2021).
21. Balakrishnan, V. S. WHO's global genomic surveillance strategy. *Lancet Infect. Dis.* **22**, 772 (2022).
22. Preparedness, P. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. https://www.who.int/publications/i/item/9789240046979 (2022).
23. Cummins, E. A., Hall, R. J., McInerney, J. O. & McNally, A. Prokaryote pangenomes are dynamic entities. *Curr. Opin. Microbiol.* **66**, 73–78 (2022).
24. McInerney, J. O., Whelan, F. J., Domingo-Sananes, M. R., McNally, A. & O'Connell, M. J. Pangenomes and Selection: The Public Goods Hypothesis. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) (Springer, 2020).

25. Gonzalez-Diaz, A. *et al.* Comparative pangenome analysis of capsulated Haemophilus influenzae serotype f highlights their high genomic stability. *Sci. Rep.* **12**, 3189 (2022).

26. Brockhurst, M. A. *et al.* The Ecology and Evolution of Pangenomes. *Curr. Biol.* **29**, R1094–R1103 (2019).

27. Bedoya-Correa, C. M., Rincón Rodríguez, R. J. & Parada-Sanchez, M. T. Genomic and phenotypic diversity of Streptococcus mutans. *J. Oral Biosci.* **61**, 22–31 (2019).

28. Gómez, P. *et al.* Genomic Analysis of Staphylococcus aureus of the Lineage CC130, Including mecC-Carrying MRSA and MSSA Isolates Recovered of Animal, Human, and Environmental Origins. *Front. Microbiol.* **12**, 655994 (2021).

29. Bosi, E. *et al.* Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3801–9 (2016).

30. Smith, J. T. *et al.* Genome Evolution of Invasive Methicillin-Resistant Staphylococcus aureus in the Americas. *Microbiol Spectr* **10**, e0020122 (2022).

31. Wen, S., Feng, D., Chen, D., Yang, L. & Xu, Z. Molecular epidemiology and evolution of Haemophilus influenzae. *Infect. Genet. Evol.* **80**, 104205 (2020).

32. Van Epps, H. L. Influenza: exposing the true killer. *J. Exp. Med.* **203**, 803 (2006).

33. Musher, D. M. Haemophilus Species. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston).

34. Slack, M. P. E. Long Term Impact of Conjugate Vaccines on Haemophilus influenzae Meningitis: Narrative Review. *Microorganisms* **9**, (2021).

35. Smith, D. A. & Nehring, S. M. *Bacteremia*. (StatPearls Publishing, 2023).

36. Stieb, D. M. *et al.* Effectiveness of Haemophilus influenzae type b vaccines. *CMAJ* **142**, 719–733 (1990).

37. Foxwell, A. R., Kyd, J. M. & Cripps, A. W. Nontypeable Haemophilus influenzae: pathogenesis and prevention. *Microbiol. Mol. Biol. Rev.* **62**, 294–308 (1998).

38. Bakaletz, L. O. & Novotny, L. A. Nontypeable Haemophilus influenzae (NTHi). *Trends Microbiol.* **26**, 727–728 (2018).

39. Short, B. *et al.* Non-typeable Haemophilus influenzae chronic colonization in chronic obstructive pulmonary disease (COPD). *Crit. Rev. Microbiol.* **47**, 192–205 (2021).

40. Krumbein, H. *et al.* Respiratory viral co-infections in patients with COVID-19 and associated outcomes: A systematic review and meta-analysis. *Rev. Med. Virol.* **33**, e2365 (2023).

41. Khan, M. *et al.* COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far. *Molecules* **26**, (2020).

42. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).

43. Sharma, A., Ahmad Farouk, I. & Lal, S. K. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention. *Viruses* **13**, (2021).

44. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim. Biophys. Acta Mol. Basis Dis.* **1866**, 165878 (2020).

45. Zhang, L. *et al.* Advances in Metagenomics and Its Application in Environmental Microorganisms. *Front. Microbiol.* **12**, 766364 (2021).

46. Gu, W., Miller, S. & Chiu, C. Y. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. *Annu. Rev. Pathol.* **14**, 319–338 (2019).

# Chapter 2: NTHi

Below is an upcoming submission of my work on the NTHi outbreak.

## Comparative genomic analysis of emerging non-typeable Haemophilus influenzae (NTHi) causing emerging septic arthritis in Atlanta

Brianna J. Bixler[1] , Robert A Petit III, Andrei Bombin[1] , Abraham Moller, Sam Sefton[1,3,4], Stepy Thomas[1,3,4], Amy Tunali[1,3,4], Lauren F. Collins[1,2] , Monica M. Farley[1] , Sarah W. Satola[1] , Timothy D. Read[1 *]

[1] Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA

[2]Ponce de Leon Center, Grady Health System, Atlanta, Georgia, USA

[3]Georgia Emerging Infections Program, Atlanta, Georgia, USA

[4]Atlanta VA Health System, Decatur, Georgia, USA

(Should we add an EIP affiliation as well? Yes, see above)
* Corresponding author, Email address: tread@emory.edu

## INTRODUCTION

*Haemophilus influenzae* (Hi) is a Gram-negative bacterium that can live on human mucosal surfaces without causing an infection but can also be associated with ear and respiratory infections or more invasive diseases, such as bacteremia or meningitis. Hi strains expressing polysaccharide capsule genes a-f, containing capsule genes on the IS1060 transposon, have historically been associated with more serious invasive disease[1,2]. Since the routine use of the Hi serotype b (Hib) vaccine in the 1990s, strains lacking the intact capsule locus (NTHi; non-typeable *Haemophilus influenzae*) have replaced encapsulated strains as the leading cause of invasive *Hi* disease[3,4].

In recent years, a CDC-funded active population-based surveillance program was leveraged to evaluate a sharp increase in the rate of iNTHi infection among persons living with HIV in 2017-2018 compared to prior years evaluated 2008-2016, and identified that the cases primarily occurred in Black men who have sex with men that had a high prevalence of septic arthritis[5]. Pulsed-field gel electrophoresis typing among iNTHi cases aged 18-55 years identified two expanded NTHi clones, named clusters 1 and 2 ("C1" and "C2") as predominant in the 2017-2018 iNTHi cases. Whole genome shotgun analysis identified C1 and C2 isolates as corresponding to multilocus sequence types ST164 and ST1714, respectively.  None of the C1/C2 strains contained capsule genes but all C1 contained the IS1016 transposon gene. Additionally in the C1 isolates, there were 2 genes flanking the IS1016 gene that are homologs of genes at this locus in encapsulated strains. The presence of these genes suggests ancestry from an encapsulated strain. Although ST164 and ST1714 were close relatives within the Hi species phylogeny, their last common ancestor clearly predated the likely timing of the Atlanta outbreaks, suggesting that two independent outbreaks were occurring concurrently. Geospatial analysis of iNTHi cases in metropolitan Atlanta revealed temporal-geographic separation between cases by cluster type as C1 and C2 and further, significant temporal-geographic aggregation of C1 cases from January-December 2017 in a certain geography compared with C2 cases [5].

It was not clear whether there were unusual genetic features of the NTHi C1 and C2 isolates that prompted infection to lead to more serious invasive disease (i.e., septic arthritis) or whether their expansion reflected chance introduction into a vulnerable population and transmission within social networks. In this study, we performed a comparative genomic analysis of the NTHi C1 and C2 isolates originally identified in metropolitan Atlanta in the context of the larger pangenome of *Hi* strains globally to identify potential features that may suggest enhanced virulence in the cluster strains.

## RESULTS

### Genetic variation within the C1 and C2 clusters

We obtained Illumina shotgun sequence data of 26 C1 and 23 C2 isolates from the original Atlanta investigation from  2017-2018. One strain was randomly chosen from each cluster for hybrid assembly of Oxford Nanopore minIOn and Illumina data to produce complete reference sequences. The final C1 (GA81666) hybrid assembly included one circular 1.875 Mb contig, while C2 (GA54827) included one circular 1.885 Mb and one circular 37.6 Kb plasmid.

The C1 isolates were found to be highly related, with a maximum distance of 132 SNPs in the core genome alignment. Many of the samples had zero SNP distance **(Figure 1A)**. There were 29 deletion regions in the C1 isolates compared to the reference genome, ranging from 1 to 17 kb and encompassing 48 genes **(Figure 2A),** 6 of which had a significant match to *Hi* virulence genes.  The C2 isolates were also found to be highly related, all representing the same sequence type and having a maximum distance of 149 SNPs in the core genome alignment **(Figure 1B).** The C2 cluster separates into 2 subclades, with sample SNP distances as low as 1 and as high as 35 SNPs within these sub-groups. In a similar manner to C1 analysis, there were 13 large deletions (~1 to 32 kb), spanning 24 annotated genes **(Figure 2B),** 6 of these genes matched suspected virulence genes.  Of the genes that had annotated gene names, five were deleted in at least one sample in both C1 and C2 (*eamA*, *ninG*, sRNA-Xcc1, sRNA-Xcc1, and *tolB*). None of these genes are found in the virulence factor database.

### Few accessory genes distinguish the C1 and C2 clusters

We created a database of 4,842 publicly available Hi genomes to compare to C1 and C2 (see methods).  The public genomes represented 536 distinct MLST sequence types. From this we selected 536 randomly chosen ST representatives, in addition to 26 C1 isolates, and 23 C2 isolates and created consisted of 6,560 gene families, of which 1368 were core (>= 95% genomes), 1107 intermediate accessory (95% < x <= 5 %) and 4085 rare accessory (x > 5%) (**Figure 3**). In a phylogenetic tree of randomly chosen representative strains of each 536 MLST sequence types, C1 and C2 were part of a closely related subclade of NTHi strains (**Figure 4**).

There were few genes in the Hi pangenome that had unusual patterns of gain and loss confined to C1 and C2 and none that could be linked to a known virulence function. We found that there were no accessory genes absolutely unique to either C1 or C2, nor both C1/C2 (i.e. present in one or both clades and not found in other MLSTs). There were also no core genes missing in only C1 or C2, or both. While there were no genes absent in C1 and/or C2 that were present in 100% of the rest of the *Hi* pangenome, there was one gene family, identified as pxpB, that was lost in 100% of the C1 and C2 isolates and present in more the 90% of rest of the population of strains **(Figure 5)**. There were 20 gene families that were 'rare' in the context of the *Hi* pangenome (i.e. found in < 10% STs) present in C1 but not C2 isolates (**Table 3**). Further, there were 7 'rare' gene families unique to C2, that were not identified in C1 strains (**Table 4**). There were no rare genes shared by both C1 and C2. Based on the virulence criteria established there were no obvious links to virulence in the C1 and C2 unique or rare genes.

While examining the Hi pangenome data, we observed an interesting pattern of gains and losses that correspond to a previously undescribed mobile cassette inserted in the same ancestral region of C1, C2, and a small subset of other sequence types. There were eight genes identiifed as present in both C1 and C2 but in less than 90% of the rest of the pangenome. All 8 gene families had significant homoplasy, with a consistency index less than or equal to 0.2 on the species core genome tree **(Table 1)**. Within this subset of rare genes, we identified 5 genes that were likely acquired as a cassette in the ancestor and inserted at the ancestral site of the one lost gene family.  The gene that was the apparent site for cassette integration in C1 and C2 encoded a protein annotated as 5-oxoprolinase subunit PxpB[6]. Using comparative genomic analysis with the MAUVE tool, we found that the five inserted genes were on a cassette of 9444 bp in C1 and C2. In the same region of the outgroup that lacked the cassette (GCF_014701215.1_ASM1470121v1), there was an intact *pxpABC* gene cluster.  Four of the 5 genes within the boundary of the cassette have been assigned a gene name and function by bakta annotations. One of those gene families is the gene *tnpA*, which is crucial for IS200/IS605 family transposition [7,8]. The remaining 3 were potentially part of a sugar metabolic operon based on their annotations. These were (*ptsEII*) - a sugar transmembrane transporter; *malQ*, 4-α-glucanotransferase important in maltose metabolism;  and *treR* the repressor of the trehalose metabolic pathway. The presence of intact *pxpB* was highly homoplasic in *Hi* (consistency Index of 0.14), which suggested a history of frequent gain and loss in the species, commensurate with cassette insertion.  There was a pattern of lost and gained gene families found in other sequence types on the tree. We observed that there were 42 STs that have the exact same pattern of the lost *pxpB* gene family and concurrently gained the 5 gene families within the cassette. The pattern of gains and losses could be explained by multiple independent insertions of a cassette at the same location that disrupted *pxpB*. The metadata available for these 42 STs included isolates from both blood and respiratory infections. Many isolates were from infections of populations at risk of pulmonary infection such as patients with chronic obstructive pulmonary disease (COPD) and cystic fibrosis.

## C1 and C2 have accessory gene profiles more similar to *Hi* isolates from blood than sputum

Out of 4,842 Hi isolates with public genomes, 1,624 had metadata indicating the isolates were collected from a blood or system infection and 1,441 were labeled as being isolated from sputum (the other genomes did not have identifiable metadata). Our goal was to identify genes associated with systemic Hi infections using a Genome-Wide Association Study (GWAS) approach to determine if  C1 and C2 isolates had systemic infection genes represented. We randomly choose at most one representative sample from each ST collected from either blood or sputum to reduce bias introduced by oversampling a small number of STs with a large number of genomes. This resulted in a set of 146 blood and 87 sputum-associated genomes. Using the Scoary pangenome GWAS tool[9] we identified 24 accessory genes that have a Bonferroni p-value less than 0.05 and odds ratio greater than 2 association with blood versus sputum (**Table 2**). Compared to the *Hi* 4,842 pangenome set, representative C1 and C2 genomes were enriched in the presence of these accessory genes associated with blood infections, having 16 and 14 genes (of the 24 identified?), respectively (**Figure 6**).

## Rates of recombination and pseudogene-formation were similar in C1 and C2 to the rest of the *Hi* species

While the analysis assessing patterns of the presence and absence of genes in the pangenomes yielded limited insight about C1/C2 virulence, we hypothesized that allelic variation introduced by homologous recombination may have played a role in the potential evolution to increased pathogen virulence, as seen for example in the recent emergence of a novel urogenital *Neisseria meningitidis* strain[10]. To investigate this hypothesis, we created a core genome alignment of 50 randomly chosen STs and one representative each from C1 and C2 to predict potential regions of recombination using the Gubbins tool (**Figure 7**). As found in previous whole genome studies of,[11,12] recombination was common across *Hi* genomes but there were clade-specific patterns. Gubbins detected 57 recombination events in the C1 genome with a rho/theta of 0.53 and no likely recombination events in the C2 genome with a rho/theta of 0. In GA81666 representing C1, there were 1717 SNPs inside a region of recombination and 30 SNPs outside of the 16 recombination blocks identified. In GA54827 representing C2, there were 0 recombination blocks identified and therefore 0 SNPs within a recombination block. In terms of the number of genes overlapping SNPs, the GA81666 (C1) genome had 326 genes in regions involved, while the GA54827 (C2) genome had 292 genes similarly affected. Notably, there were no genes affected by independent events in both C1 and C2 genomes.

Finally, we found the number of pseudogenes identified in C1 and C2 strains using the Bakta tool (6 and 10 hypothetical proteins, respectively) did not significantly differ from the number found in other *Hi* strains (**supplemental data?**). No genes with a function linked to virulence were found to have acquired null mutations in either C1 or C2.

## DISCUSSION

Using innovative and comprehensive bacterial genotypic analytic methods, we investigated whether the emergence of two clones of NTHi (C1 and C2) associated with a novel clinical presentation of invasive disease in metropolitan Atlanta (primarily occurring among persons with well-controlled HIV)  was associated with genomic changes that could have increased the virulence of clones. We showed that both clusters consisted of closely related strains, with few core chromosome SNPs but with some gene loss occurring.  Because of the limited within-clade diversity, we compared representative strains to an extensive public dataset of *Hi* genomes to gather clues about potential clade-level adaptations. We evaluated patterns of gene gain and loss involving the C1/C2 isolates in the context of an *Hi* species pangenome deriving from 4,842 publicly available genomes.  Based on the pangenome analysis, we identified 24 accessory genes enriched in Hi strains associated with blood/ systemic over sputum infections and found that the genomic composition of the C1/C2 strains resembled those causing invasive disease.

We examined whether specific gene losses or gains may be linked to virulence and discovered a previously undescribed mobile cassette in *Hi* as part of the C1 and C2 genomes but otherwise did not see evidence of unusual patterns in C1/C2. Finally, we determined that rates of homologous recombination or pseudogenization in the C1/C2 genomes were not outliers compared to other *Hi* clades.  While our investigation did not reveal genomic changes in C1 and C2 that could be directly associated with traditionally defined serotyped *Hi* virulence factors, we did identify intriguing changes potentially meriting further exploration through laboratory-based analysis.

While we did not find genes with known functions unique to C1 and/or C2, we did identify a novel cassette encoding a polysaccharide metabolism cluster inserted in C1/C2 genomes in a manner that disrupted the *pxpB* gene. This disruption could itself be linked to a pathoadaptive phenotype. The *pxpB* gene is part of an operon including *pxpA* and *pxpC*. Single gene mutations in *B. subtilis* showed that each of the *pxpA*, *pxpB*, and *pxpC* genes were necessary and sufficient for 5-Oxoproline (OP) metabolism, and deletion of any resulted in OP accumulation and slowed growth (Niehaus et al. 2017).  Accumulation of OP causes a number of cellular responses in prokaryotes, including growth inhibition [6]. Deletion of *pxpB* showed aberrant DNA recombination within a large genetic interaction screen in *E. coli*[13].  The disruption of pxpB should therefore be associated with a fitness deficit, so it is interesting that it is the target for disruption in several *Hi* lineages.  This might suggest a possible tradeoff for a pathoadaptive trait and should be investigated further.

Of the 24 genes that were identified by SCOARY as potentially discriminatory for bloodstream infection, only two were known virulence factors in the *Haemophilus* genus according to the VFDB. The *lsgB*,  gene found in both C1 and C2 samples, is associated with *Haemophilus parasuis* virulence through its involvement in lipooligosaccharide biosynthesis sialylation[14]. It is one of several virulence genes in *Haemophilus parasuis*, contributing to the bacterium's pathogenicity by influencing sialylated lipooligosaccharide production. Another of the 24, an *igaA1* gene, found in C1 but not C2 genomes, is a homolog of *Salmonella* membrane protein IgaA [15]. IgA regulates bacterial regulons like RcsC-YojN-RcsB and PhoP-PhoQ, with the igaA1 allele (due to an R188H mutation) altering the expression of PhoP-PhoQ-activated (pag) genes, such as *ugd*, which is linked to lipopolysaccharide modification and colanic acid capsule synthesis[15].

Three other genes (*dacB*, *fbp,* and *sbcB*) associated with blood infections have been associated with virulence mechanisms for other pathogens, dacB and fbp are the only genes also found core genes in our pangenome analysis. The dacB gene encodes for a serine-type D-Ala-D-Ala carboxypeptidase, and it appears to have the potential to influence virulence in the context of peptidoglycan[16,17]. Studies have shown that disabling this gene, as well as its counterpart *dacA*, in pneumococci led to significant attenuation of the bacteria in infected mice[17]. Additionally, mutants lacking dacB and dacA exhibited enhanced uptake by professional phagocytes and decreased adherence to lung epithelial cells. In another context, a mutation in dacB was associated with changes in peptidoglycan structure, including the release of different peptidoglycan fragments, highlighting its role in peptidoglycan metabolism[16]. The *fbp* gene, which encodes fructose-1,6-bisphosphatase, exhibits varying effects on virulence in different bacterial and protozoan species. In *Brucella*, the loss of *fbp* does not impact virulence, suggesting that it is not essential for full virulence in laboratory models[18]. Similarly, in *Brucella suis* biovar 5, fbp is not required for full virulence in laboratory models[18]. However, in *Leishmania*, the gluconeogenic enzyme fructose-1,6-bisphosphatase encoded by fbp is essential for virulence, as mutants lacking this enzyme can persist in mice but fail to generate normal lesions[19]. This suggests that *Leishmania* relies on fructose-1,6-bisphosphatase for virulence, possibly due to its dependence on non-glucose carbon sources in glucose-poor phagosomes. Additionally, fbp has been identified in a screen in *Staphylococcus*, but its specific role in virulence in this context is not detailed in the provided information[20]. The *sbcB* gene, which encodes exodeoxyribonuclease I, is a recognized component of virulence in *Salmonella*[21]. Research has shown that mutants of *Salmonella* lacking the RecBC function, in which sbcB is involved, are avirulent in mice and incapable of growing inside macrophage [21]. This finding highlights the critical role of the RecBCD recombination pathway, in which sbcB plays a part, in Salmonella's virulence. This pathway is essential for repairing double-strand breaks generated during DNA replication and is proposed to be necessary for systemic infection by *S. enterica*, as it likely facilitates DNA replication within phagocytes during infection, notably the other pathogens included don't necessarily cause joint manifestations.

The comparative genomic approach described here is limited to events involving gain and loss or recombination-driven allelic change in genes with known virulence functions; within those limitations, our data do not reveal unusual patterns of virulence genes in C1/C2. There are genomic changes that could cause hypervirulence but would not be detected using the methodology implemented in this analysis, such as rare SNPs, particularly in regulatory genes, genomic rearrangements, and the gain of virulence genes of unknown function. One future exploratory approach may include evaluating potential differences in gene expression in virulence models between C1/ C2 and non-pathogenic *Hi*. From these data, hypotheses on genetic changes responsible for the phenotypic effects could be generated. Finally, our analysis focused on a C1/C2 genomic-derived mechanism underlying the emergence of two novel NTHi strains leading to invasive disease with a relatively high prevalence of joint involvement among primarily persons with HIV and Black men who have sex with men who resided in geographic proximity. However, this unique clinical presentation in a particular demographic warrants additional investigation into host factors as well as potential transmission modes including anatomic sites as well as the potential role of social networks.

# METHODS

## Oxford nanopore sequencing and hybrid assembly with Illumina data

*Haemophilus influenzae* strains GA81666 and GA54827 genomic DNA were extracted using the Promega Wizard Genomic DNA Purification Kit. Sequencing libraries prepared using the SQK-LSK109 1D ligation sequencing kit and sequenced on a FLO-FLG001 Flongle flow cell, yielding 496.9 Mb and 552.6 Mb of raw reads (~267x and ~297x coverage) for GA81666 and GA54827, respectively. The *Hi* GA81666 and GA54827 genomes were then assembled from Nanopore and Illumina paired-end reads using Unicycler[22].

## Downloading public *Hi* genome data

One of the strengths of this study is setting up not only a database of *Hi* samples to compare against the clusters identified in the original investigation in Atlanta, but also assembling resources and a pipeline to assess their uniqueness and potential virulence. 4,842 samples were downloaded from NCBI SRA database and run through the bactopia pipeline to ensure consistency. The output of this pipeline were used as the inputs for the pangenome, pan-GWAS, and recombination analysis that is to follow.

## *Hi* virulence gene database

We created a blast database of all of the virulence factors defined in the virulence factor database (http://www.mgc.ac.cn/VFs/main.htm accessed October 2023) and the Victors database (https://phidias.us/victors/ accessed October 2023)[23,24]. To match a protein against our database we used blastp with default parameters and used a threshold of 80% identity to define a hit.

We defined a gene as being “potentially linked to virulence” if it was either 1) in the virulence factor database and the Victors database blast database, 2) one of the 24 gene families potentially linked to systemic infection from our SCOARY analysis described in the next section or 3) its annotated gene name contained the terms “virulence”, “pathogenicity”, “capsule” or “toxin”.

## Processing whole genome data

For the remaining *Haemophilus influenzae* strains we used Bactopia (v1.6.0, [25]) to process the Illumina data. The multi-locus sequence type (MLST) schema for *Hi* from PubMLST.org[26] was included. In addition to the public datasets included with the command “bactopia datasets”, we added gene and protein sequences of 105 *H. influenzae* reference genes described in Pinto et. al. [27], and completed genomes for GA81666 and GA54827 as optional datasets.

With Bactopia the reads were cleaned and error-corrected using BBDuk (v38.86, [28]) and Lighter (v1.1.2, [29]). Process reads were assembled with SKESA (v2.4.0, [30]) using Shovill (v1.1.0, [31]. Assembly quality metrics were determined with assembly-scan (v0.3.0 [32]) and CheckM (v1.1.3, [33]). The species composition of the assembly was determined by screening against a minmer sketch of GenBank [34] using sourmash (v3.5.0, [35]). The MLST was determined using BLAST+ (V2.10.1, [36]) and Ariba (v2.14.6, [37]).

To supplement our study we used “bactopia search” to identify all publicly available *H. influenzae* genomes from the Sequence Read Archive in December 2020. Each of the public genomes with Illumina sequencing was also processed through Bactopia..

The “summary” tool from Bactopia was used to aggregate the results for all genomes into a single table (nthi-report.txt). The average nucleotide identity (ANI) between GA81666 and all other genomes was calculated with FastANI (v1.32, [38]). A Python script ('generate-representative-set.py') was created to identify genomes to exclude from further comparative analysis as well as generate a set of genomes that represented all identified sequence types at least once.

Any genomes with more than 500 assembled contigs, did not have a sequence type determined, or not screened as "Haemophilus influenzae" by sourmash, were excluded from further analysis.

The representative set included all Georgia genomes from this study (that were not excluded) and all genomes that had an ANI greater than the floor of the ANI of the most distant C2 genome from GA81666 (a member of C1). If an ST was identified, but not yet included, a genome was picked prioritizing its quality rank, assembly completeness, and total number of contigs.

Comparative analyses were done using available Bactopia. A tree was constructed with Mashtree (v1.2.0, [39]). The representative set was used to determine the pan-genome using PIRATE (v1.0.4, [40]). A phylogenetic tree based on the recombination masked core-genome alignment was created with IQ-TREE (v2.0.3, [41,42]).

### Intra-clade comparisons

SNPs were called using the bactopia-tools workflow for snippy. To understand if there was any variability across the entire genomes of C1 and C2 samples, reads from each sample were individually mapped to the completed reference genome of each and potentially deleted regions were assessed by lower than expected coverage in regions. In 1000 bp windows, sliding every 250 bp, reads were counted, adjusted for rpkm, and selected if one of the samples had 0 reads over that bin. Adjacent regions of zero coverage were combined into a single region of interest and coverage was converted into a binary presence or absence.

### GWAS

In order to gather association data from bacterial genomes of nontypeable Hflu (NTHi) in the bloodstream versus sputum, we used publicly available samples from multiple studies and the Roary/Scoary software[9]. Roary is a software that creates a pangenome from gff annotation files for each sample. The software collects the DNA sequence of each annotated gene, translates to the protein sequence, blasts protein sequences against each other to create gene families of highly similar sequences, and creates a count table of gene families that are present or absent in each sample. Scoary inputs include the roary pangenome csv that annotated the presence or absence of each gene family in every sample and a manifest file separating the samples into the two conditions. The first step is to identify gene families in the accessory genome and calculate initial associations to the conditions. A second association is calculated by incorporating the phylogenetic structure of the samples. This is followed by a permutation test. The output is a table of genes and their calculated p-values, odd ratio, and another metric of significance. From this, we hope to identify genes positively associated with septic infection, query the genome of C1/C2 clusters to identify which are present in these samples.

### Recombination and Pseudogene analysis

The core-genome alignment of 50 randomly chosen STs had recombination events predicted by Gubbins (ref) and masked with maskrc-svg (v0.5) [44].Pseudogenes were detected using bakta reporting of pseudogenes, a new feature to version v1.8.2(Schwengers et al. 2021).

### Data availability

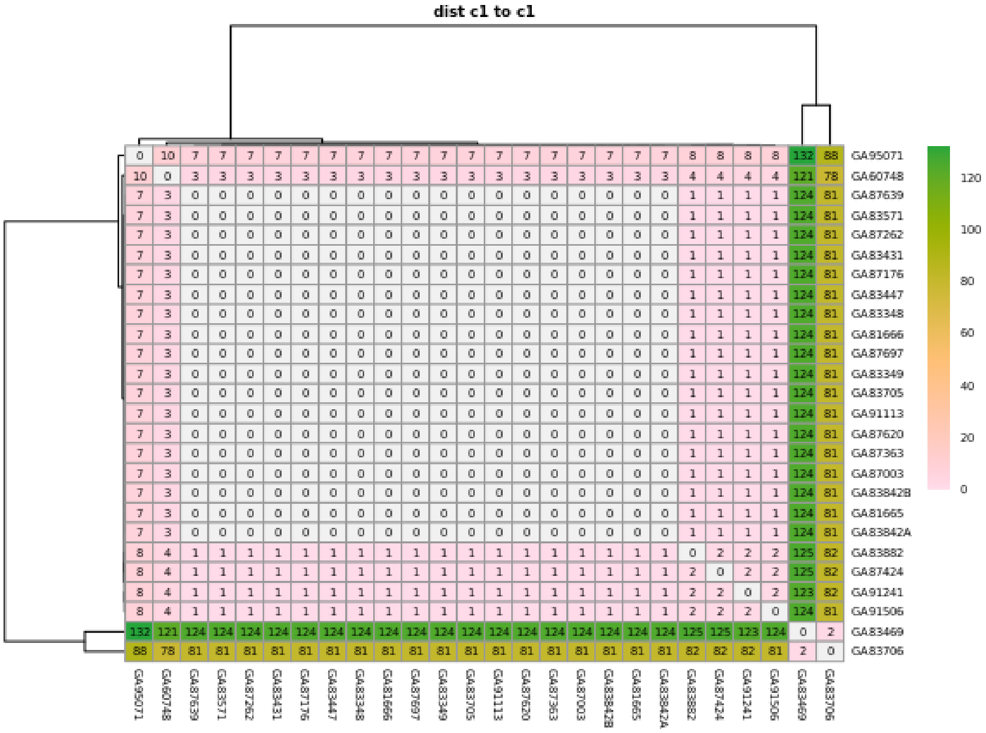
Summary data is available at **/mnt/tiramisu/emergent/projects/NTHI/ena-results.txt**

The exact commands and code used in this study are available at

https://github.com/Read-Lab-Confederation/gaeip-nthi.

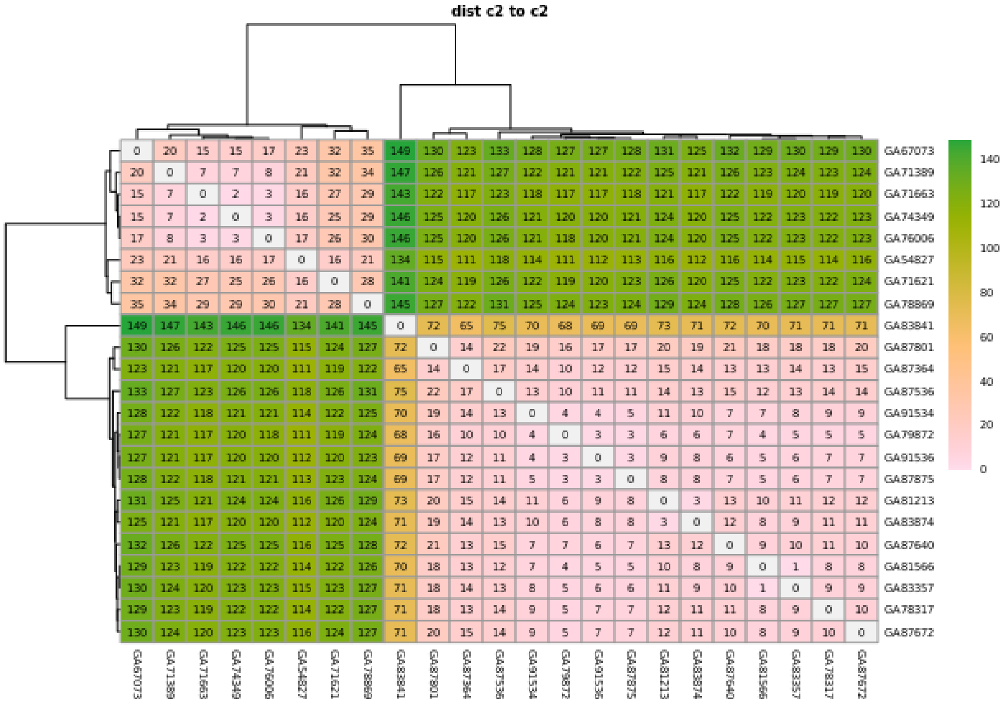
## FIGURES

### Figure 1A: SNPs in the core alignment of C1



| | GA95071 | GA60748 | GA87639 | GA83571 | GA87262 | GA83431 | GA87176 | GA83447 | GA83348 | GA81666 | GA87697 | GA83349 | GA83705 | GA91113 | GA87620 | GA87363 | GA87003 | GA83842B | GA81665 | GA83842A | GA83882 | GA87424 | GA91241 | GA91506 | GA83469 | GA83706 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GA95071 | 0 | 10 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | 8 | 132 | 88 |
| GA60748 | 10 | 0 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 121 | 78 |
| GA87639 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83571 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87262 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83431 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87176 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83447 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83348 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA81666 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87697 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83349 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83705 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA91113 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87620 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87363 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA87003 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83842B | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA81665 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83842A | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 124 | 81 |
| GA83882 | 8 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 2 | 125 | 82 |
| GA87424 | 8 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 2 | 2 | 125 | 82 |
| GA91241 | 8 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 0 | 2 | 123 | 82 |
| GA91506 | 8 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 124 | 81 |
| GA83469 | 132 | 121 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 124 | 125 | 125 | 123 | 124 | 0 | 2 |
| GA83706 | 88 | 78 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 81 | 82 | 82 | 82 | 81 | 2 | 0 |

## Figure 1B: SNP distance in C2 genome



| | GA67073 | GA71389 | GA71663 | GA74349 | GA76006 | GA54827 | GA71621 | GA78869 | GA83841 | GA87801 | GA87364 | GA87536 | GA91534 | GA79872 | GA91536 | GA87875 | GA81213 | GA83874 | GA87640 | GA81566 | GA83357 | GA78317 | GA87672 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GA67073 | 0 | 20 | 15 | 15 | 17 | 23 | 32 | 35 | 149 | 130 | 123 | 133 | 128 | 127 | 127 | 128 | 131 | 125 | 132 | 129 | 130 | 129 | 130 |
| GA71389 | 20 | 0 | 7 | 7 | 8 | 21 | 32 | 34 | 147 | 126 | 121 | 127 | 122 | 121 | 121 | 122 | 125 | 121 | 126 | 123 | 124 | 123 | 124 |
| GA71663 | 15 | 7 | 0 | 2 | 3 | 16 | 27 | 29 | 143 | 122 | 117 | 123 | 118 | 117 | 117 | 118 | 121 | 117 | 122 | 119 | 120 | 119 | 120 |
| GA74349 | 15 | 7 | 2 | 0 | 3 | 16 | 25 | 29 | 146 | 125 | 120 | 126 | 121 | 120 | 120 | 121 | 124 | 120 | 125 | 122 | 123 | 122 | 123 |
| GA76006 | 17 | 8 | 3 | 3 | 0 | 17 | 26 | 30 | 146 | 125 | 120 | 126 | 121 | 118 | 120 | 121 | 124 | 120 | 125 | 122 | 123 | 122 | 123 |
| GA54827 | 23 | 21 | 16 | 16 | 17 | 0 | 16 | 21 | 134 | 115 | 111 | 118 | 114 | 111 | 112 | 113 | 116 | 112 | 116 | 114 | 115 | 114 | 116 |
| GA71621 | 32 | 32 | 27 | 25 | 26 | 16 | 0 | 28 | 141 | 124 | 119 | 126 | 122 | 119 | 120 | 123 | 126 | 120 | 125 | 122 | 123 | 122 | 124 |
| GA78869 | 35 | 34 | 29 | 29 | 30 | 21 | 28 | 0 | 145 | 127 | 122 | 131 | 125 | 124 | 123 | 124 | 129 | 124 | 128 | 126 | 127 | 127 | 127 |
| GA83841 | 149 | 147 | 143 | 146 | 146 | 134 | 141 | 145 | 0 | 72 | 65 | 75 | 70 | 68 | 69 | 69 | 73 | 71 | 72 | 70 | 71 | 71 | 71 |
| GA87801 | 130 | 126 | 122 | 125 | 125 | 115 | 124 | 127 | 72 | 0 | 14 | 22 | 19 | 16 | 17 | 17 | 20 | 19 | 21 | 18 | 18 | 18 | 20 |
| GA87364 | 123 | 121 | 117 | 120 | 120 | 111 | 119 | 122 | 65 | 14 | 0 | 17 | 14 | 10 | 12 | 12 | 15 | 14 | 13 | 13 | 14 | 13 | 15 |
| GA87536 | 133 | 127 | 123 | 126 | 126 | 118 | 126 | 131 | 75 | 22 | 17 | 0 | 13 | 10 | 11 | 11 | 14 | 13 | 15 | 12 | 13 | 14 | 14 |
| GA91534 | 128 | 122 | 118 | 121 | 121 | 114 | 122 | 125 | 70 | 19 | 14 | 13 | 0 | 4 | 4 | 5 | 11 | 10 | 7 | 7 | 8 | 9 | 9 |
| GA79872 | 127 | 121 | 117 | 120 | 118 | 111 | 119 | 124 | 68 | 16 | 10 | 10 | 4 | 0 | 3 | 3 | 6 | 6 | 7 | 4 | 5 | 5 | 5 |
| GA91536 | 127 | 121 | 117 | 120 | 120 | 112 | 120 | 123 | 69 | 17 | 12 | 11 | 4 | 3 | 0 | 3 | 9 | 8 | 6 | 5 | 6 | 7 | 7 |
| GA87875 | 128 | 122 | 118 | 121 | 121 | 113 | 123 | 124 | 69 | 17 | 12 | 11 | 5 | 3 | 3 | 0 | 8 | 8 | 7 | 5 | 6 | 7 | 7 |
| GA81213 | 131 | 125 | 121 | 124 | 124 | 116 | 126 | 129 | 73 | 20 | 15 | 14 | 11 | 6 | 9 | 8 | 0 | 3 | 13 | 10 | 11 | 12 | 12 |
| GA83874 | 125 | 121 | 117 | 120 | 120 | 112 | 120 | 124 | 71 | 19 | 14 | 13 | 10 | 6 | 8 | 8 | 3 | 0 | 12 | 8 | 9 | 11 | 11 |
| GA87640 | 132 | 126 | 122 | 125 | 125 | 116 | 125 | 128 | 72 | 21 | 13 | 15 | 7 | 7 | 6 | 7 | 13 | 12 | 0 | 9 | 10 | 11 | 10 |
| GA81566 | 129 | 123 | 119 | 122 | 122 | 114 | 122 | 126 | 70 | 18 | 13 | 12 | 7 | 4 | 5 | 5 | 10 | 8 | 9 | 0 | 1 | 8 | 8 |
| GA83357 | 130 | 124 | 120 | 123 | 123 | 115 | 123 | 127 | 71 | 18 | 14 | 13 | 8 | 5 | 6 | 6 | 11 | 9 | 10 | 1 | 0 | 9 | 9 |
| GA78317 | 129 | 123 | 119 | 122 | 122 | 114 | 122 | 127 | 71 | 18 | 13 | 14 | 9 | 5 | 7 | 7 | 12 | 11 | 11 | 8 | 9 | 0 | 10 |
| GA87672 | 130 | 124 | 120 | 123 | 123 | 116 | 124 | 127 | 71 | 20 | 15 | 14 | 9 | 5 | 7 | 7 | 12 | 11 | 10 | 8 | 9 | 10 | 0 |

## Figure 2A: Regions of low coverage within C1

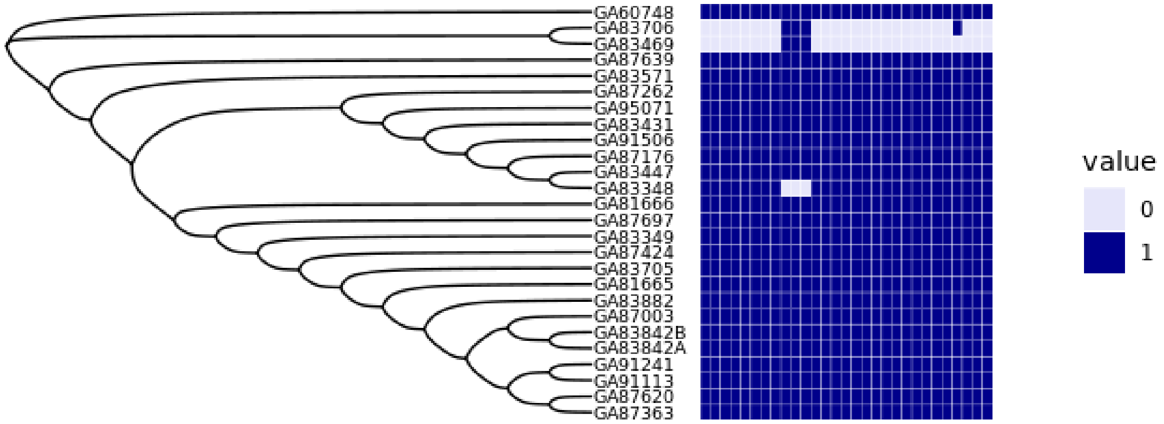


Figure 2B: Regions of low coverage within C2

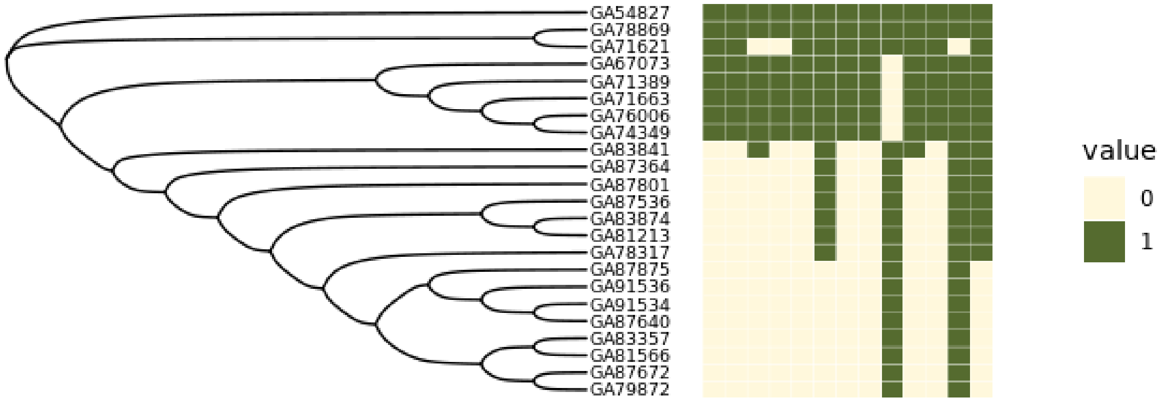


Figure 3: The pangenome consisted of 6,560 gene families, of which 1368 were core (>= 95% genomes), 1107 intermediate accessory (95% < x <= 5 %) and 4085 rare accessory (x > 5%)

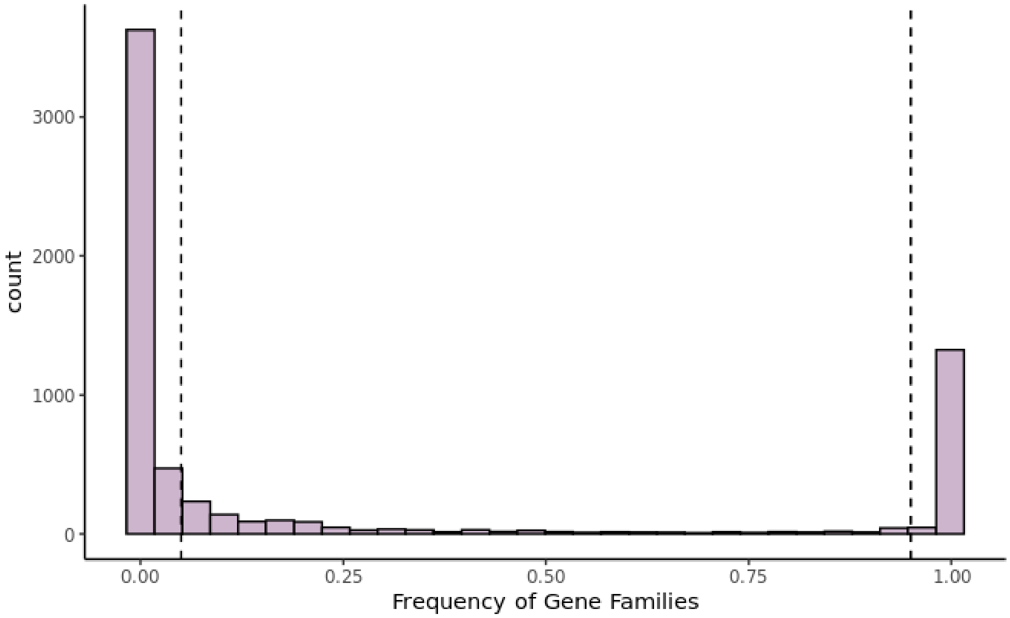


Figure 4: C1 and C2 isolates from 2017-2018 are from a closely related clade

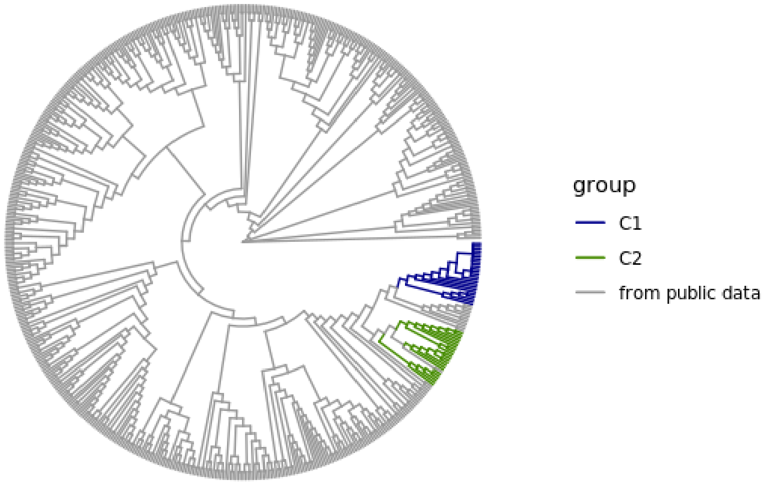


Figure 5: Heatmap of the gained and lost gene families

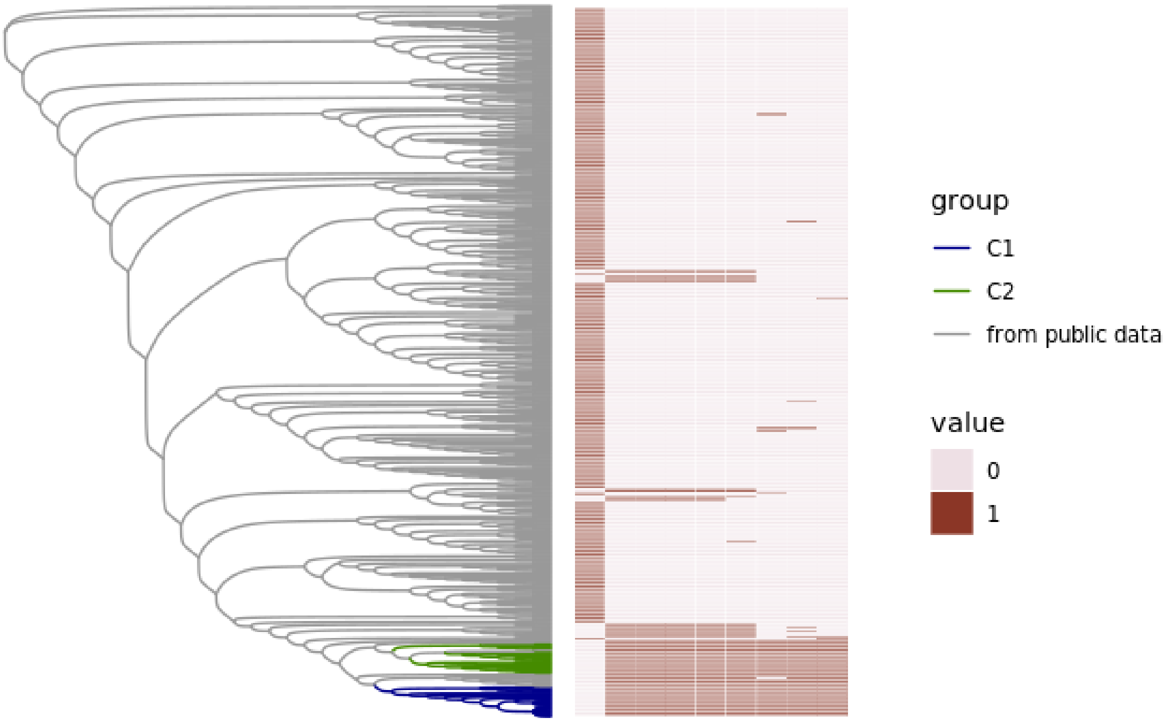
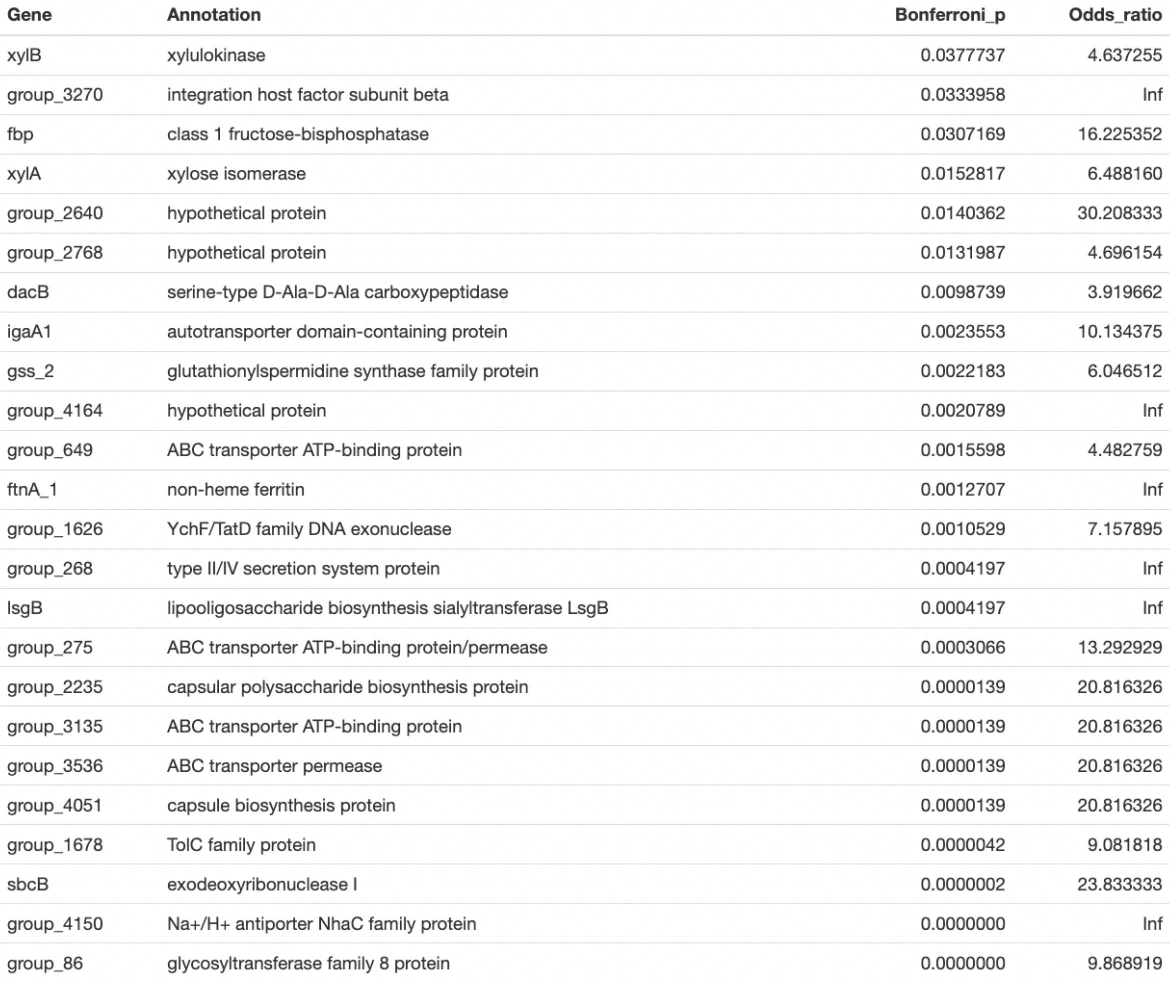


Table 1: Value the 1 lost and 8 gained gene families by consistency index

| gene_family | ID | ConsistencyIndex |
|---|---|---:|
| g00535 | lost | 0.1428571 |
| g05330 | gained | 0.1428571 |
| g02873 | gained | 0.1428571 |
| g04075 | gained | 0.1428571 |
| g02801 | gained | 0.1428571 |
| g06057 | gained | 0.1428571 |
| g02241 | gained | 0.1111111 |
| g03808 | gained | 0.1111111 |
| g01853_1 | gained | 0.2000000 |

## Table 2: 24 significant genes from pan-GWAS



| Gene | Annotation | Bonferroni_p | Odds_ratio |
|---|---|---|---|
| xylB | xylulokinase | 0.0377737 | 4.637255 |
| group_3270 | integration host factor subunit beta | 0.0333958 | Inf |
| fbp | class 1 fructose-bisphosphatase | 0.0307169 | 16.225352 |
| xylA | xylose isomerase | 0.0152817 | 6.488160 |
| group_2640 | hypothetical protein | 0.0140362 | 30.208333 |
| group_2768 | hypothetical protein | 0.0131987 | 4.696154 |
| dacB | serine-type D-Ala-D-Ala carboxypeptidase | 0.0098739 | 3.919662 |
| igaA1 | autotransporter domain-containing protein | 0.0023553 | 10.134375 |
| gss_2 | glutathionylspermidine synthase family protein | 0.0022183 | 6.046512 |
| group_4164 | hypothetical protein | 0.0020789 | Inf |
| group_649 | ABC transporter ATP-binding protein | 0.0015598 | 4.482759 |
| ftnA_1 | non-heme ferritin | 0.0012707 | Inf |
| group_1626 | YchF/TatD family DNA exonuclease | 0.0010529 | 7.157895 |
| group_268 | type II/IV secretion system protein | 0.0004197 | Inf |
| lsgB | lipooligosaccharide biosynthesis sialyltransferase LsgB | 0.0004197 | Inf |
| group_275 | ABC transporter ATP-binding protein/permease | 0.0003066 | 13.292929 |
| group_2235 | capsular polysaccharide biosynthesis protein | 0.0000139 | 20.816326 |
| group_3135 | ABC transporter ATP-binding protein | 0.0000139 | 20.816326 |
| group_3536 | ABC transporter permease | 0.0000139 | 20.816326 |
| group_4051 | capsule biosynthesis protein | 0.0000139 | 20.816326 |
| group_1678 | TolC family protein | 0.0000042 | 9.081818 |
| sbcB | exodeoxyribonuclease I | 0.0000002 | 23.833333 |
| group_4150 | Na+/H+ antiporter NhaC family protein | 0.0000000 | Inf |
| group_86 | glycosyltransferase family 8 protein | 0.0000000 | 9.868919 |

## Figure 6: 24 significant genes count in all samples

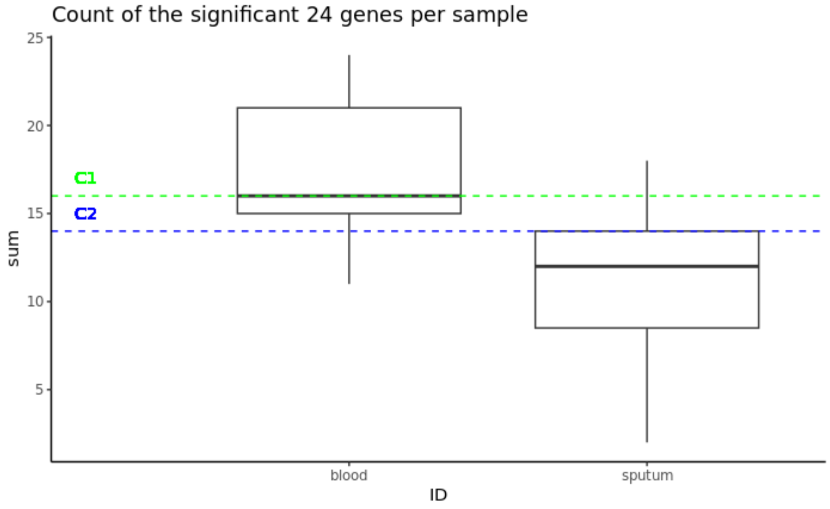


Figure 7: rho/theta values of recombination detected in 50 sequence types of *Haemophilus influenzae*.

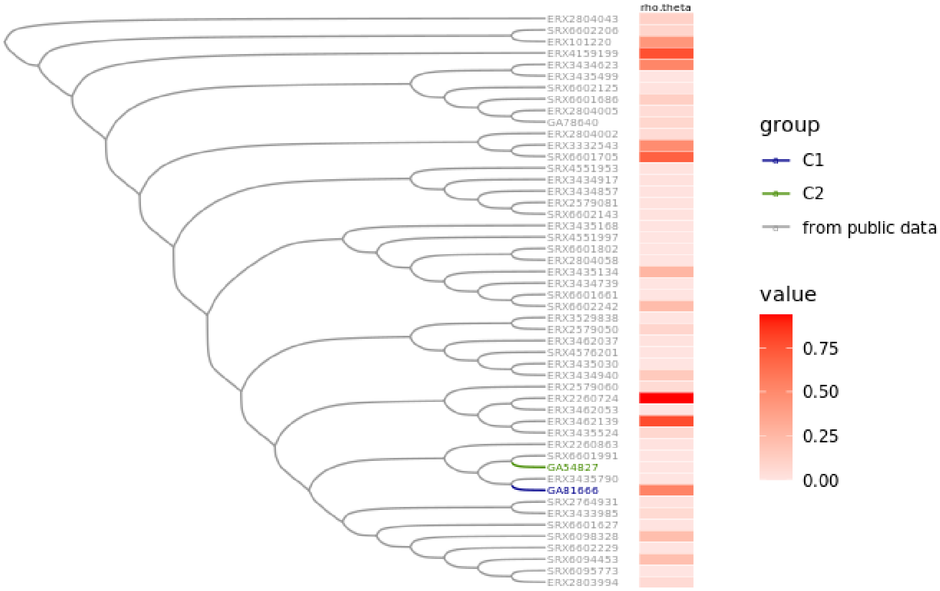
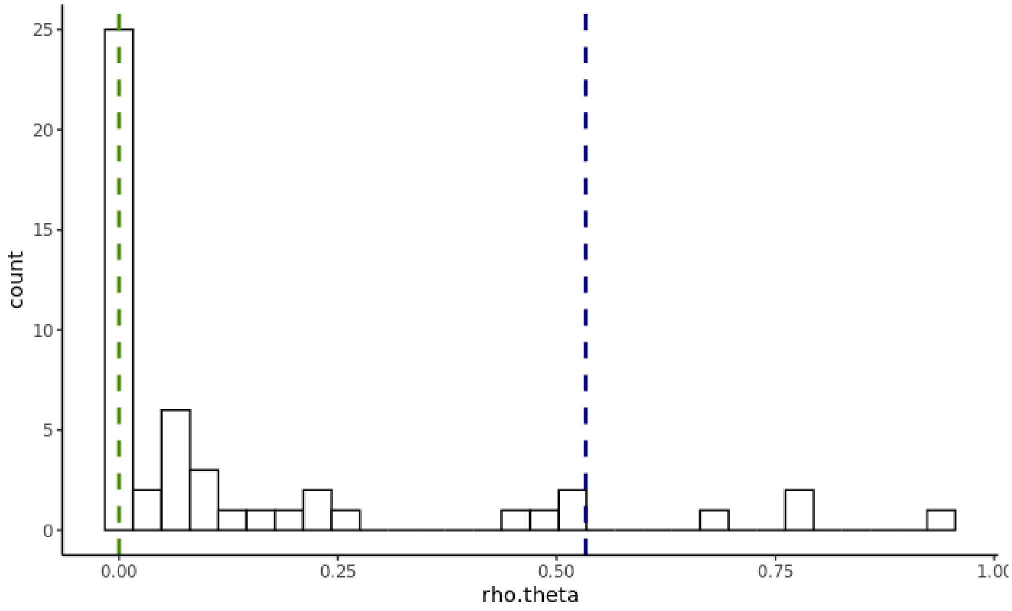
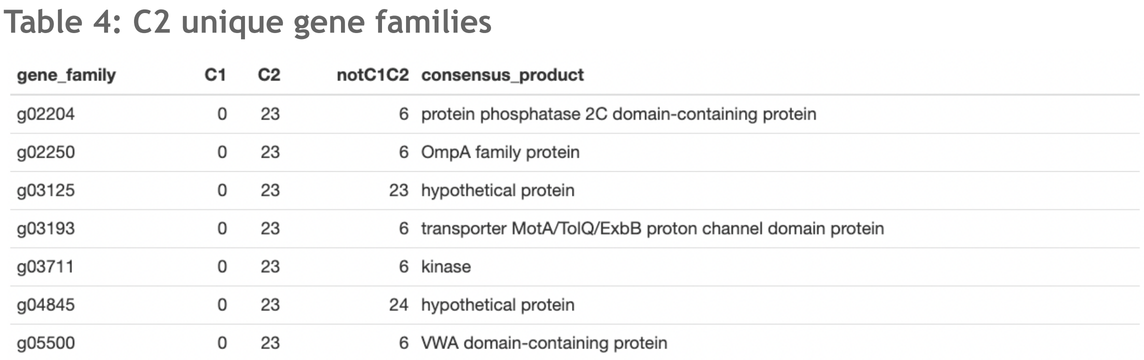


Figure 8: Distribution of rho/theta values with C1 and C2 values represented by vertical lines, where the blue line represents the value of C1 and the green line represents the value of C2.



## Table 3: C1 unique gene families

| gene_family | C1 | C2 | notC1C2 | consensus_product |
|---|---|---|---|---|
| g00892 | 26 | 0 | 34 | hypothetical protein |
| g01570 | 25 | 0 | 31 | hypothetical protein |
| g01607 | 25 | 0 | 46 | hypothetical protein |
| g01616 | 25 | 0 | 39 | helix-turn-helix domain-containing protein |
| g01925 | 25 | 0 | 34 | type II toxin-antitoxin system RelE/ParE family toxin |
| g01950 | 25 | 0 | 34 | hypothetical protein |
| g02180 | 26 | 0 | 10 | hypothetical protein |
| g02274 | 26 | 0 | 15 | WYL domain-containing protein |
| g02446 | 25 | 0 | 30 | hypothetical protein |
| g02457 | 25 | 0 | 35 | hypothetical protein |
| g02469 | 25 | 0 | 30 | membrane protein insertion efficiency factor YidD |
| g02484 | 25 | 0 | 50 | hypothetical protein |
| g02486 | 25 | 0 | 30 | hypothetical protein |
| g02892 | 26 | 0 | 11 | hypothetical protein |
| g04177 | 26 | 0 | 18 | transcriptional regulator |
| g04683 | 26 | 0 | 11 | YkgJ family cysteine cluster protein |
| g04692 | 26 | 0 | 18 | hypothetical protein |
| g05380 | 25 | 0 | 27 | hypothetical protein |
| g05657 | 25 | 0 | 1 | type 1 fimbrial protein |
| g06203 | 25 | 0 | 27 | hypothetical protein |

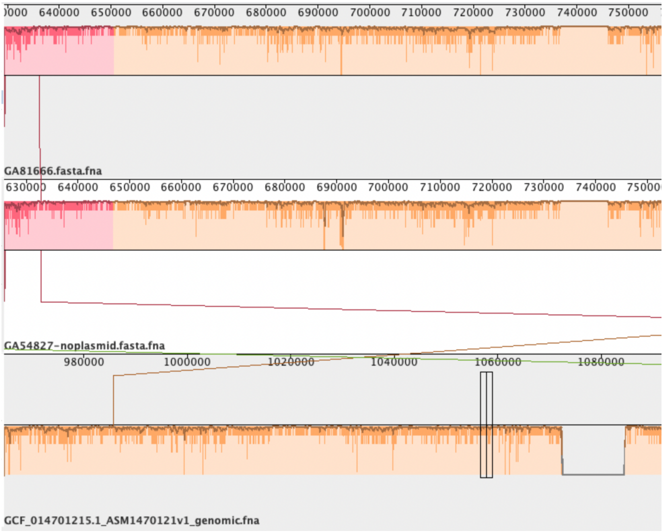


## Table 4: C2 unique gene families

| gene_family | C1 | C2 | notC1C2 | consensus_product |
|---|---|---|---|---|
| g02204 | 0 | 23 | 6 | protein phosphatase 2C domain-containing protein |
| g02250 | 0 | 23 | 6 | OmpA family protein |
| g03125 | 0 | 23 | 23 | hypothetical protein |
| g03193 | 0 | 23 | 6 | transporter MotA/TolQ/ExbB proton channel domain protein |
| g03711 | 0 | 23 | 6 | kinase |
| g04845 | 0 | 23 | 24 | hypothetical protein |
| g05500 | 0 | 23 | 6 | VWA domain-containing protein |

## Supplemental figures

Output from mauve for C1, C2, and an outgroup



## REFERENCES

1 Satola Sarah W., Napier Brooke, Farley Monica M. Association of IS1016 with the hia Adhesin Gene and Biotypes V and I in Invasive Nontypeable Haemophilus influenzae. *Infect Immun* 2008;**76**:5221–7. https://doi.org/10.1128/IAI.00672-08.

2 Satola SW, Collins JT, Napier R, Farley MM. Capsule gene analysis of invasive Haemophilus influenzae: accuracy of serotyping and prevalence of IS1016 among nontypeable isolates. *J Clin Microbiol* 2007;**45**:3230–8. https://doi.org/10.1128/JCM.00794-07.

3 Eskola J, Käyhty H, Takala AK, Peltola H, Rönnberg P-R, Kela E, *et al.* A Randomized, Prospective Field Trial of a Conjugate Vaccine in the Protection of Infants and Young Children against Invasive Haemophilus influenzae Type b Disease. *N Engl J Med* 1990;**323**:1381–7. https://doi.org/10.1056/NEJM199011153232004.

4 Soeters HM, Blain A, Pondo T, Doman B, Farley MM, Harrison LH, *et al.* Current Epidemiology and Trends in Invasive Haemophilus influenzae Disease—United States, 2009–2015. *Clin Infect Dis* 2018;**67**:881–9. https://doi.org/10.1093/cid/ciy187.

5 Collins LF, Havers FP, Tunali A, Thomas S, Clennon JA, Wiley Z, *et al.* Invasive Nontypeable Haemophilus influenzae Infection Among Adults With HIV in Metropolitan Atlanta, Georgia, 2008-2018. *JAMA* 2019;**322**:2399–410. https://doi.org/10.1001/jama.2019.18800.

6 Niehaus TD, Elbadawi-Sidhu M, de Crécy-Lagard V, Fiehn O, Hanson AD. Discovery of a widespread prokaryotic 5-oxoprolinase that was hiding in plain sight. *J Biol Chem* 2017;**292**:16360–7. https://doi.org/10.1074/jbc.M117.805028.

7 Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, *et al.* Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed

target site selection. *Cell* 2008;**132**:208–20. https://doi.org/10.1016/j.cell.2007.12.029.

8 Morero NR, Zuliani C, Kumar B, Bebel A, Okamoto S, Guynet C, *et al.* Targeting IS608 transposon integration to highly specific sequences by structure-based transposon engineering. *Nucleic Acids Res* 2018;**46**:4152–63. https://doi.org/10.1093/nar/gky235.

9 Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;**17**:238. https://doi.org/10.1186/s13059-016-1108-8.

10 Tzeng Y-L, Bazan JA, Turner AN, Wang X, Retchless AC, Read TD, *et al.* Emergence of a new Neisseria meningitidis clonal complex 11 lineage 11.2 clade as an effective urogenital pathogen. *Proc Natl Acad Sci U S A* 2017;**114**:4237–42. https://doi.org/10.1073/pnas.1620971114.

11 Gonzalez-Diaz A, Carrera-Salinas A, Pinto M, Cubero M, van der Ende A, Langereis JD, *et al.* Comparative pangenome analysis of capsulated Haemophilus influenzae serotype f highlights their high genomic stability. *Sci Rep* 2022;**12**:3189. https://doi.org/10.1038/s41598-022-07185-5.

12 Carrera-Salinas A, González-Díaz A, Calatayud L, Mercado-Maza J, Puig C, Berbel D, *et al.* Epidemiology and population structure of Haemophilus influenzae causing invasive disease. *Microb Genom* 2021;**7.**: https://doi.org/10.1099/mgen.0.000723.

13 Kumar A, Beloglazova N, Bundalovic-Torma C, Phanse S, Deineko V, Gagarinova A, *et al.* Conditional Epistatic Interaction Maps Reveal Global Functional Rewiring of Genome Integrity Pathways in Escherichia coli. *Cell Rep* 2016;**14**:648–61. https://doi.org/10.1016/j.celrep.2015.12.060.

14 Wang H, Liu L, Cao Q, Mao W, Zhang Y, Qu X, *et al.* Haemophilus parasuis α-2,3-sialyltransferase-mediated lipooligosaccharide sialylation contributes to bacterial pathogenicity. *Virulence* 2018;**9**:1247–62. https://doi.org/10.1080/21505594.2018.1502606.

15 Tierrez A, García-del Portillo F. The Salmonella membrane protein IgaA modulates the

activity of the RcsC-YojN-RcsB and PhoP-PhoQ regulons. *J Bacteriol* 2004;**186**:7481–9. https://doi.org/10.1128/JB.186.22.7481-7489.2004.

16 Lerner TR, Lovering AL, Bui NK, Uchida K, Aizawa S-I, Vollmer W, *et al.* Specialized peptidoglycan hydrolases sculpt the intra-bacterial niche of predatory Bdellovibrio and increase population fitness. *PLoS Pathog* 2012;**8**:e1002524. https://doi.org/10.1371/journal.ppat.1002524.

17 Abdullah MR, Gutiérrez-Fernández J, Pribyl T, Gisch N, Saleh M, Rohde M, *et al.* Structure of the pneumococcal l,d-carboxypeptidase DacB and pathophysiological effects of disabled cell wall hydrolases DacA and DacB. *Mol Microbiol* 2014;**93**:1183–206. https://doi.org/10.1111/mmi.12729.

18 Zúñiga-Ripa A, Barbier T, Lázaro-Antón L, de Miguel MJ, Conde-Álvarez R, Muñoz PM, *et al.* The Fast-Growing Brucella suis Biovar 5 Depends on Phosphoenolpyruvate Carboxykinase and Pyruvate Phosphate Dikinase but Not on Fbp and GlpX Fructose-1,6-Bisphosphatases or Isocitrate Lyase for Full Virulence in Laboratory Models. *Front Microbiol* 2018;**9**:641. https://doi.org/10.3389/fmicb.2018.00641.

19 Naderer T, Ellis MA, Sernee MF, De Souza DP, Curtis J, Handman E, *et al.* Virulence of Leishmania major in macrophages and mice requires the gluconeogenic enzyme fructose-1,6-bisphosphatase. *Proc Natl Acad Sci U S A* 2006;**103**:5502–7. https://doi.org/10.1073/pnas.0509196103.

20 Srinivasan M, Muthukumar S, Rajesh D, Kumar V, Rajakumar R, Akbarsha MA, *et al.* The Exoproteome of Staphylococcus pasteuri Isolated from Cervical Mucus during the Estrus Phase in Water Buffalo (Bubalus bubalis). *Biomolecules* 2022;**12.**: https://doi.org/10.3390/biom12030450.

21 Cano DA, Pucciarelli MG, García-del Portillo F, Casadesús J. Role of the RecBCD recombination pathway in Salmonella virulence. *J Bacteriol* 2002;**184**:592–5. https://doi.org/10.1128/JB.184.2.592-595.2002.

22 Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

23 Sayers S, Li L, Ong E, Deng S, Fu G, Lin Y, *et al.* Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res* 2019;**47**:D693–700. https://doi.org/10.1093/nar/gky999.

24 Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 2005;**33**:D325–8. https://doi.org/10.1093/nar/gki008.

25 Petit RA 3rd, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* 2020;**5.**: https://doi.org/10.1128/mSystems.00190-20.

26 Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;**3**:124. https://doi.org/10.12688/wellcomeopenres.14826.1.

27 Pinto M, González-Díaz A, Machado MP, Duarte S, Vieira L, Carriço JA, *et al.* Insights into the population structure and pan-genome of Haemophilus influenzae. *Infect Genet Evol* 2018. https://doi.org/10.1016/j.meegid.2018.10.025.

28 Bushnell B. *BBMap short read aligner, and other bioinformatic tools*. SourceForge. n.d. URL: https://sourceforge.net/projects/bbmap/ (Accessed 24 September 2019).

29 Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol* 2014;**15**:509. https://doi.org/10.1186/s13059-014-0509-9.

30 Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 2018;**19**:153. https://doi.org/10.1186/s13059-018-1540-z.

31 Seemann T. *Shovill: De novo assembly pipeline for Illumina paired reads*. Github; n.d.

32 Petit RA III. *assembly-scan: generate basic stats for an assembly*. Github; n.d.

33 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**:1043–55. https://doi.org/10.1101/gr.186072.114.

34 Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2016;**44**:D67–72. https://doi.org/10.1093/nar/gkv1276.

35 Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. *JOSS* 2016;**1**:27. https://doi.org/10.21105/joss.00027.

36 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421.

37 Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;**3**:e000131. https://doi.org/10.1099/mgen.0.000131.

38 Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**:5114. https://doi.org/10.1038/s41467-018-07641-9.

39 Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, *et al.* Mashtree: a rapid comparison of whole genome sequence files. *JOSS* 2019;**4**:1762. https://doi.org/10.21105/joss.01762.

40 Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 2019;**8.**: https://doi.org/10.1093/gigascience/giz119.

41 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74. https://doi.org/10.1093/molbev/msu300.

42 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 2018;**35**:518–22.

https://doi.org/10.1093/molbev/msx281.

43 Watts SC, Holt KE. hicap: In Silico Serotyping of the Haemophilus influenzae Capsule Locus. *J Clin Microbiol* 2019;**57**:2001. https://doi.org/10.1128/JCM.00190-19.

44 Kwong J. *maskrc-svg - Masks recombination as detected by ClonalFrameML or Gubbins and draws an SVG*. Github; n.d.

# Chapter 3: COVID-19 co-infection

## INTRODUCTION

### Overview of COVID-19 metatranscriptomic project

Shortly after the discovery of the novel SAR-CoV-2 virus (COVID-19), it spread to over 200 countries and was declared a pandemic by March 2020 by the World Health Organization (WHO) [1–3]. By June of that same year, the United States reported 2 million cases of COVID-19 and by August COVID-19 became the third leading cause of death in the United States. By September 2020 the global death toll reached 1 million people[4]. Within the same year, there were also documented cases of people in the hospital suffering from bacterial or fungal co-infection [5–10]. Our primary question is to characterize the rate of viral co-infection at the onset of COVID-19 infection. We are able to do this by metatranscriptomic sequencing, which gives us the transcriptomic data of the host and microbiome, and the genomic data of the RNA virus infecting that patient.

### COVID-19 genomics

Although the COVID-19 virus was novel, it is related to other positive-sense RNA viruses that have circulated in our populations and have been well studied in the past[1,11,12]. It is part of the coronaviruses (CoVs) family of viruses belonging to the Coronaviridae family of order Nidovirales [11]. They are broken into 4 genera: alpha, beta, gamma, and delta[1,13]. Each of these genera has hosts that they are more likely to infect, alpha and beta primarily affecting mammals, gamma primarily affecting birds, and delta affecting both mammals and birds[1].

CoVs have relatively large RNA genomes, which can be from 26 to 32 kb [14,15]. CoV genomes are single-stranded, positive-sense RNAs. This means that the RNA resembles an mRNA and can be translated into proteins directly. Their relatively large size also means that they can store a range of protein-coding sequences, making their genomes more diverse and interesting than some other viruses that are constrained by the compact and relatively simple structure of their genomes. RNA is translated into proteins in units of open-reading frames (ORFs) and the CoV genomes have several [15]. Specifically, COVID-19 has 6 highly conserved ORFs, including ORF1ab, ORF3, ORF6, 7a, 8, and ORF10. Some of the functional products are transcribed and translated into spike proteins, nucleocapsid, envelope, and membrane proteins[15]. CoVs are also genetically plastic, leading to altered transmission, virulence, host-specificity, and other factors of viral evolution. These changes in the genome can arise from point mutations, insertions, deletions, and recombination events, both within and between strains ([16].

## COVID-19 and Co-infection

Bacterial, fungal, and viral co-infection was observed in the hospital setting early during the COVID-19 pandemic [5]. Fungal infections of invasive pulmonary aspergillosis and candidiasis were observed in immunocompromised or diabetic patients also infected with COVID-19, and it was established that prolonged hospitalization and use of corticosteroids were associated with a higher risk of fungal co-infection [7];[6,8]. These cases were primarily reported during the second wave of COVID-19 in patients in India[7].

A slightly more common occurrence observed in many different countries is co-infection with COVID-19 and a bacterial or viral pathogen. A meta-analysis assessing the prevalence of co-infections available patient hospitalization records of patients admitted for COVID-19 infection was published in May 2020 [5]. Of the bacterial co-infections identified in this study, over

40% of them can be attributed to *Mycoplasma pneumoniae.* Of the viral co-infections identified, RSV, Influenza A, Rhino/Enteroviruses, and Influenza B were the most common, responsible for 10-15% of the viral co-infections. Although they were able to detect some co-infections, they found them to be relatively rare in the hospital setting; their meta-analysis indicated that 7% of hospitalized patients were co-infected with a bacterial pathogen and only 3% were infected with a viral pathogen. Since we know co-infections can lead to more severe symptoms, we might expect the rate of co-infection to be enriched in this population of patients. This study was limited to hospitalized patients that were reported in the literature and was not an unbiased sampling of all patients experiencing COVID-19. Another meta-analysis published in 2022 confirmed that viral co-infection rates were low (5.01%) and the most prevalent viral pathogen observed was influenza viruses and enteroviruses[17].

Our study differs from these in that we are interested in what viral pathogens may be present in patients at the onset of symptoms, without any bias towards people who were hospitalized with severe infection. Also, this study uses the metagenomic approach to capture all possible infectious agents in the nasal cavity, instead of being limited to medical diagnostic methods.

## Candidate Viruses

We selected a set of viruses that are common and well-studied causes of respiratory disease, including adenoviridae, coronaviridae, paramycoviridea, picornaviridea, and orthomyovirdea. Below I will describe some biological context for each family, highlighting the partial strains we selected as co-infection candidates.

Adenovirus (Human mastadenovirus C from the Adenoviridae family of viruses) is a double-stranded DNA genome of 26 to 48 kb [18]. This family of viruses is responsible for a range of diseases, including respiratory infection, gastroenteritis, and epidemic keratoconjunctivitis[19]. Infections affect both adults and children which can be severe, especially in immunocompromised individuals. These viruses, belonging to the Adenoviridae family, are responsible for a significant portion of upper and lower respiratory tract infections, including the common cold, bronchitis, and pneumonia.  Adenoviral respiratory diseases often spread through respiratory droplets, making them highly contagious.

Coronaviridae is another family of viruses that we are widely interested in, including Coronavirus HKU, Coronavirus NL63, Coronavirus 299E, Coronavirus OC43, and SARS-CoV-2 as candidates. This is the same family of positive-sense RNA viruses that were introduced earlier in this chapter because COVID-19 is in the same taxa [14]. This is a diverse group of viruses that primarily cause respiratory illness, ranging from mild to severe presentation.

Respiratory illnesses caused by viruses within the Paramyxoviridae family, namely Human metapneumovirus (HMPV), Parainfluenza virus, and Respiratory syncytial virus (RSV), represent a significant burden on public health. These viruses have a single-stranded, negative-sense RNA genome[20]. These viruses are prominent culprits behind a range of respiratory infections, from mild cold-like symptoms to more severe respiratory distress, especially in vulnerable populations such as young children and the elderly[20]. HMPV, a relatively recently discovered pathogen, shares clinical similarities with RSV and Parainfluenza virus, often causing bronchitis and pneumonia[21]. Parainfluenza viruses are a common cause of croup and bronchiolitis in children, while RSV is known for its impact on infants and can lead to severe bronchiolitis and pneumonia.

Respiratory illnesses caused by very small, RNA viruses within the Picornaviridae family, particularly Rhinovirus and Enterovirus, are a common occurrence and a major contributor to upper respiratory tract infections [22,23]. Rhinoviruses are colloquially known as ‘common cold’, leading to symptoms like runny nose, sore throat, and cough[23]. They are highly contagious and thrive in temperate climates, especially during the cooler months. In contrast, enteroviruses can cause a broader range of respiratory and systemic symptoms, from mild respiratory issues to more severe conditions, such as pneumonia, myocarditis, or meningitis[22]. They are typically spread by fecal-hand-oral contamination and are endemic in warmer climates.

Respiratory illnesses caused by viruses within the Orthomyxoviridae family, specifically Influenza A and Influenza B, are regularly a public health concern [24]. They are a negative-sense, single-stranded RNA virus that regularly circulates in our population[25]. Influenza A and Influenza B are responsible for the annual flu outbreaks, which can vary in severity and are commonly vaccinated against[26]. These infections can vary in severity but typically lead to a wide range of respiratory symptoms, fever, cough, sore throat, and muscle aches[24]. Influenza A has the potential to cause more severe and widespread epidemics and pandemics due to its genomic plasticity, while influenza B is generally responsible for milder outbreaks.

This curated list creates a starting point to explore potential co-infection in our metatranscriptomic data and offers a focused hypothesis to answer from the mountain of information that is retained in these samples.

## Metagenomic Libraries

Metatranscriptomics is a powerful approach used in the field of genomics and microbiology to study the gene expression profiles of entire microbial communities within a specific environment, such as the nasal cavity. When applied to RNA viral genomics, metatranscriptomics provides valuable insights into the diversity, activity, and functional roles of RNA viruses present in the nasal microbiome. Samples were collected at the onset of COVID-19 infection for diagnostic purposes, as patients with COVID-19-like symptoms visited a clinic in the metropolitan Atlanta area. When a COVID-19 infection was confirmed, RNA was extracted from nasal cavity samples and used to create a metatranscriptomic library. The primary goal of the study was to sequence SARS-CoV-2 genomes for public health surveillance. Initially we used a metagenomic approach because there were not robust SARS-CoV-2 specific sequencing approaches - and even after these became available, we continued to perform some metagenomic sequencing to study coinfectionThis RNA sample represents the patient's transcriptome, the microbiome's transcriptome, and RNA genomes of potential nasal RNA viruses.

Host cells in the nasal cavity include epithelial cells, immune cells (neutrophils, macrophages, dendritic cells, and lymphocytes), and mucosal gland cells. These cells play critical roles in immune defense and maintaining the nasal barrier.  I expect the RNA from this group to represent the steady-state measure of all the genes being expressed in these populations of these host cells as they go about their function and life cycle. In addition to host cells, the nasal cavity hosts a diverse microbiome of bacteria and fungi, including Staphylococcus, Streptococcus, Corynebacterium, Candida, Malassezia, and Aspergillus. RNA sequences from this microbiome provide insights into gene expression and species identification. The metatranscriptomic data also includes viral genomes, such as COVID-19. These viral transcripts represent the virus's genome as it replicates and is distinct from host and microbiome RNA. This comprehensive analysis helps us understand the interactions of these components during COVID-19 infection.

# METHODS

## Building the Pipeline

The goal here is to build a pipeline that is not just useful for this study, but would have broad applicability across other projects in clinical metagenomics. In order to achieve this, I spent time making this script as portable and generalizable as possible. The goal of this pipeline is to classify reads by their taxonomy, separate those reads by group, create QC measure, and create output that can be read into R in order to make figures.

There are 4 scripts within this pipeline that are all orchestrated by a shell script that requires the path to a manifest file that stores the following information in a space-delimited fashion:

1. R1 file path: Path to the first read file (fastq format).
2. R2 file path: Path to the second read file (fastq format).
3. Prefix: A prefix to be used in the output file names.
4. Batch: Information about the dataset batch or any other relevant details.

The runner script will call and orchestrate the inputs and outputs of 4 scripts that pre-process the data, classify the reads, separate the reads by their classification, calculate QC, and modify the reports so that downstream figures and analysis can be done. The script logs the progress of each dataset processing by writing information to a file called progress. This includes the commands executed and the number of files generated for each dataset and provides a record of the script's execution. This script automates the processing and analysis of paired-end sequencing data. It iterates through a list of data manifests, runs several processing and analysis scripts on each dataset, and logs the results. This is particularly useful for batch processing multiple sequencing datasets with consistent file structures.

The first script that the runner script calls does most of the heavy lifting of this pipeline. This script is designed to process paired-end sequencing data, deduplicate and trim reads, classify them using Kraken2, and extract specific taxonomic groups of interest[27–29]. Kraken2 is a taxonomic sequence classier that examines k-mers within a fastq against a database in order to add taxonomic labels to those sequences. The first section uses Kraken2 to classify the reads and Bracken to summarize the unfiltered dataset, producing Kraken and Bracken reports for further analysis. Bracken, Bayesian Reestimation of Abundance with KrakEN, is a software package by the same group that adjusts abundance of taxonomic groups that were outputed by Kraken2[30]. The next section of the script extracts reads corresponding to specific taxonomic groups of interest (e.g., human, bacteria, fungus, virus, COVID) and writes them to separate fastq files. Reads that do not belong to the specified group are also extracted. The split files are then re-run through Kraken2 and Bracken to generate report and output files for each split file. This output would be important to use if i was interested in the taxonomic grouping of all reads classified as viruses for each sample.

The second script that the runner script calls is a script that counts the reads in the fastq file as it progresses through the pre-processing and quality control steps. The third and fourth script reformat the Kraken and Bracken reports for figures.

All downstream data manipulation was done through R and custom awk/shell scripts. One of these manipulations was to format a csv file with demographic and clinical data to join this to the output reports from Kraken and Bracken.

## Running the Pipeline

For the 846 metatranscriptomic samples, the reads were first passed through a pre-processing pipeline where the reads were deduplicated with Clumpify.sh in the BBMap tools [31]. Deduplicated reads were trimmed with Trimmomatic Version 0.40 and filtered for quality, with flags leading:3, trailing:3, slidingwindow:4:15, minlen:36[32]. Each of these pre-processing steps are designed to drop duplicated or low-quality reads, making the total read count in each sample drop. In order to assess how many reads were being lost at each step compared to the original, we counted the reads and plotted the output numbers of each step in a boxplot.

Pre-processed reads were run through kraken2 v2.1.3 against the k2_pluspf_20210127 database to assign each read to a taxonomic group, then adjusted for significance with Bracken. The k2_pluspf_20210127 database is a pre-made kraken2 database that stores information to identify taxonomic groups that could be represented in the sample. This particular database can classify Refeq archaea, bacteria, viral, plasmid, human1, UniVec_Core, Refeq protozoa and fungi[33]. Within the Kraken Tools packages, the extract_kraken_reads.py script was used to separate reads by taxonomic ID for human taxID_hg="9606", bacteria taxID_bac="2", fungus taxID_fungus="4751", viruses taxID_virus="10239", and COVID-19 taxID_COVID="2697049".

Our main questions were to identify potential coinfections with the metatranscriptomic data and understand if these cases lead to more severe clinical outcomes. Custom shell and R scripts were used to determine if the following viruses, bacteria, and fungi were found in each sample:

| Pathogen | Type | Tax ID |
|---|---|---|
| Human mastadenovirus C | virus | 129951 |
| Coronavirus HKU1 | virus | 443239 |
| Coronavirus NL63 | virus | 277944 |
| Coronavirus 299E | virus | 11137 |

| | | |
|---|---|---:|
| Coronavirus OC43 | virus | 31631 |
| SARS-CoV-2 | virus | 2697049 |
| Paramyxoviridae | virus | 11158 |
| Human metapneumovirus | virus | 162145 |
| Parainfluenza virus | virus | 2905673 |
| Respiratory syncytial virus | virus | 12814 |
| Picornaviridae | virus | 12058 |
| Rhinovirus | virus | 31708 |
| Enterovirus | virus | 12059 |
| Orthomyxoviridae | virus | 11308 |
| Influenza A | virus | 382835 |
| Influenza B | virus | 11520 |
| Mycoplasma pneumoniae | bacteria | 2104 |
| Pseudomonas aeruginosa | bacteria | 287 |
| Haemophilus influenzae | bacteria | 727 |
| Klebsiella pneumoniae | bacteria | 573 |
| Enterobacter | bacteria | 547 |
| Acinetobacter baumannii | bacteria | 470 |
| Chlamydia | bacteria | 810 |
| Enterococcus faecium | bacteria | 1352 |
| Staphylococcus aureus | bacteria | 1280 |
| Serratia marscecens | bacteria | 615 |
| Aspergillus | fungus | 5052 |
| Candidia | fungus | 160764 |

## Processing samples in AWS

Samples were processed with an EC2 instance on AWS. To start and use a Linux EC2 (Elastic Compute Cloud) instance on AWS (Amazon Web Services) cloud computing, begin by signing into your AWS account and accessing the AWS Management Console. Launch an EC2 instance, selecting a Linux-based Amazon Machine Image (AMI) that suits your needs. Configure the instance type, adjust network settings, and allocate storage as required. Don't forget to configure security groups to allow SSH access. You can use default settings in many cases but customize them to fit your specific use case. Optionally, add tags for organization and identification. Review your settings, create a key pair for SSH access, and proceed to launch the instance. After the instance starts, connect to it using SSH, and you'll have complete control over your Linux EC2 instance. Install software, configure settings, and run applications as needed. Secure your instance further, set up user accounts, and manage OS updates. AWS offers comprehensive documentation and resources to help you manage and maintain your EC2 instance, allowing you to harness the scalability and power of cloud computing for your Linux-based projects and applications.

# RESULTS

## Sample demographics

We have processed 846 samples from 2021 processed through this pipeline. The output files were joined with the metadata that was collected by our lab and our collaborators. Although there is a portion of the samples that do not have corresponding metadata, there are enough samples that we can infer general trends in the data.  Samples that have available metadata were collected from June to September, primarily in September **(Figure 1)**. These patients were also predominantly women **(Figure 1).**

Another piece of metadata to consider is the clinical Ct values for each patient's sample. Clinical Ct (cycle threshold) values represent a measure of the number of PCR (Polymerase Chain Reaction) cycles required for the amplification of a target nucleic acid sequence in a patient's sample to reach a detectable level. Lower Ct values indicate a higher quantity of the target sequence in the sample, meaning the virus is present at a higher level in the sample. Higher Ct values indicate the opposite, lower concentration of target sequence and therefore lower amount of viral load. In the context of testing for COVID-19, a lower Ct value can represent a higher viral load, more active or severe infection, in this case at the onset of symptoms. When we observe clinical ct values by age, there is a general trend toward lower Ct values in the young and old populations of people **(Figure 2)**. This is what we might expect considering that this is the population of people that typically develop severe infections. It is also interesting that we observe this trend at the onset of symptoms. We also observe that early infections have a wider and lower range of clinical Ct values than later infections, suggesting the possibility of acquired immunity or better interventions later in the pandemic **(Figure 3).**

## Assessing the quality of the reads

As part of quality control measures, reads were de-duplicated, trimmed, and then filtered for read quality and length post-trimming. In order to understand if the reads within the sample withstood these quality control steps, reads in each sample were counted. As shown in figure 4, the majority of the reads were kept in the deduplicated and trimmed-pair intermediate files **(Figure 4)**. This also suggests that the samples were good quality and the libraries were not over-amplified in their preparation steps.

Another important component of sample quality is to determine how many of the reads are able to be identified in the Kraken database and if we are observing the taxa we might expect given that these samples were collected from the nasal canal of patients. Kraken is a software designed to assign every read within a sample to a taxonomic level and identification, giving a profile of what cells contributed DNA or RNA to that library[27]. The software generates a report that summarizes the taxons present, the number of reads assigned to this level, the sum of the reads at this level and below it, and the proportion of the reads that are unable to be assigned to any taxon. To understand what proportion of the reads were able to be identified as something within our Kraken database, I plotted that value for each value on a histogram and observed that the majority of our samples had high identification rates **(Figure 5).** To confirm that our samples contained COVID-19 reads and to understand what proportion of those reads were contained in each sample, I plotted the percentage of COVID-19 reads on a histogram **(Figure 6).** We can observe that most samples have less than 10% of their reads assigned to COVID, but some of the samples had an incredibly high proportion of those reads assigned to COVID-19.

## Validating Kraken2 assignments

To validate the Kraken assignments, I plotted the percentage of COVID-19 reads assigned by Kraken against the Clinical Ct value of those same samples **(Figure 8).** In line with our expectations, at low Ct values there is a very high percentage of COVID-19 reads assigned by Kraken, and as the Ct value increases the percentage of reads assigned to COVID-19 decreases **(Figure 8).** This negative correlation is exactly what we might expect to see in our data and is a solid and independent confirmation that what we are observing in the Kraken data is accurate. To understand what percentage of reads were being assigned to the major groups we expect in our samples, Figure 7 plots the human, bacterial, or viral reads for every sample in either a stacked bar plot to highlight the total amount these 3 taxa are responsible for **(Figure 7).** There is also the same plot split into 3 bar plots for each taxa, where we can observe that for the majority of the samples human reads dominate the sample, viral reads are present in high amounts but are consistently less than human, and bacterial reads typically account for a small portion of the reads **(Figure 7).**

## Searching for Viral Co-infection

Most, but not all, of the samples, had at least 1 read assigned to COVID-19, with an average of 34,354 reads assigned to this taxa **(Table 1)**. A small portion of the samples also contained at least one read assigned to a candidate virus, 39 samples contained Human mastadenovirus C reads, 27 samples contained Paramyxoviridae, 5 samples contained Picornaviridea, 3 samples contained Enterovirus, 2 samples contained Human coronavirus NL63, and 1 sample contained Human coronavirus 229E. It should be noted that although these samples were positively identified with Kraken, some had a very low read count **(Table 1)**. When the read counts of these viruses were correlated with the read counts of COVID-19 in the same sample, we found that every one of the samples that came up as positive for another virus was negative (read count of 0) for COVID-19 **(Figure 9)**. Therefore, this suggests that the cause of their illness is this or another virus and not COVID-19.

## Expanding the search to Bacterial and Fungal Co-infections

Candidates for bacterial co-pathogens were chosen directly from a meta-analysis published early in the pandemic in 2020 [5]. Fungal species that were candidates for were also identified directly from what was observed and published in the literature [8]. We detected 3 bacterial species that were present in patients with varying COVID-19 reads, however, there was no correlation between the number of COVID-19 reads and the number of reads from that pathogen **(Figure 10)**. There are also 6 bacterial pathogens that are identified in our samples, but only in samples with 0 COVID-19 reads **(Figure 11)**. There was one fungal pathogen identified in the samples, but again no correlation between the number of reads assigned as Aspergillus and viral load **(Figure 12).**

# CONCLUSIONS

## No detected viral co-infections

To find no viral co-infections of these candidates in over 846 patients at the onset of COVID-19 symptoms is a surprising negative result. While the instances of co-infection in COVID-19 patients was lower than what can be observed in previous influenza outbreak, it was still prevalent enough to be observed in the hospitalized population and even make the New York Times headlines [5,7,34].

The high rates of reads that were able to be positively identified within Kraken2 database suggest that this finding could be a true negative result **(Figure 5**). There are times when a negative result leaves you with the disjunction of there being accurately negative/not present or I am not able to detect anything to confirm its presence. Given that we were able to classify almost all of the reads within these samples, the latter seems less likely to be true.

## Bacterial and Fungal pathogens

While there were some fungal and bacterial pathogens identified, they were either in samples with 0 COVID-19 reads or they showed no correlation with viral load. While it is interesting that we are able to observe some of these species in the nasal microbiome at the onset of COVID-19 infection, there is no evidence to support that the presence leads to a higher viral burden in the patient. If we were to explore this hypothesis more these identifications would need to be confirmed with an identification tool independent from Kraken2.

## Challenges with meta-transcriptomic data

The major challenge for this project was the sheer volume of comparisons that can be made from the data. Another challenge is determining which ones were false positives, especially with viral pathogens because the detectable genomic material is so low that we are getting very few reads above the noise.

## Future directions

Metatrancripttomic data is incredibly rich, giving us insights into host immune response, and bacterial and fungal microbiome, on top of our initial question of viral co-infection. It would be interesting to take an agnostic approach to correlate the presence or absence of any constituent of the microbiome with the increase of COVID-19. While this question is simple to pose, it has proved itself to be a very challenging question to code and assess at a large scale. There are so many identifications within one sample at different taxonomic levels obscuring any real signal with a mountain of noise.

Another possible experiment would be to rank samples by the number of COVID reads to a human gene set enrichment to understand what pathways are activated during an immune response that is able to keep COVID growth at bay (or just an early response) compared to an immune response when viral replication has progressed.

# FIGURES

**Figure 1**: Sample demographics.

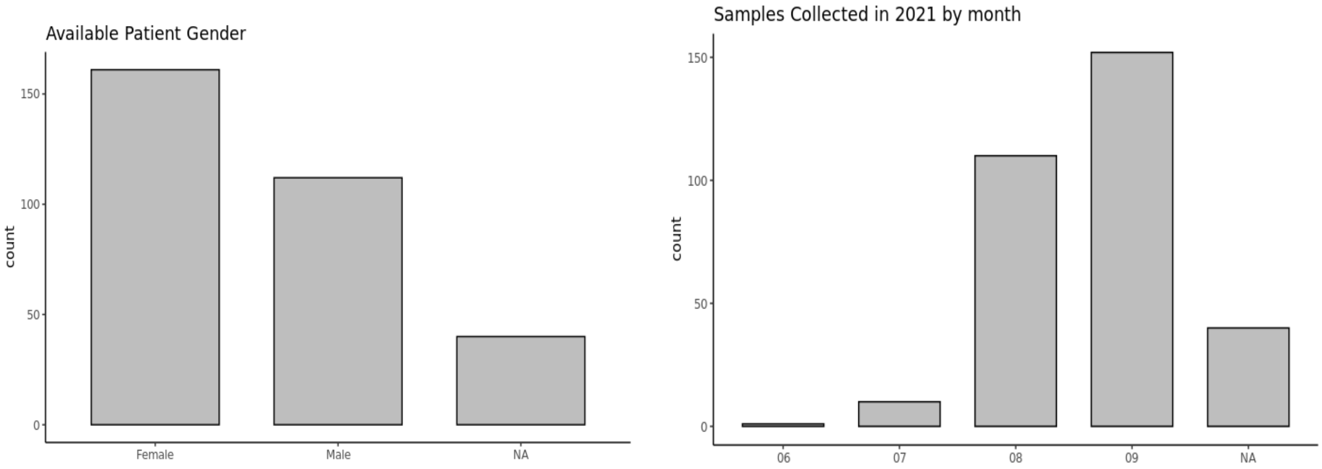


**Figure 2**: COVID Ct by age

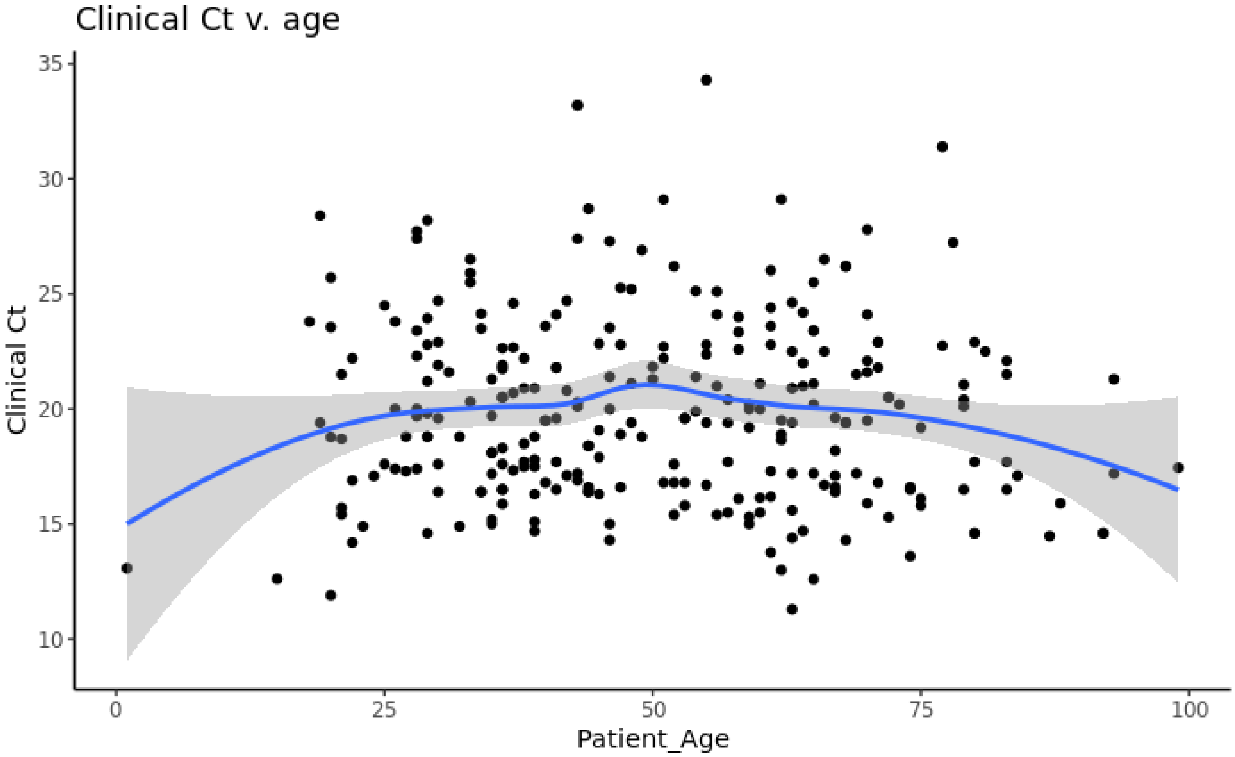


**Figure 3**: COVID Ct over time

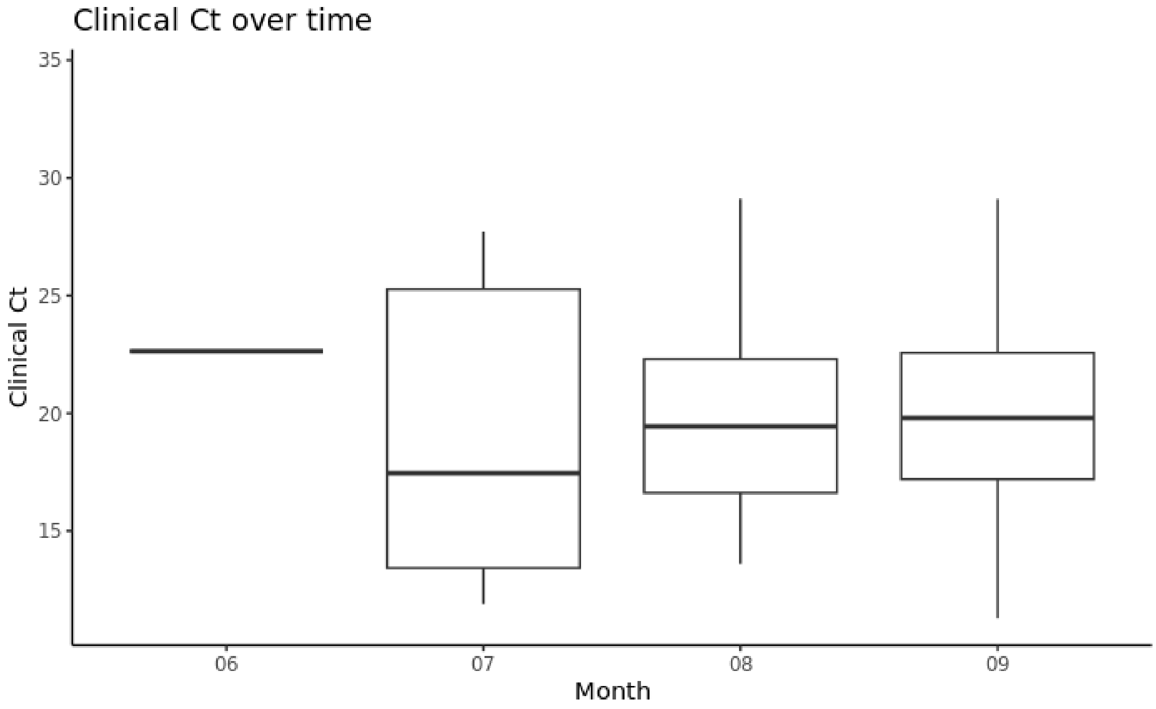


**Figure 4:** As part of quality control steps, the number of reads were counted at each preprocessing step to determine the number of reads that are being retained.

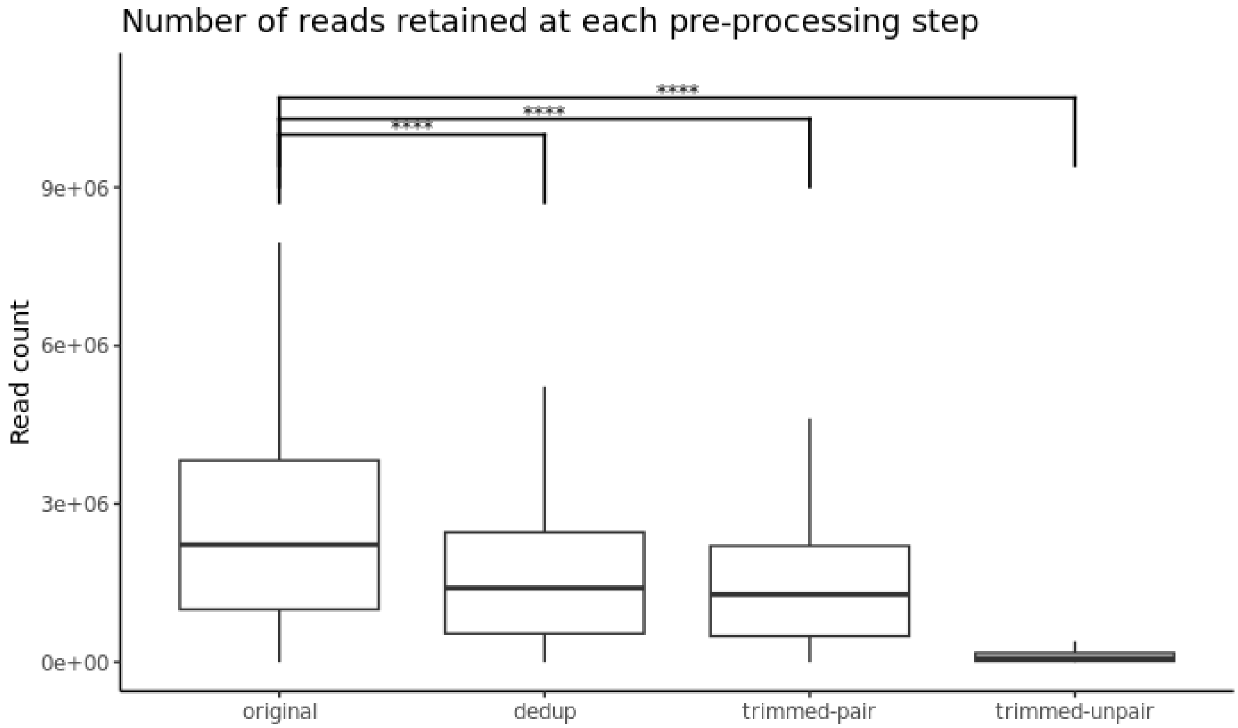


Figure 5: Reads identified by Kraken2

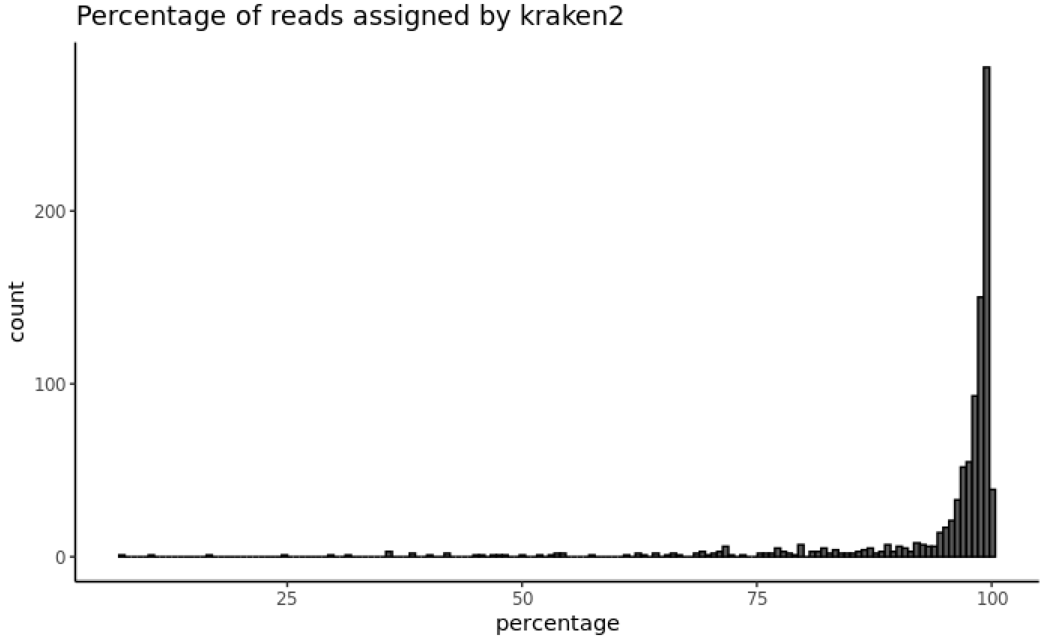


Figure 6: COVID-19 reads identified by Kraken2

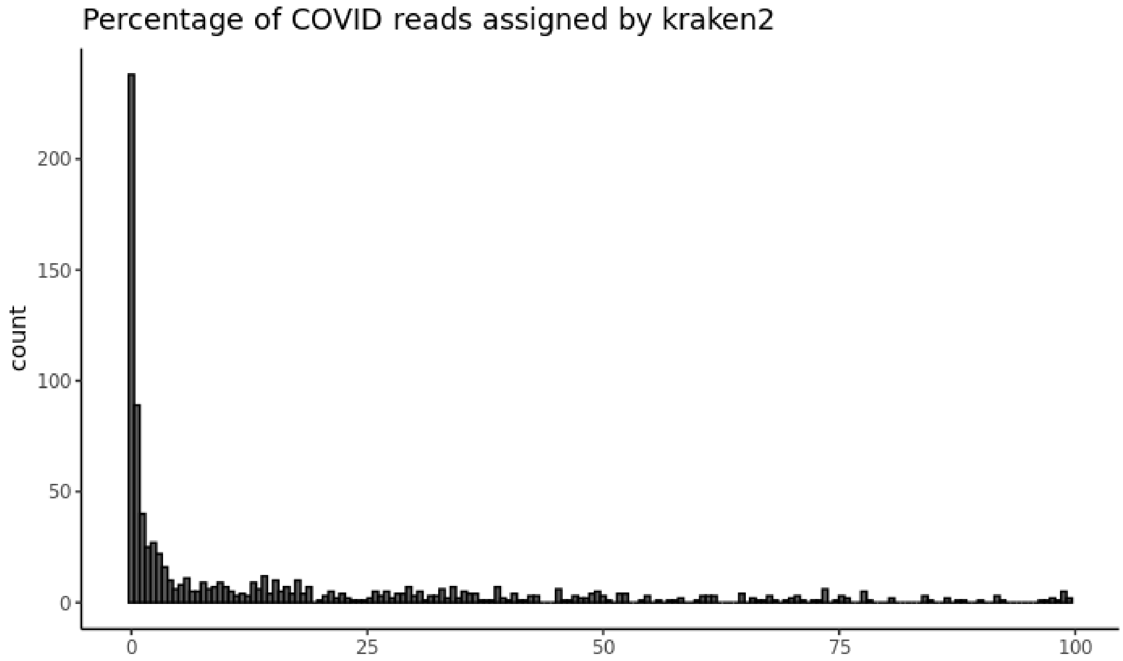


Figure 7: Bacterial, Human, and Viral reads identified by Kraken2

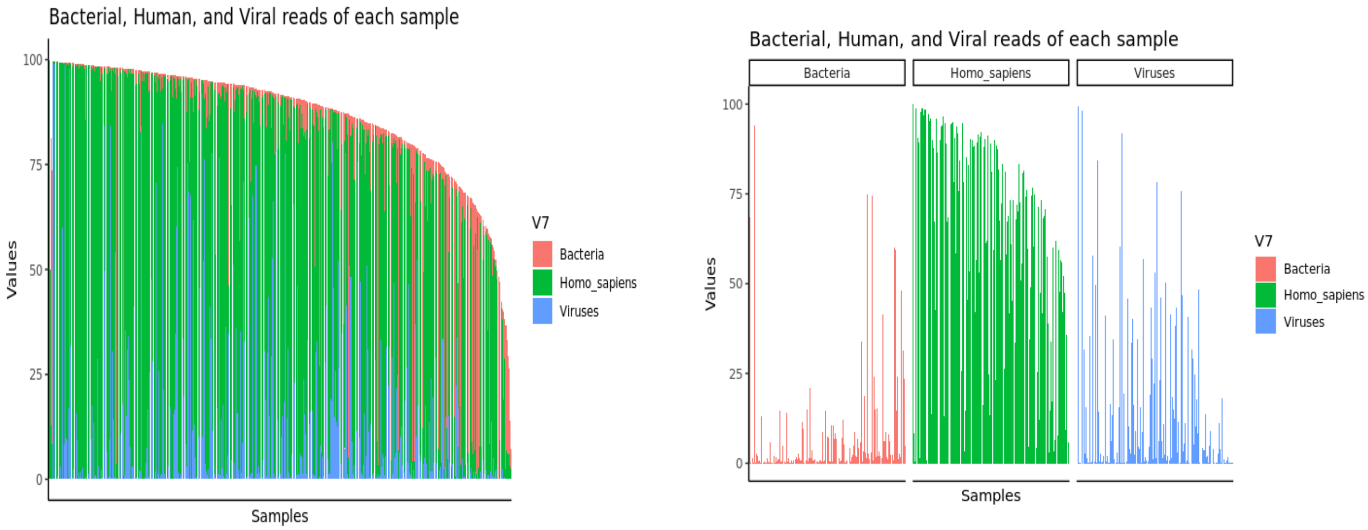


Figure 8:Clinical Ct obtained by the lab against reads identified as COVID-19 by Kraken2 in the meta-transcriptomic libraries.

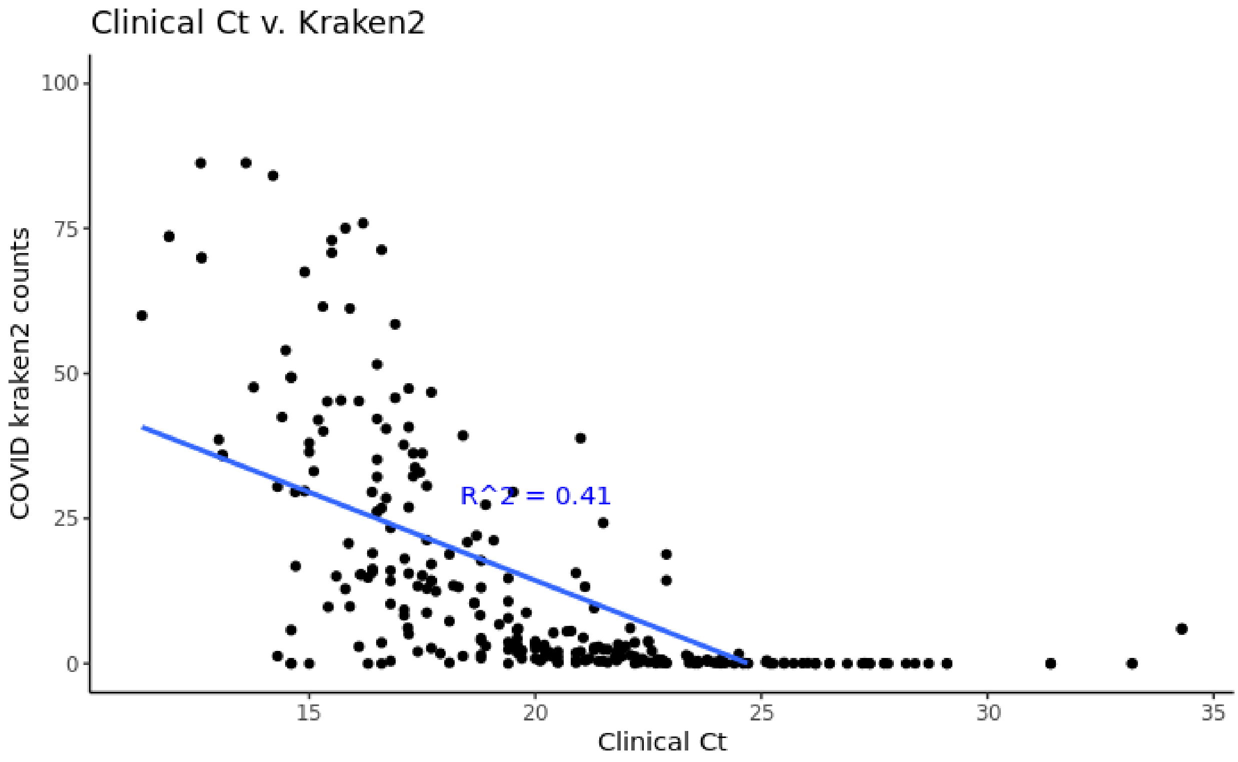
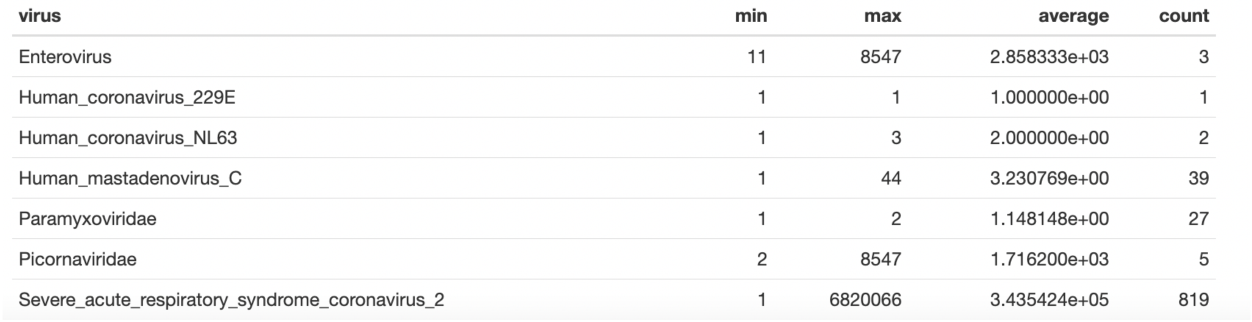
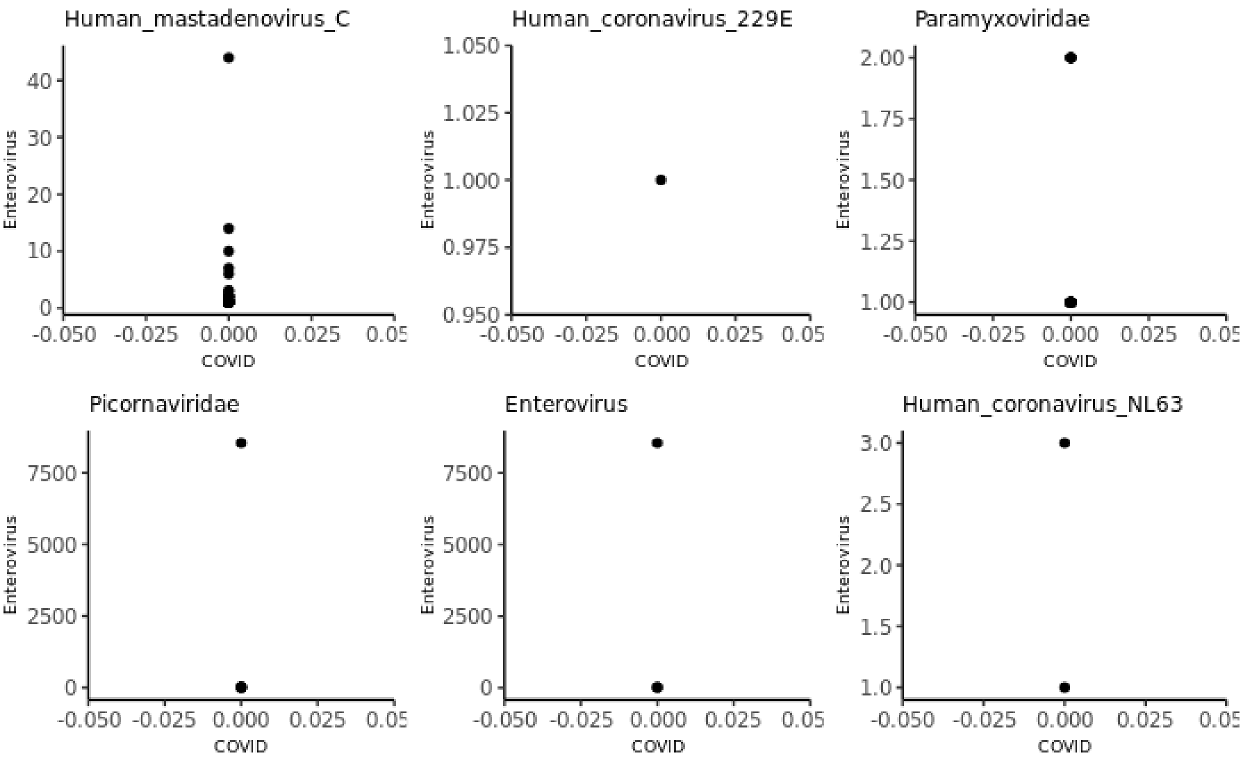


Table 1: Candidate viruses detected in meta-transcriptomic samples

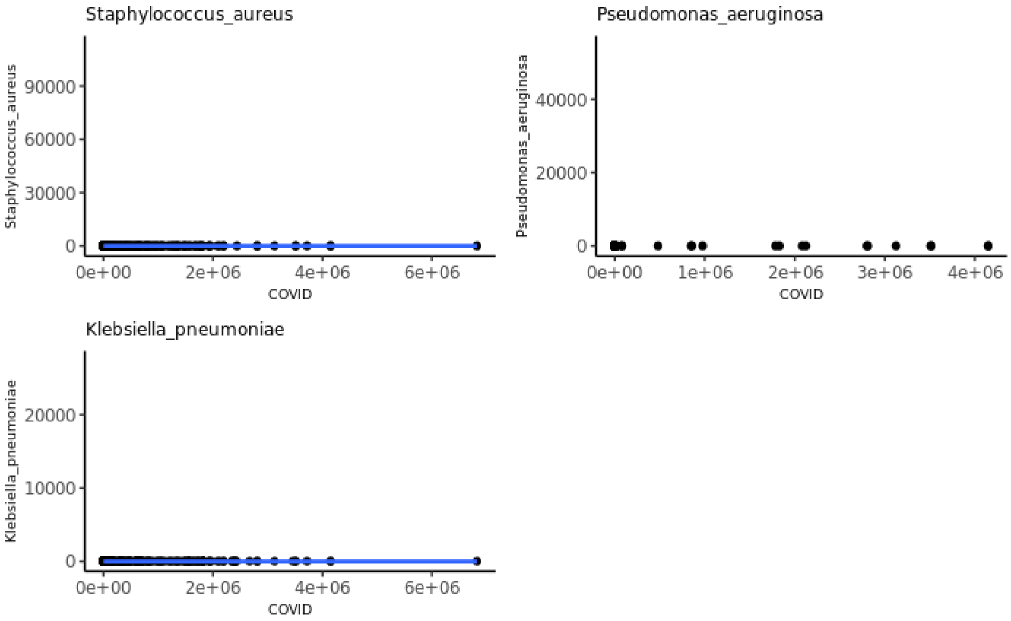


| virus | min | max | average | count |
|---|---:|---:|---:|---:|
| Enterovirus | 11 | 8547 | 2.858333e+03 | 3 |
| Human_coronavirus_229E | 1 | 1 | 1.000000e+00 | 1 |
| Human_coronavirus_NL63 | 1 | 3 | 2.000000e+00 | 2 |
| Human_mastadenovirus_C | 1 | 44 | 3.230769e+00 | 39 |
| Paramyxoviridae | 1 | 2 | 1.148148e+00 | 27 |
| Picornaviridae | 2 | 8547 | 1.716200e+03 | 5 |
| Severe_acute_respiratory_syndrome_coronavirus_2 | 1 | 6820066 | 3.435424e+05 | 819 |

Figure 9: Reads assigned to candidate viruses by Kraken2 against reads assigned to COVID-19 by Kraken2.


Human_mastadenovirus_C
Human_coronavirus_229E
Paramyxoviridae
Picornaviridae
Enterovirus
Human_coronavirus_NL63
Enterovirus
COVID
-0.050
-0.025
0.000
0.025
0.05
0
10
20
30
40
0.950
0.975
1.000
1.025
1.050
1.00
1.25
1.50
1.75
2.00
0
2500
5000
7500
1.0
1.5
2.0
2.5
3.0

## Figure 10: bacteria in COVID-19 positive patients



## Figure 11: Bacteria in patients with no COVID-19 reads

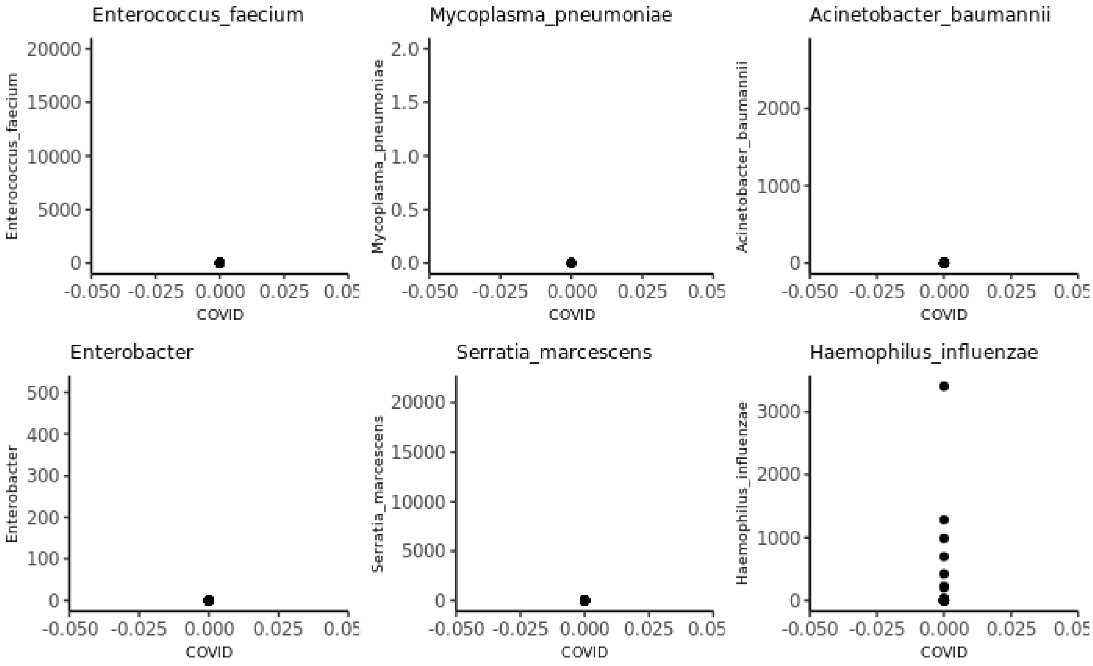
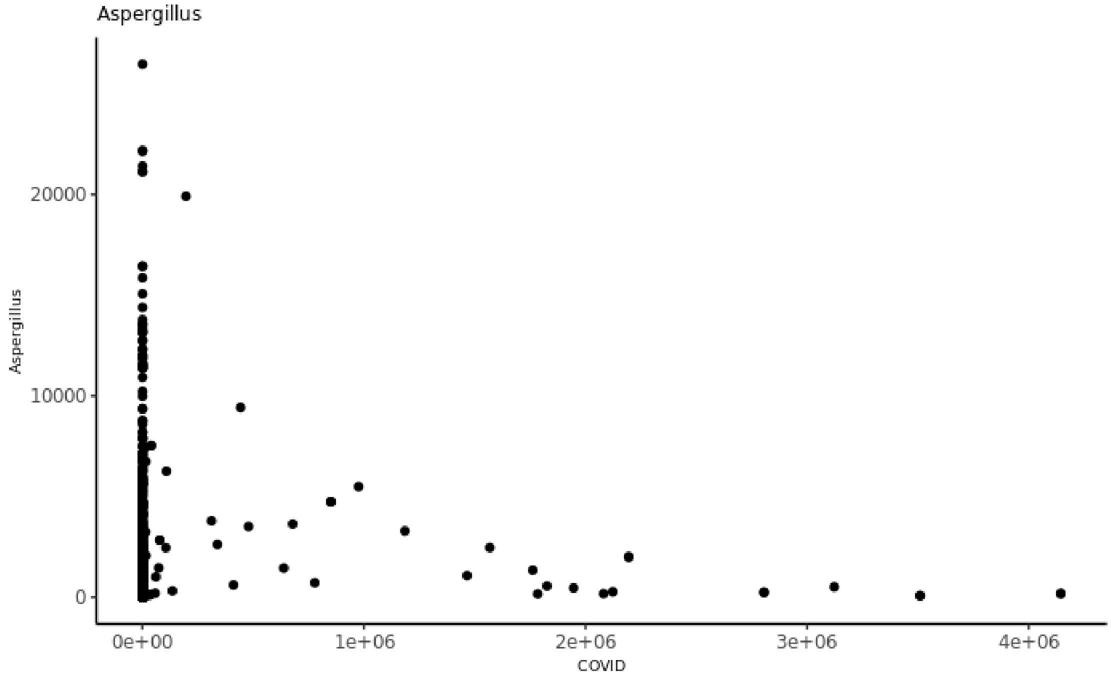


Figure 12: Fungal



# REFERENCES

1. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim. Biophys. Acta Mol. Basis Dis.* **1866**, 165878 (2020).
2. Khan, M. *et al.* COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far. *Molecules* **26**, (2020).
3. Liu, Y.-C., Kuo, R.-L. & Shih, S.-R. COVID-19: The first documented coronavirus pandemic in history. *Biomed. J.* **43**, 328–333 (2020).
4. Hadj Hassine, I. Covid-19 vaccines and variants of concern: A review. *Rev. Med. Virol.* **32**, e2313 (2022).
5. Lansbury, L., Lim, B., Baskaran, V. & Lim, W. S. Co-infections in people with COVID-19: a systematic review and meta-analysis. *J. Infect.* **81**, 266–275 (2020).
6. Rawson, T. M. *et al.* Bacterial and Fungal Coinfection in Individuals With Coronavirus: A

Rapid Review To Support COVID-19 Antimicrobial Prescribing. *Clin. Infect. Dis.* **71**, 2459–2468 (2020).

7. Al-Tawfiq, J. A. *et al.* COVID-19 and mucormycosis superinfection: the perfect storm. *Infection* **49**, 833–853 (2021).
8. Pemán, J. *et al.* Fungal co-infection in COVID-19 patients: Should we be concerned? *Rev. Iberoam. Micol.* **37**, 41–46 (2020).
9. Hoenigl, M. *et al.* COVID-19-associated fungal infections. *Nat Microbiol* **7**, 1127–1140 (2022).
10. Mishra, A., George, A. A., Sahu, K. K., Lal, A. & Abraham, G. Tuberculosis and COVID-19 Co-infection: An Updated Review. *Acta Biomed.* **92**, e2021025 (2020).
11. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
12. Sharma, A., Ahmad Farouk, I. & Lal, S. K. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention. *Viruses* **13**, (2021).
13. Malik, Y. A. Properties of Coronavirus and SARS-CoV-2. *Malays. J. Pathol.* **42**, 3–11 (2020).
14. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
15. Tsai, P.-H. *et al.* Genomic variance of Open Reading Frames (ORFs) and Spike protein in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *J. Chin. Med. Assoc.* **83**, 725–732 (2020).
16. Focosi, D. & Maggi, F. Recombination in Coronaviruses, with a Focus on SARS-CoV-2. *Viruses* **14**, (2022).
17. Krumbein, H. *et al.* Respiratory viral co-infections in patients with COVID-19 and associated outcomes: A systematic review and meta-analysis. *Rev. Med. Virol.* **33**, e2365 (2023).
18. Harrach, B. Adenoviruses: General Features☆. in *Reference Module in Biomedical Sciences* (Elsevier, 2014).

19. Chen, S. & Tian, X. Vaccine development for human mastadenovirus. *J. Thorac. Dis.* **10**, S2280–S2294 (2018).

20. Enders, G. *Paramyxoviruses*. (University of Texas Medical Branch at Galveston, 1996).

21. Uddin, S. & Thomas, M. Human Metapneumovirus. in *StatPearls* (StatPearls Publishing, 2023).

22. Jubelt, B. & Lipton, H. L. Enterovirus/picornavirus infections. *Handb. Clin. Neurol.* **123**, 379–416 (2014).

23. Esneau, C., Duff, A. C. & Bartlett, N. W. Understanding Rhinovirus Circulation and Impact on Illness. *Viruses* **14**, (2022).

24. Pormohammad, A. *et al.* Comparison of influenza type A and B with COVID-19: A global systematic review and meta-analysis on clinical, laboratory and radiographic findings. *Rev. Med. Virol.* **31**, e2179 (2021).

25. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* **16**, 47–60 (2018).

26. Jones-Gray, E., Robinson, E. J., Kucharski, A. J., Fox, A. & Sullivan, S. G. Does repeated influenza vaccination attenuate effectiveness? A systematic review and meta-analysis. *Lancet Respir Med* **11**, 27–44 (2023).

27. Lu, J. *et al.* Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).

28. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

29. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

30. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

31. Bushnell, B. *BBMap: A fast, accurate, splice-aware aligner*.

https://www.osti.gov/biblio/1241166 (2014).

32. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

33. Wood, D. *Manual*. (Github).

34. Richtel, M. Deadly Fungus Spread Rapidly During the Pandemic, C.D.C. Says. *The New York Times* (2023).

# Chapter 4: Conclusion

## NTHi study

In this study we found a number of genes associated with septic infection of *Haemophilus influenzae* (*Hi*). I am particularly interested in this set of genes because some were identified as virulence factors and less was known about others. I can imagine a set of benchwork experiments aimed at understanding the mechanism of virulence is an important advancement in the field.

As more samples of septic arthritis NTHi infections come in, we have a database of factors and genes to compare them against. The 5 gene families that are traveling on the cassette knocking out pxpB are of particular interest to me, even though there is no known virulence factors in this set of gene families. I think benchwork experiments to understand how these genes affect virulence, coupled with a PCR screen of affected patients, could be a really powerful next step for understanding severe NTHi infections.

## COVID-19 study

In the COVID-19 study there is a lot of work left to be done on these samples, mainly because there is so much information within a single sample. There is a lot of validation of Kraken2 calls that could be made, which can be done by funneling these outputs through blast or aligning to reference genomes. This could help us validate samples that seem to be negative by Kraken's more conservative calls and confirm the presence of other co-infecting pathogens in the nasal canal.

I would love to look at the host and microbiome's transcriptomic, to see if any genes are associated with higher viral load. This would give us an idea of what is going on in the host's body during the early stages of severe and less severe cases. It would be so interesting to see if any bacterial virulence factors are over-represented, especially antibiotic resistance genes.

# Appendix

## Previous publications

### Undergraduate researcher in the Ané lab, University of Wisconsin-Madison (2012-2016)

During my time as an undergraduate, I spent over 3 years working on collaborative projects in the labs of Prof. Jean-Michel Ané and Prof. Charles Kaspar. I worked on projects designed to understand how plants perceive and communicate with microbes in the environment. The main model system I used was *Medicago truncatula* and *Rhizobia*. I also studied over 50 different species of plants (algae, liverworts, gymnosperms, angiosperms) in order to analyze the most abundant root exudate and how it differed between plants able to establish symbiotic relationships with bacteria and fungi. My last year of this collaboration focused on the colonization of Medicago by Salmonella, and the study of genomic instability of bacteria in the presence of environmental stresses

a. Wahlig TA, Bixler BJ, Valdés-López O, Mysore KS, Wen J, Ané JM, Kaspar CW. Salmonella enterica serovar Typhimurium ATCC 14028S is tolerant to plant defenses triggered by the flagellin receptor FLS2. FEMS Microbiol Lett. 2019 Feb 1;366(4) PubMed Central PMCID: PMC6420342.

### **Research Technician at Van Andel Research Institute, Grand Rapids, MI (2016, 2017)**

As an intern and research associate at Van Andel Research Institute, I developed a project to understand how differentially methylated regions that are necessary for genomic imprinting persist despite rapid and global DNA demethylation in the zygote and pre-implantation embryo. With the help of my mentor, this project was awarded an institutional innovation award which funded my return to Van Andel to continue developing the methods and aims of the project. I characterized candidate proteins biochemically and designed a method to better understand their binding potential in the genomic material in the zygote.

a. Meng Y, Wang G, He H, Lau KH, Hurt A, Bixler BJ, Parham A, Jin SG, Xu X, Vasquez KM, Pfeifer GP, Szabó PE. Z-DNA is remodelled by ZBTB43 in prospermatogonia to safeguard the germline genome and epigenome. Nat Cell Biol. 2022 Jul;24(7):1141-1153. doi: 10.1038/s41556-022-00941-9. Epub 2022 Jul 4. PMID: 35787683; PMCID: PMC9276527.

## Initial research at Emory University (2017-2020)

My research was focused on understanding the principles of chromatin organization in the germline and the pre-implantation embryo. My contribution to both papers cited was to characterize how certain transcription factors might contribute to chromatin organization. In Jung et al. 2019 I found that in mature sperm CTCF, Znf143, and Smc1 were co-localized at enhancer-enhancer, enhancer-promoter, and promoter-promoter interactions. In Rowley et al. 2019, I used a method developed by our lab to identify loops in a Hi-C matrix, which in mammals are intense point-to-point interactions that are often mediated by CTCF and cohesin. I characterized how the strength of candidate transcription factors that co-localize with CTCF corresponded with the strength of looping in the Hi-C matrix. I also contributed to the analysis of recently published Hi-C datasets in pachytenes in order to understand how compartmentalization and histone modifications can affect interactions between homologs during meiosis.

a. Rowley MJ, Poulet A, Nichols MH, Bixler BJ, Sanborn AL, Brouhard EA, Hermetz K, Linsenbaum H, Csankovszki G, Lieberman Aiden E, Corces VG. Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals. Genome Res. 2020 Mar;30(3):447-458. PubMed Central PMCID: PMC7111518.
b. Jung YH, Kremsky I, Gold HB, Rowley MJ, Punyawai K, Buonanotte A, Lyu X, Bixler BJ, Chan AWS, Corces VG. Maintenance of CTCF- and Transcription Factor-Mediated Interactions from the Gametes to the Early Mouse Embryo. Mol Cell. 2019 Jul 11;75(1):154-171.e5. PubMed Central PMCID: PMC6625867.
c. Jung YH, Wang HV, Ruiz D, Bixler BJ, Linsenbaum H, Xiang JF, Forestier S, Shafik AM, Jin P, Corces VG. Recruitment of CTCF to an Fto enhancer is responsible for transgenerational inheritance of BPA-induced obesity. Proc Natl Acad Sci U S A. 2022 Dec 13;119(50):e2214988119. doi: 10.1073/pnas.2214988119. Epub 2022 Dec 5. PMID: 36469784.

## Guide to mindful server usage

This guide was included as part of our lab manual as commands that can be used to make sure the resources requested by your job does not surpass what could be considered reasonable for a shared resource.

### Example of how to run a shared server

1. Activate conda virtual env
2. Install needed softwares and check installs (usually command --help is sufficient)
3. Run a pilot- only one sample to get an idea of how long/how many cpus is needed, you can monitor by htop -u <userID>
4. Create a loop or shell script to go through commands one-by-one, instead of running them all at once in the background. If you need to parallelize, consider nextflow or using the “nice” command to set the priority of your jobs lower when the server is being used by others.
5. Run with specified cpus whenever possible (even if it is and optional input)
6. Check htop and watch to make sure the cores are not overwhelmed, this means that only a portion of the cores are being used, leaving some of the computing space for the processes that keep the server running.
7. If overworked, pid kill immediately
8. Check outputs and their size
9. Immediately delete intermediates after you have confirmed outputs. This is the most important step because in a few weeks from now you may have a hard time remembering which are extra files and which are the finalized outputs

    a. Note: if you are running a pipeline for the first time and are not sure what can/can't be deleted, create a gzip step for all files and use zcat to comb through them

10. Immediately move raw data to long term storage and check your disk space footprint. This is a shared resource that will only be useful if we keep enough room for everyone to run and create outputs

## AWS guide

### Setting up your EC2 with extra EBS storage

This guide was created with inspiration from generic AWS guide, an email from AWS help at Emory, and brute force. Pink highlighted text is the actual command run on the commandline either on your home computer or within the EC2 instance.

### Mounting EBS for the first time

1. Download *.pem file locally
2. Change the permissions of *.pem file with command chmod 400 *pem
3. Ssh into EC2.
    a. At this point your EBS is visible but not mounted
4. Use the df command to get your starting baseline
    a. the EBS will not be visible
5. Use lsblk command
    a. Note the first column, Name, is the relevant identifier for your EBS. For t3 series this name will be something along the scheme of "nvme1n1" and for this example is will use "EBS_lsblk_name" as it's variable.

   b. Note the last column, Mountpoint, is blank. Again this is a starting baseline that will change if mounting is successful

6. For ubuntu use the sudo apt-get install xfsprogs
   a. It is likely already installed but just in case
7. Get things ready with sudo mkfs -t xfs /dev/EBS_lsblk_name
8. Create the destination of the mount-point by sudo mkdir /data
9. Mount the EBS with sudo mount /dev/EBS_lsblk_name /data
10. You should be good to go
    a. Use lsblk command to confirm that the mountpoint entry in the last column of EBS_lsblk_name row is now populated with /data
    b. Use df to confirm that you can see the /data mount point
    c. Use ls -lt /data to confirm the state of that mountpoint

### Setting up EBS to be automatically mounted in subsequent logins

1. Copy your fstab start-up file for safekeeping with sudo cp /etc/fstab /etc/fstab.orig
2. Create identifier for EBS_lsblk_name with sudo blkid
   a. Copy the UUID=”...” identifier for /dev/EBS_lsblk_name
3. Add the UUID info to lsblk readout by using the command sudo lsblk -o +UUID
   a. There should be an additional column titled “UUID” added
4. Carefully(!) update fstab with this command sudo nano /etc/fstab
   a. This will open the fstab file within the nano text editor
5. Within the fstab document opened with nano, add a newline of information to be read upon start up of this server. The syntax of this line will be UUID=... /data xfs defaults,nofail
   a. This is newline, space delimited, no quotations around UUID number
6. Use sudo umount /data to unmount your EBS

7. Use sudo mount -a to confirm that your addition to th
    a. If not, you can go within the fstab and edit the line you created
    b. If nothing works, please remove the modified fstab and use the mv /etc/fstab.orig /etc/fstab to replace the modified file with the original

## Setting up /data

1. Change ownership of that directory and its contents with common sudo chown -R ubuntu /data
    a. Unless the username was change by the owner of the AMI, it can be foun within the ssh common, i.e. ssh -i "bbixler_1.3.pem" ubuntu@10.66.123.132
    b. Check the ownership before and after with ls -lt /data

## Useful commands to remember

1. To copy files to bucket
    a. aws s3 cp <your_file>  s3://<your_bucket>
2. To create bucket from the commandline of EC2
    a. aws s3 mb s3://<bucket_name>
3. To exchange files between your home computer and an EC2 scp will not work, you need to use an sftp command
    a. sftp username@ip.address:/path/to/files /path/to/destination