

**Metatranscriptomic analysis of pediatric acute
sinusitis: pathogen detection and host response
profiling**

by

Nooran Abu Mazen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2024

© Nooran Abu Mazen 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The work presented in this thesis has been submitted for publication and deposited as a pre-print (see below):

Nasopharyngeal metatranscriptomics reveals host-pathogen signatures of pediatric sinusitis. Nooran AbuMazen, Vivian Chu, Manjot Hunjan, Briallen Lobb, Sojin Lee, Marcia Kurs-Lasky, John V. Williams, William MacDonald, Monika Johnson, Jeremy A. Hirota, Nader Shaikh, Andrew C. Doxey. *bioRxiv* 2024.03.03.24303663; doi: <https://doi.org/10.1101/2024.03.03.24303663>

I would like to thank the following individuals for their collaboration and contributions to this work:

- Vivian Chu, who developed and ran scripts for the initial taxonomic and host transcript analysis on the first two batches of data.
- Dr. Briallen Lobb, who gave expert feedback and recommendations during countless discussions.

Abstract

Acute sinusitis (AS) is the fifth leading cause of antibiotic prescriptions in children. Distinguishing bacterial AS from common viral upper respiratory infections in children is crucial to prevent unnecessary antibiotic use but is challenging with current diagnostic methods. Despite its speed and cost, untargeted RNA sequencing (RNA-seq) of clinical samples from children with suspected AS has the potential to overcome several limitations of other methods. However, the utility of sequencing based approaches in analysis of AS has not been fully explored. Here, we performed RNA-seq of nasopharyngeal samples from 221 children with clinically diagnosed AS to characterize their pathogen and host-response profiles. Results from RNA-seq were compared with those obtained using culture for three common bacterial pathogens and qRT-PCR for 12 respiratory viruses. Metatranscriptomic pathogen detection showed high concordance with culture or qRT-PCR, showing 87%/81% sensitivity (sens) / specificity (spec) for detecting bacteria, and 86%/92% (sens/spec) for viruses, respectively. 22 additional pathogens not tested for in the clinical panel were detected, and plausible pathogens were identified in 11/19 (58%) of cases where no organism was detected by culture or qRT-PCR. 205 viruses were assembled across the samples including novel strains of coronaviruses, respiratory syncytial virus (RSV), and enterovirus D68. By analyzing host gene expression, host-response signatures were identified that distinguished bacterial and viral infections and correlated with pathogen abundance. Ultimately, this study demonstrates the potential of untargeted metatranscriptomics for in depth analysis of the etiology of AS, comprehensive host-response profiling, and using these together to work towards optimized patient care.

Acknowledgements

I would like to thank my supervisor, Dr. Andrew C. Doxey, for providing me the opportunity and support to work on this research project, and from whom I've learned a lot over these 2 years. Your passion for research and commitment to excellence has been a source of inspiration. I would also like to thank my committee members, Dr. Trevor Charles and Dr. Brendan McConkey, for their insightful and constructive feedback. I would like to thank our collaborators at Pittsburgh University, Dr. Nader Shaikh and his colleagues, for their support of and contributions to this research.

I acknowledge NSERC, the Government of Ontario, the University of Waterloo, and the Digital Research Alliance of Canada for their funding and resources, without which my project could not have been as successful and rewarding as it has been.

Lastly, I would like to extend a big thank you to the members of the Doxey lab with whom I have had the pleasure of working alongside. A special shoutout to Harold Hodgins, who helped me many a time with code gone awry and has been a great friend to me.

Table of Contents

Author's Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
List of Figures	ix
List of Abbreviations	xi
1 Introduction and Literature Review	1
1.1 Respiratory Infections	1
1.1.1 Common Respiratory Viruses	2
1.1.2 Sinusitis	5
1.2 Current Pathogen Detection Methods	6
1.3 Metatranscriptomics for Pathogen Detection	7

1.3.1	Overview of Methods	8
1.3.2	Previous work	11
1.4	Hypothesis and Objectives	18
2	Metatranscriptomic Analysis and Results	19
2.1	Methods	19
2.1.1	Study design and description of the cohort	19
2.1.2	Culture and sensitivity pattern of bacterial pathogens	20
2.1.3	qRT-PCR for viral co-infection	20
2.1.4	RNA-seq library generation, sequencing, and data processing	21
2.1.5	Taxonomic classification of RNA-seq reads for detection of bacterial and viral pathogens	22
2.1.6	Detecting beta-lactamase genes using RNA-seq	22
2.1.7	Viral genome assembly and phylogenetic analysis	23
2.1.8	Host response gene expression analysis	23
2.1.9	Statistical analysis	24
2.2	Results	24
2.2.1	Cohort characteristics	24
2.2.2	Bacterial pathogen detection by metatranscriptomic analysis of NP samples	26
2.2.3	Beta-lactamase gene detection in HFLU positive samples	28
2.2.4	Metatranscriptomic detection and analysis of respiratory viruses	29
2.2.5	RNA-seq uncovers additional pathogens and alternate explanations of disease etiology	32

2.2.6	Viral genome assembly and subtyping from host-derived metatranscriptomes	35
2.2.7	Host-response expression profiles distinguish bacterial from viral infections	37
2.2.8	Magnitude of host responses correlates with viral and bacterial pathogen abundance	40
2.2.9	RNA-seq classifies patients into distinct groups with unique pathogen-host response profiles	44
3	Discussion	46
	Conclusion	50
	Future Work	51
	References	52
	Appendix	66
	Supplementary Data	66

List of Figures

Figure 1. Overview of study design.	25
Figure 2. Metatranscriptomic detection of bacterial pathogens in NP samples from children with clinically diagnosed acute sinusitis.	27
Figure 3. Detected beta-lactamase genes by CARD in resistant versus non-resistant HFLU samples.	28
Figure 4. Detection of common respiratory viruses in NP metatranscriptomes.	31
Figure 5. Metatranscriptomics of NP samples from children with acute sinusitis identified organisms not detected by qRT-PCR or culture.	34
Figure 6. Assembled genomes of viruses from children with clinically diagnosed acute sinusitis.	36
Figure 7. Identification of differentially expressed host genes indicative of host-responses to bacterial and viral infection in acute sinusitis patients.	39
Figure 8. Host-response correlates with relative abundance of bacterial and viral pathogens.	42
Figure 9. Differential host response expression analysis based on patients' symptom severity score (PRSS) at time of sample collection.	44

List of Abbreviations

ADV	Human adenovirus.
AMR	Antimicrobial resistance.
ANI	Average nucleotide identity.
ARI	Acute respiratory infection.
AS	Acute sinusitis.
AUC	Area under the curve.
AUROC	Area under the receiver operator curve.
CARD	Comprehensive Antibiotic Research Database.
COV	Human coronavirus.
Ct	Cycle threshold.
DEG	Differentially expressed gene.
DNA	Deoxyribonucleic acid.
DNA-seq	Deoxyribonucleic acid sequencing.
dsRNA	Double stranded ribonucleic acid.
EV	Human enterovirus.
HFLU	<i>Haemophilus influenzae</i> .
HRV	Human rhinovirus.
INF	Influenza virus.
LRTIs	Lower respiratory tract infections.
MCAT	<i>Moraxella catarrhalis</i> .
MPV	Human metapneumovirus.

NCBI	National Center for Biotechnology Information.
NGS	Next-generation sequencing.
NP	Nasopharyngeal.
PCR	Polymerase chain reaction.
PIV	Human parainfluenzavirus.
PRSS	Patient symptom severity scores.
qRT-PCR	Real-time quantitative reverse transcription PCR.
RAM	Random Access Memory.
RGI	Resistance gene identifier.
RNA	Ribonucleic acid.
RNA-seq	Ribonucleic acid sequencing.
ROC	Receiver operator curve.
RPKM	Reads per kilobase million.
RPM	Reads per million.
RSV	Human respiratory syncytial virus.
SPN	<i>Streptococcus pneumoniae</i> .
upDEG	Upregulated differentially expressed gene.
URTI	Upper respiratory tract infection.

Chapter 1

Introduction and Literature Review

1.1 Respiratory Infections

Acute respiratory infections (ARIs) are some of the most frequent types of infections worldwide as well as the most common reason for seeking medical care such as visits to clinics or hospitals (Charlton et al., 2018). An ARI is defined as an infection causing one or more respiratory symptoms such as coughing, sore throat, and difficulty to breathe (Kang et al., 2016). These infections cause significant burden to health systems and economies due to their prevalence and ability to cause epidemics as well as pandemics as we have recently seen with COVID-19. (Zhang et al., 2020). They can also significantly burden their host as a result of their morbidity (symptoms often include fever, cough, sore throat, difficulty breathing, or more specific manifestations such as pneumonia and bronchiolitis) (Khomich et al., 2018) and their mortality rate (especially in young children and older adults) (Charlton et al., 2018). Infections can occur in the upper respiratory tract (upper respiratory tract infections, or URTIs) and/or the lower respiratory tract (lower respiratory tract infections, or LRTIs). URTIs occur in sites such as the nasal cavities, throat, sinuses, and inner ear; they typically manifest as a ‘common cold’, sinusitis, sore/dry throat, and eye or ear infections (Charlton et al., 2018). LRTIs typically occur in the trachea and lungs and can manifest as bronchitis or pneumonia (Charlton et al., 2018).

ARIs can be caused by pathogenic viruses, bacteria, or fungi (Dasaraju & Liu, 1996). Of these 3 pathogen groups, viruses cause the largest proportion of respiratory illness (80%) (Zhang et al., 2020) with Influenza virus, Respiratory Syncytial virus, Coronavirus, Adenovirus, Rhinovirus, Metapneumovirus, Enterovirus, and Parainfluenza virus being the most common viruses (Charlton et al., 2018). With the exception of Adenovirus which has a double stranded DNA genome, the above mentioned viruses have single stranded RNA genomes (Knipe & Howley, 2013). Bacterial and fungal respiratory pathogens include: *Haemophilus influenzae* (HFLU), *Moraxella catarrhalis* (MCAT), *Streptococcus pneumoniae* (SPN), *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, *Legionella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Bordetella pertussis*, *Aspergillus*, *Cryptococcus*, *Mucor*, *Candida*, and *Histoplasma capsulatum* (Zhu et al., 2020).

Young children, older adults, and the immunocompromised are most at risk of infection and have higher risk of complications and increased infection severity due to their immune systems not being fully developed or declining in functionality, respectively (Charlton et al., 2018). Respiratory illnesses are so frequent in younger children that by the age of 5, close to 100% of children have been infected at some point of their life with Rhinovirus and Respiratory Syncytial virus (Rajagopala et al., 2021), 80% have been infected with a parainfluenza viruses (Schomacker et al., 2012), and 90-100% have been infected with human Metapneumovirus (Panda et al., 2014).

1.1.1 Common Respiratory Viruses

Human Influenza virus (INF) is categorized into 3 types: A, B, and C, and infects up to 10% of the global population annually (Javanian et al., 2021). Influenza A and B cause mild to severe illness such as pneumonia, with influenza A being more prevalent and known to cause epidemic outbreaks while influenza B is responsible for localized outbreaks (Javanian et al., 2021). Influenza C typically causes illness with mild symptoms and can sometimes lead to localized outbreaks (Javanian et al., 2021). INF infections are most prevalent during the winter season.

Respiratory syncytial virus (RSV) is categorized into types A and B. Some studies have shown that type A generally causes more severe illness than type B (Borchers et al., 2013). RSV can manifest as mild URTIs but also causes a substantial proportion of severe LRTIs that lead to hospitalization such as bronchiolitis in infants (Borchers et al., 2013). Like INF, RSV infections are most common during winter.

Seven species of coronavirus (COV) are known to infect humans, of these, four are common and usually cause mild URTIs typically manifesting as a common cold: 229E, NL63, HKU1, and OC43. Co-infections of these mild coronavirus species with other coronaviruses or respiratory viruses is common, with some studies reporting close to 50% co-infection incidence. Recently, three new coronaviruses have emerged which can cause severe illness and have caused pandemics of large global burden: SARS-CoV-1, MERS-CoV, and SARS-CoV-2 (COVID-19). Although severe illness and death are rarely likely to manifest in children and instead occur mostly in adults, children still contract and readily transmit these viruses (Rajapakse & Dixit, 2021).

Adenovirus (ADV) respiratory infections are typically mild and self limiting and therefore rarely a cause for seeking medical care (Khanal et al., 2018). Adenoviruses are split into several species, A to G. They are not as prevalent as other viruses, causing an estimated 5-10% of respiratory illnesses in children (Knipe & Howley, 2013). However, localized outbreaks have occurred with more severe symptoms that can be a threat to young children and immunocompromised individuals (Khanal et al., 2018). Adenovirus B, C, and E can manifest as respiratory illness with types 4,7, and 14 being implicated in outbreaks (Knipe & Howley, 2013).

Human rhinoviruses (HRVs) are the most common cause of URTIs and more specifically, the common cold. They are split into three groups: A, B, and C. Unlike most respiratory viruses which have seasonal patterns and commonly cause outbreaks during winter months (such as Influenza and RSV), Rhinoviruses cause illness year-round and are responsible for most viral ARIs in the spring, summer, and fall. (Jacobs et al., 2013). While Rhinoviruses are largely responsible for mild infections such as the common cold and sinusitis, they have been implicated in more severe outcomes such as pneumonia and in asthma development and exacerbation (Jacobs et al., 2013).

Human metapneumoviruses (MPVs), which are classified into either group A or B, most commonly infect young children but can also infect the elderly and immunocompromised (Panda et al., 2014). MPV season peaks in the spring and can cause mild URTIs or serious LRTI infections leading to pneumonia (Panda et al., 2014).

Enteroviruses (EVs) are a large group of viruses causing a wide range of illness such as polio, meningitis, encephalitis, eye infections, flaccid paralysis, and respiratory illness (Knipe & Howley, 2013). EVs are grouped into four species, A to D. Most frequently the respiratory illnesses they cause are mild, self-limiting URTIs. Enterovirus D68 is of note as it has caused outbreaks of severe and fatal LRTI (Holm-Hansen et al., 2016).

Parainfluenza (PIV) types 1 to 4 usually cause URTI infections such as croup (swelling of the upper airways) but can also cause serious LRTIs especially in young children and older adults (Branche & Falsey, 2016). PIV accounts for up to 40% of hospitalizations of children due to LRTIs. PIV 3 is the most prevalent and associated mostly with LRTIs while the other 3 serotypes are associated mostly with URTIs (Branche & Falsey, 2016).

Generally, RSV, PIV, MPV, and INF are associated with occurrences of severe illness which requires hospitalization (Flynn et al., 2021). On the other hand, RHV, ADV, COR, and EV typically manifest as mild illness such as the common cold (Flynn et al., 2021). Along with virus type, co-infection of viral-viral or viral-bacterial pathogens is another factor that contributes to the severity of illness. Some studies using newer, more sensitive PCR (polymerase chain reaction) techniques have reported 40% or more co-infection rate of 2 or more viruses while older methods reported 10-30% (Scotta et al., 2016). There has been some contradiction in the literature about the association of viral-viral co-infections with the severity of illness (Scotta et al., 2016). A review of 43 papers that studied a total of 17,234 patients found that there was no significant association between the presence/absence of viral-viral co-infections and risk of outcomes such as severity, length of stay, and need for supplemental oxygen (Scotta et al., 2016). Conversely, numerous studies have reported that viral-bacterial co-infections increase the severity of respiratory illness. Viruses have been reported to increase the presence of pathogenic bacteria in the microbiome of infected patients via several proposed mechanisms (Bosch et al., 2013; Brealey et al., 2015; Jacobs et al., 2013). The incidence rate of viral-bacterial co-infections in

hospitalized children with ARI varies quite significantly between studies, with incidence rates from 2%-77% being reported (Brealey et al., 2015). However, the studies also varied widely in the viruses, bacteria, and type of ARI being examined. Secondary infection of *Streptococcus pneumoniae* with influenza virus is a well known example of bacterial infection increasing the mortality rate of viral illness such as the case of the highly deadly and contagious Spanish flu in 1918 (Bosch et al., 2013). The bacteria most commonly causing secondary infections with viral respiratory pathogens are *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Haemophilus influenzae* and (Brealey et al., 2015).

1.1.2 Sinusitis

Sinusitis is a condition where the sinus mucosal lining is inflamed, commonly due to a viral infection or an allergen; acute sinusitis is distinguished by symptoms lasting up to 4 weeks whereas symptoms of chronic sinusitis persist for at least 12 weeks (Brook, 2011, 2013; Quintanilla-Dieck & Lam, 2018). Acute sinusitis is one of the most common diagnoses in primary care settings in the U.S. with 31 million individuals being diagnosed annually (Anzai & Paladin, 2010). In many cases, sinusitis is caused by common respiratory viruses mentioned above and the patient recovers without the need for antibiotics or other therapy (Leung et al., 2020). However, the inflammation caused by viral sinusitis can lead to fluid build up and viscous secretions which obstruct the removal of bacteria from the nasal passages (Brook, 2011; DeMuri & Wald, 2012). Some viruses (such as HRV) have even been shown to promote the colonization of opportunistic pathogens such as *Staphylococcus aureus* and *Streptococcus pneumoniae* in the respiratory tract (Jacobs et al., 2013). These factors result in the increased presence of pathogens in the sinuses and subsequently provide increased opportunity for the development of bacterial sinusitis infection. As a result, in a subset (0.5-2% in adults and 5-10% in children) of viral sinusitis, secondary single or polymicrobial bacterial sinusitis infections may develop and cause complications (Brook, 2011; Leung et al., 2020). Bacterial sinusitis is commonly marked by pain in the sinus area, nasal congestion, headaches, fever, and thick, often yellow, nasal discharge (Brook, 2013). However, these symptoms are often shared with viral sinusitis, and children

often exhibit less specific symptoms (Leung et al., 2020). *Haemophilus influenzae*, *Streptococcus pneumoniae*, *Moraxella catarrhalis*, *Staphylococcus aureus*, *Streptococcus pyogenes*, group C *Streptococcus*, and some anaerobe species are possible pathogens causing bacterial sinusitis, with the first 3 being the most common (Leung et al., 2020).

An estimated 10% of children have had bacterial sinusitis at some point by the age of 3, but bacterial sinusitis is commonly overlooked as a secondary infection to viral URTIs or confused with viral URTIs due to symptom similarity. This can result in chronic bacterial sinusitis and other serious complications (Leung et al., 2020). This has led to the overuse of antibiotic prescriptions to treat sinusitis (Hersh et al., 2013). Despite secondary bacterial sinusitis infections typically only occurring in up to 10% of pediatric viral URTIs and even less commonly in adult URTIs, as many as 50-60% of sinusitis cases are diagnosed as bacterial and subsequently prescribed antibiotics. In the U.S. in 1992, 2.2 billion dollars were spent in medications to treat sinusitis indicating a large economic and health impact (Anzai & Paladin, 2010). The prevalence of sinusitis only continues to increase due to pollution and antimicrobial resistance, highlighting the need for more accurate diagnosis to limit the use of antibiotics (Anzai & Paladin, 2010).

1.2 Current Pathogen Detection Methods

Oftentimes, clinicians are not able to distinguish bacterial from viral infections in pediatric acute sinusitis using clinical presentation alone due to the overlap in symptoms (Charlton et al., 2018). Molecular or culture-based testing is typically done to identify the infectious agent(s) and provide a diagnosis. Culture is used to test for the presence of bacteria while qRT-PCR, a molecular technique, is typically used to test for viruses. Culture and qRT-PCR are both targeted techniques which require prior knowledge of possible infectious agents and therefore typically do not detect rare and novel pathogens. Although PCR tests have high sensitivity especially when compared to antigen or culture testing (for example, a multiplex qRT-PCR test demonstrated 96.9% sensitivity overall for 10 viruses (Choudhary et al., 2016)), PCR tests are a targeted approach that can only test a limited number of pathogens are subject to false negatives due to viral mutation (Nakamura et al., 2009).

In fact, even with the use of PCR testing, a specific pathogen is unable to be detected in 20-60% of ARIs (Graf et al., 2016). Additionally, this current routine testing done by microbiological laboratories takes approximately three days (Sinha et al., 2018), which is too slow to affect clinical intervention and results in children being prescribed antibiotics by default during the first visit to the clinic.

1.3 Metatranscriptomics for Pathogen Detection

Current clinical methods do not differentiate between bacterial and viral sinusitis reliably, necessitating the development of an accurate and precise diagnostic method (Brook, 2011; Knipe & Howley, 2013; Tada & Hanada, 2010). Next generation sequencing (NGS) is a high-throughput sequencing approach which uses the concept of massively parallel processing; it has become cheaper, faster, and more accurate, which has led to recent research in using it as a diagnostic method (Chiu & Miller, 2019). Numerous studies have already investigated pathogen detection by NGS in various types of illness such as pneumonia (Xie et al., 2021), bone and joint infections (Ramchandar et al., 2021), COVID-19 (Tan et al., 2020), meningitis and encephalitis (Piantadosi et al., 2021), and gastroenteritis (Mohammad et al., 2020).

Metatranscriptomics, in this context, is the use of NGS to sequence of all the RNA from a patient sample (a random ‘shotgun’ approach). RNA sequencing is favorable over DNA sequencing for several reasons: in addition to being able to detect pathogens and having the potential to identify novel pathogens, metatranscriptomics can provide other information such as profiling the microbiome, pathogen characterization, detecting antimicrobial resistance genes, and profiling the host response to elucidate infection type.

1.3.1 Overview of Methods

Sampling

First, an appropriate specimen type must be chosen according to the illness being investigated. For respiratory illnesses and particularly URTIs, nasopharyngeal aspirates, nasopharyngeal (NP) swabs, or nasal swabs are commonly used (Charlton et al., 2018). The choice of sample type matters as it affects ease of collection, patient comfort, and sensitivity of detection. Aspirates are collected by inserting a catheter deeply through the nostril and requires a suction device. NP swabs are more simply collected by inserting a swab deep into the nostril. Nasal swabs require only shallow insertion of a swab into the nostrils. Nasopharyngeal swabs are easier to collect than aspirates, do not have the risk of aerosol production associated with aspirates, have similar sensitivity to aspirates, and higher sensitivity than nasal swabs (Charlton et al., 2018; Flynn et al., 2021). Therefore, it is of interest to investigate the diagnostic potential of using metatranscriptomics to sequence nasopharyngeal swabs of patients experiencing ARI.

RNA Sequencing

Initial RNA studies employed low throughput techniques such as PCR and sanger sequencing. Currently, NGS machines producing short reads (35-300 base pairs), or micro-arrays, are the most commonly used techniques (Dulanto Chiang & Dekker, 2020). NGS has advantages over micro-arrays that make it better suited for metatranscriptomic analysis such as not requiring prior knowledge of the sequences in the sample, and better handling cases of very low or very high sequence expression (Kukurba & Montgomery, 2015). In order to sequence RNA, the RNA is isolated from a sample and converted to DNA, which is then used to construct a library, followed by PCR amplification, and finally, sequenced using a sequencing machine (Kukurba & Montgomery, 2015). Depending on the sample type and study, additional steps can be taken to deplete or enhance detection of certain transcripts, such as depleting ribosomal RNA (Kukurba & Montgomery, 2015). After sequencing and before subsequent analysis of reads, quality control is performed to remove

low quality/complexity reads, contaminants, and to remove human reads if appropriate (de Vries, Brown, Couto, et al., 2021).

Taxonomic Classification

Classifiers can match sequences based on nucleotide similarity, amino acid similarity, or both (de Vries, Brown, Fischer, et al., 2021). BLAST is commonly used to classify or cluster a sequence to its closest match in the database and has good accuracy but high computational cost (de Vries, Brown, Couto, et al., 2021; S. H. Ye et al., 2019). However, RNA sequencing of clinical samples typically yields millions of reads per sample, making tools such as BLAST impractical. Therefore, software which use approaches designed to classify sequences quickly and efficiently have been developed. The most common algorithm approach is using k-mers, which are short nucleotide sequences, to search for perfect matches within a database (de Vries, Brown, Couto, et al., 2021; S. H. Ye et al., 2019). Tools which employ this algorithm include Kraken2 (Wood et al., 2019), Centrifuge (Kim et al., 2016), DIAMOND (Buchfink et al., 2015), Kaiju (Menzel et al., 2016), MetaPhlan (Truong et al., 2015), and more. When choosing which of these tools to use, variables such as speed, sensitivity vs specificity, database available, and type of sample being investigated are considered. Several papers have benchmarked and compared these different tools and concluded that the tools had their respective strengths and weaknesses, but Kraken and its derivatives performed the best overall in terms of speed, accuracy, and database availability and customization (de Vries, Brown, Fischer, et al., 2021; Vollmers et al., 2017; S. H. Ye et al., 2019).

Pathogen Profiling

Classification levels reported by taxonomy classification tools are largely dependent on the database used. A larger database is needed to classify organisms on the serotype/strain/isolate level vs the species level, which increases the computing resources and computing time required (de Vries, Brown, Fischer, et al., 2021). A study found that using the NCBI nucleotide database with various classification tools provided the best classification beyond

species level (de Vries, Brown, Fischer, et al., 2021), however, this results in requirements of >100GB RAM in order to run. Another method entails de novo assembling sequences of interest (which creates longer stretches of sequences and allows for more accurate classification) followed by using BLAST or another similar tool to compare the sequences against the non-redundant sequence database (Rajagopala et al., 2021). The top hit with high sequence identity can be used to infer the serotype or strain of the species of interest. The added benefit of assembling prior to further classification is the assemblies can be used to create phylogeny trees or calculate metrics such as average nucleotide identity and genome coverage, and longer sequences can be classified more accurately compared to shorter sequences.

AMR detection

Detecting the presence of antimicrobial resistance (AMR) genes typically uses one of two methods: de novo assembly of reads into contigs followed by searching contigs against a reference database of resistance genes, or, mapping reads directly onto genes from the gene reference database. The read-based approach is faster but does not allow for detection of resistance genes not included in the database and does not allow for corrections to false positives using genomic context. On the other hand, while the assembly approach is slower and more expensive computationally, it allows for the detection of resistance genes which are novel or have low similarity to genes in the database (Boolchandani et al., 2019; “Omics of antimicrobials and antimicrobial resistance”, 2019; Waskito et al., 2022). There are several databases available such as CARD (Comprehensive Antibiotic Research Database) (Alcock et al., 2023), Resfinder (Florensa et al., 2022), Resfams (Gibson et al., 2015), and SARG(Structured Antibiotic Resistance Genes) (Yin et al., 2018). Some of these databases, such as CARD, also come with bioinformatics tools to use with the database for the identification of AMR genes.

Host Response Profiling

The first step in host response profiling to detect differentially expressed genes (DEGs) involves quantification of human transcript abundances. Tools such as Salmon (Patro et al., 2017) or Kallisto (Caldeweyher, 2021) can be used to quantify transcripts; these tools accomplish this by mapping reads onto a reference human genome. R packages such as edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014) are used to identify DEGs between two or more experimental groups (such as healthy vs infected, or viral vs bacterial infection). The identified DEGs can then be used as input for further analysis such as gene set enrichment analysis for identifying pathways or functions that are upregulated or downregulated. EnrichR (E. Y. Chen et al., 2013) is a popular and powerful tool for gene set enrichment analysis.

1.3.2 Previous work

Taxonomic Classification

A strength of metatranscriptomic analysis is the ability to profile in an unbiased manner the species present in a sample. For clinical diagnosis, detecting the pathogen(s) causing infection is important, but profiling commensal organisms may also contribute insights to disease type or progression which can aid with selecting appropriate treatment (Chiu & Miller, 2019; de Vries, Brown, Couto, et al., 2021; Xie et al., 2021). Current studies focus on taxonomic classification to identify causative pathogens, but some studies have also looked for patterns in the microbiome. Several of these studies also compared metatranscriptomics to traditional clinical methods such as PCR to compare accuracy and sensitivity of detecting common respiratory pathogens.

Rajagopala et. al. (Rajagopala et al., 2021) analyzed nasal swab samples from 58 children (healthy $n = 19$, respiratory syncytial virus (RSV) positive $n = 39$) to capture their virome and microbiome. In addition to RSV, they identified nine other human viruses such as coronavirus, human rhinovirus, and bocavirus in both healthy and RSV positive patients. Non-human viruses identified included multiple plant viruses, prophages, and

bacteriophages. They compared their metatranscriptomic virus detections to a multiplex panel and found that metatranscriptomics was more sensitive for the detection of RNA viruses. Microbiome analysis revealed the detection of 88 bacteria and 3 fungal species with high confidence; the topmost abundance species were *Moraxella catarrhalis*, *Streptococcus pneumoniae*, *Streptococcus mitis*, *Haemophilus influenzae*, and *Cutibacterium acne*. Their study demonstrated the feasibility of capturing the RNA virome and microbiome from a low biomass respiratory sample and reported on the presence and frequency of potential pathogens in both healthy and infected patients, as well as the occurrence of co-infections.

Toma et. al. (Toma et al., 2022) did unbiased metatranscriptomic analysis of oropharyngeal swabs from patients (n = 6) before, during, and after acute infection to study the throat microbiome in relation to respiratory illness. Microbiome analysis revealed an average species richness of 479 with *Streptococcus salivarius*, *Veillonella atypica*, *Prevotella shahii*, and *Streptococcus mitis* being included in the topmost abundant microbial species. Five out of six patients had a significant change in abundance of a pathogen during sickness compared to before or after. The respiratory pathogens identified as having changed in abundance included human coronavirus, *Moraxella catarrhalis*, *Klebsiella pneumoniae*, rhinovirus A, *Streptococcus pneumoniae*, and *Haemophilus influenzae*. The pathogens identified and the types of respiratory illness those pathogens commonly cause corresponded with the patients' symptoms. They also studied the change in the abundance of non-pathogenic taxa and found that three patients had significant shifts between healthy and sickness timepoints in microbial abundance of certain species. They showed that metatranscriptomic analysis of the throat microbiome could be used to give high resolution detection of pathogenic organisms that cause respiratory illness, as well as other species which shift in abundance in relation to sickness and can help shed light on disease mechanisms or progression.

Graf et. al. (Graf et al., 2016) compared detection of viruses in 42 patients positive for one or more virus using a commercial multiplex panel to RNA sequencing of nasopharyngeal swabs. RNA-seq had 95% agreement with the panel, with the remaining 5% discrepancy being due to two cases of rhinovirus that had high cycle threshold values from PCR (35 and 33 Ct). For a different subset of 65 samples, which were selected at random, RNA-seq

had 92% agreement. However, PCR testing confirmed RNA-seq findings and show that the panel had several false positive results. In addition, RNA-seq detected 12 additional viruses (some of which were targeted by the panel, and some not) which the panel failed to detect.

With SARS-CoV-2 being an important area of research recently, several papers have been published which focus on the use of RNA-seq to detect the virus and/or to further profile and characterize the virus in patients with COVID-19. Lu et al. (Lu et al., 2021) sequenced throat swabs from confirmed or suspected COVID-19 patients and analyzed RNA-seq data to detect SARS-CoV-2 as well as other putative pathogens. They identified several bacteria that were opportunistic pathogens that may have been causing co-infections. Butler et al. (Butler et al., 2021) tested 669 clinical samples from patients confirmed or suspected to have COVID-19. They used shotgun RNA-sequencing to detect SARS-CoV-2 and to profile the microbiome. They identified 17 species that were depleted in the microbiome of COVID-19 patients. In addition, they identified other human respiratory pathogens such as coronaviruses, influenza A, mastadenovirus, rhinovirus, and metapneumovirus. When the detection of these viruses was compared to a standard PCR panel, metatranscriptomics had an accuracy of 99.4% and sensitivity of 75.8%.

These studies demonstrate the sensitivity of metatranscriptomics in detecting respiratory RNA viruses and its ability to capture both virome and microbiome data from clinical samples. The studies report that metatranscriptomics reveals shifts in microbial abundance during sickness and highlights the potential of metatranscriptomics in high-resolution pathogen detection and understanding disease mechanisms. Comparison with traditional methods like PCR shows high accuracy and sensitivity, with some discrepancies attributed to lack of sensitivity due to insufficient sequencing depth. Other less relevant studies have looked at chronic or lower respiratory infections, and/or have used metagenomics instead of metatranscriptomics for their analysis (Castañeda-Mogollón et al., 2021; Mostafa et al., 2020; Schlaberg et al., 2017; Wang et al., 2016; Xie et al., 2021; Xu et al., 2017).

Pathogen Profiling

Metatranscriptomics has the ability to characterize a species at a much deeper level than typical clinical tests such as PCR. Full or partial genomes can be assembled and then utilized for further analysis such as detecting mutations, predicting phenotype, identifying strain/isolate, and tracking the spread or origin of infections (Byron et al., 2016; Chiu & Miller, 2019; Mulcahy-O’Grady & Workentine, 2016). This kind of pathogen profiling is necessary for implementing more effective measures to control the spread of infection and to better inform decisions made on public health matters.

Rajagopala et. al. (Rajagopala et al., 2021) assembled a total of 79 viral genomes from their 58 samples (minimum 90% coverage), many of which were complete genomes. The viruses assembled included mostly RSV-A and RSV-B, as well as other RNA viruses such as human coronavirus, influenza virus, and human rhinovirus. When comparing metatranscriptomics to their Luminex pathogen panel, they found that metatranscriptomics was more sensitive since it was able to distinguish between human rhinovirus and human enterovirus strains whereas the Luminex panel could not. Interestingly, they employed phylogenetic analysis of assembled RSV genomes to rule out the possibility of internal cross contamination during sequencing protocol.

Graf et. al. (Graf et al., 2016) did sequence-based strain typing by blasting the largest contig in the assembly. They typed influenza A, rhinovirus, respiratory syncytial virus, and coronavirus, and compared subtyping results between their clinical panel and metatranscriptomic analysis where applicable. Perfect agreement between the two methods was achieved for influenza A subtyping. They analyzed their rhinovirus assemblies with complete coverage to assess genetic diversity and placed them in a phylogenetic tree to assess similarity and lineages between the assemblies compared to reference rhinovirus genomes.

Butler et. al. (Butler et al., 2021) assembled 155 full length SARS-CoV-2 assemblies with sufficient information to achieve 10x coverage or higher. They examined these assemblies and found 165 unique variants which included single nucleotide variants and deletions. Phylogenetic analysis revealed that a high proportion of their assemblies were associated with clade 20C which likely originated in western Europe. They performed additional

analysis including variant calling to assess intrahost diversification.

Lie et. al. (Li et al., 2020) identified genotypes/subtypes as well as epidemiological origins of the viruses detected in multiple patient samples. This included both Yamagata and Victoria lineages of influenza B, and A and B genotypes of human metapneumovirus. Phylogenetic analysis revealed that the sequences formed several closely related clusters, suggesting they may originate from the same outbreak. Additionally, they discovered that an echovirus 6 detected in the dataset had less than 90% nucleotide identity to known viruses, and likely represents a new variant of echovirus 6.

In summary, researchers successfully utilized metatranscriptomics to assemble numerous viral genomes, including those of RSV, human coronavirus, influenza virus, and rhinovirus, with high coverage. Additionally, it enabled the identification of variants, assessment of intrahost diversification, tracing of epidemiological origins, and even identified potentially novel virus variants, which are all essential for understanding viral diversity, lineages, and origins.

AMR Detection

Metatranscriptomics offers several advantages when used for the detection of AMR genes. Typical clinical procedure uses bacterial isolates to test for the presence of targeted resistance genes such as beta lactamases, which does not offer a broader view of the presence of resistance genes in the microbial community as a whole (the resistome) and which may be relevant for treatment choice (Dulanto Chiang & Dekker, 2020; Mulcahy-O'Grady & Workentine, 2016). Additionally, unlike genomic approaches, metatranscriptomics allows for the capture of AMR genes which are being actively expressed.

Analysis in a study by Li et. al. (Li et al., 2020) revealed the presence of AMR genes from a total of 8 classes of antibiotics across their samples. This included aminoglycosides, beta-lactamases, and fluoroquinolones, to name a few. They also reported a significant difference in abundance and diversity of AMR genes between pediatric patients with respiratory infection and healthy controls, with healthy controls having lower abundance and diversity.

Graf et. al. (Graf et al., 2016) examined a specific mutation site in their influenza A assemblies that typically confers oseltamivir resistance, and determined that none of the assemblies possessed the mutation. However, they were not able to analyze 2 of 8 influenza assemblies due to coverage being insufficient.

While the study by Lu et.al. (Lu et al., 2021) focused on patients with COVID-19, they profiled antimicrobial resistance of the microbiome using CARD and report findings that are relevant to other types of respiratory infections. They detected AMR genes belonging to 28 classes of antibiotics overall, with healthy patients having significantly fewer AMR gene transcripts detected than patients positive for SARS-CoV-2. The most abundance genes belonged to classes including beta-lactams, aminoglycosides, and tetracyclines classes.

The collective findings of these papers highlight the ability of metatranscriptomics in detecting AMR genes across various antibiotic classes, although coverage limitations may pose challenges. Metatranscriptomics offers a broader view of the resistome compared to traditional methods, as it captures actively expressed AMR genes from the entire microbial community. Studies demonstrate significant differences in the abundance and diversity of AMR genes between infected and healthy individuals, indicating its potential in understanding resistance dynamics in respiratory infections.

Host Response Profiling

Profiling host transcripts provides a promising avenue for directing the use of antibiotics. By analyzing the host response through genes being actively expressed, it is possible to detect whether the host is fighting an illness of viral, bacterial, or non-infectious origin (Ko et al., 2015; Ross et al., 2019; Troy & Bosco, 2016). This can then guide the selection of an appropriate treatment as to avoid the use of antibiotics when unnecessary. Recent publications have reported on functional pathway patterns and gene signatures derived from the analysis of samples from patients infected with respiratory infections that allow for distinguishing between infection types.

Rajagopala et. al. (Rajagopala et al., 2021) compared the nasal mucosal cell transcriptome between healthy patients and patients infected with respiratory syncytial virus.

They report 2,878 upregulated and 1,746 genes downregulated in infected patients. Notably, upregulated genes included interferon response genes and chemokines. Pathways enrichment analysis revealed upregulation of 69 pathways related to anti-viral response such as interferon signaling, chemokine signaling, and the inflammasome pathway.

Several studies have looked at analyzing host response to differentiate between SARS-CoV-2 and other viruses or healthy cases. Lu et.al. (Lu et al., 2021) analyzed differences and similarities in host response between patients that were positive vs negative for SARS-CoV-2, and between patients with low vs high viral load. They found that they could clearly differentiate between these different groups. They report genes which are enriched during infection, as well as genes which are enriched in either low viral load or high viral load states. Pathway enrichment analysis of these genes provided insights into how the host response changes in relation to viral load. Butler et. al. (Butler et al., 2021) reported 757 differentially expressed genes between samples that were positive or negative for SARS-CoV-2 (350 upregulated, 407 downregulated). The upregulated genes corresponded to a range of antiviral response pathways. Interestingly, they identified a common interferon response that was significantly higher when compared with samples that were negative for SARS-CoV-2 but positive for other respiratory viruses, suggesting a host response unique to COVID-19 infection. Pathways that were downregulated included an olfactory receptor pathway, and pathways involved with lung cell growth and regulation of heme regulation; this is consistent with phenotypes observed during infection.

Applications of analyzing host response are currently centralized on the development of predictive models to differentiate between bacterial, viral, and healthy states. Landry et. al. (Landry & Foxman, 2018) sequenced and analyzed nasopharyngeal swabs from patients with suspected viral respiratory infection to analyze the host transcriptome and create a signature of 3 mRNA biomarkers to predict viral infection. They found that the levels of these 3 transcripts highly correlated with viral infection; they reported 97% accuracy with 100% positive predictive value. All samples which were negative for viruses was also negative for the biomarker test. Conversely, RNA-seq analysis performed by Bhattacharya et. al. (Bhattacharya et al., 2017) was used to select eleven genes as markers to detect bacterial infection. The classifier achieved 90% sensitivity and 83% specificity. While the

study focused on patients with lower respiratory tract infections and used whole blood for sequencing and profiling, their results demonstrate a proof of concept that can then be tested on different types of infections or clinical samples.

The papers reviewed in this section demonstrate that metatranscriptomics can be used effectively for the analysis of host response, aiding in the differentiation of viral, bacterial, and healthy states. Studies reveal distinct gene expression patterns in response to respiratory infections, with upregulated genes related to antiviral or antibacterial pathways. Researchers developed and tested predictive models using host transcriptome signatures, achieving high accuracy in discriminating infections. These models hold promise for improving diagnostic accuracy and informing clinical decisions for respiratory infections.

1.4 Hypothesis and Objectives

This thesis explores the hypothesis that metatranscriptomics can be used to accurately diagnose and characterize pathogens causing acute pediatric sinusitis from nasopharyngeal swabs. Key objectives include:

- Taxonomic classification to detect bacterial pathogens causing acute sinusitis, followed by comparing these detection to those of traditional clinical culture plate testing done in parallel. Detection of beta-lactamase transcripts in samples positive for HFLU will also be analyzed and compared.
- Taxonomic classification to detect viral pathogens causing acute sinusitis, followed by comparing these detection to those of traditional clinical qRT-PCR testing done in parallel. This will be followed by characterizing the detected viruses: calculating viral load, assembly of viral genomes, and assessing sequence similarity to known sequences.
- Analysis of host transcripts to uncover patterns in host response that might distinguish bacterial from viral acute sinusitis infections or predict other clinically relevant features.

Chapter 2

Metatranscriptomic Analysis and Results

Material in this chapter has been prepared for publication and is available as a pre-print in *bioRxiv* accessible at the following DOI: [10.1101/2024.03.03.24303663](https://doi.org/10.1101/2024.03.03.24303663)

2.1 Methods

2.1.1 Study design and description of the cohort

Sample collection was performed by collaborators (Dr. Shaikh Nader and his colleagues) from Pittsburgh University. Between February 2016 and April 2022, 510 children 2 to 11 years of age (inclusive) with clinically diagnosed acute sinusitis were enrolled in a randomized multicenter double-blind trial (ClinicalTrials.gov number, NCT02554383). Exclusion criteria have been previously described (Shaikh et al., [2023](#)). Children were recruited from 6 outpatient centers. A total of 204 patients did not have a NP sample collected, or their sample was not preserved in RNA buffer and were excluded. Of the remaining 306 patients' samples, 61 were not sequenced due to low RNA yield. Although 245 samples underwent RNA-sequencing, batch 1 was prepared with a different kit/protocol and when analyzed

displayed a strong batch effect and was thus removed, leaving 221 patients. Children were randomly assigned to receive 10 days of amoxicillin-clavulanate or matching placebo. The primary outcome, symptom burden, was assessed by having parents complete the Pediatric Rhinosinusitis Symptom Scale (PRSS) electronically every evening on Days 2 to 11. As previously described (Shaikh et al., 2023) the PRSS is a validated scale that assesses symptoms of sinusitis.

2.1.2 Culture and sensitivity pattern of bacterial pathogens

This was performed by collaborators (Dr. Shaikh Nader and his colleagues) from Pittsburgh University. They collected NP swabs from all children at study entry. As previously described (Lopez et al., 2019), the tip of the swab was cut, placed in DNA/RNA shield (Zymo, R1100), and transported on ice to the lab. The remainder of the swab was placed into Amies transport medium and transported on ice to the Clinical Laboratory at UPMC Children’s Hospital of Pittsburgh within 48 hours and plated on blood and chocolate agars. Identification of SPN, HFLU, and MCAT on culture was accomplished using standard microbiological techniques. HFLU isolates were tested for the beta-lactamase production using a cefinase disk.

2.1.3 qRT-PCR for viral co-infection

PCR was performed by collaborators (Dr. Shaikh Nader and his colleagues) from Pittsburgh University. Using an aliquot of Amies transport media plus MagMax lysis/binding buffer, nucleic acid extraction was performed for viral identification using the ABI MagMax96 Express automated instrument and the MagMax 96 Viral Isolation Kit (Thermo Fisher, AMB 18365) (Lopez et al., 2019). Adenovirus, influenza subtypes A/B/C, human metapneumovirus (MPV), human rhinovirus (HRV), parainfluenza virus (PIV) subtypes 1-4, Enterovirus D68, and respiratory syncytial virus (RSV) were tested for using individual real-time qRT-PCR assays. A Ct threshold of 40 was used for all viruses and positive and negative controls were included in each run.

2.1.4 RNA-seq library generation, sequencing, and data processing

Library generation and sequencing was performed by collaborators (Dr. Shaikh Nader and his colleagues) from Pittsburgh University, and data processing was performed by the author of this thesis. RNA was assessed for quality using a Fragment Analyzer 5300 and RNA concentration was quantified on a Qubit FLEX fluorometer. Libraries were generated with either the Illumina TruSeq Stranded Total RNA prep (20020599) or the Illumina Stranded Total Library Prep kit (Illumina: 20040529) according to the manufacturer's instructions, after using the Illumina Ribo Zero Plus rRNA Depletion Kit (20037135). Batch 5 was additionally treated with the Illumina Ribo-Zero Plus Microbiome rRNA Depletion Kit (20072062). For library generation, 100 ng of input was used for the Illumina TruSeq Stranded Total RNA protocol with 15 cycles of indexing PCR, and 20-100 ng of RNA input was used for the Illumina Stranded Total Library Prep protocol with 15 cycles of indexing PCR for 100ng of RNA input and 17 cycles of indexing PCR for input RNA >100ng. Library quantification and assessment was done using a Qubit FLEX fluorometer and the Fragment Analyzer 5300. Libraries were normalized and pooled to 2 nM by calculating the concentration based off the fragment size (base pairs) and the concentration (ng/l) of the libraries. Sequencing was performed on an Illumina NextSeq 2000, using a P3 200 flow cell with sequencing read lengths of 2x101bp, with a target of 40 million reads per sample. Sequencing data was demultiplexed by the Illumina on-board DRAGEN FASTQ Generation software. Library generation and sequencing was performed by the University of Pittsburgh Health Sciences Sequencing Core (HSSC), Rangos Research Center, UPMC Children's Hospital of Pittsburgh, Pittsburgh, Pennsylvania, United States of America.

Fastp v0.23.1 (S. Chen et al., 2018) was used for quality trimming and adapter removal on default parameters. FastQC v0.11.9 (Andrews, 2010) and MultiQC v1.12 (Ewels et al., 2016) were used to check the quality of all sequence files before and after processing to ensure data was ready for analysis.

2.1.5 Taxonomic classification of RNA-seq reads for detection of bacterial and viral pathogens

Taxonomic classification of sequencing reads was performed using Kraken 2 v2.1.2 (Wood et al., 2019) with default parameters. The PlusPF database dated 9/8/2022 (<https://benlangmead.github.io/awsindexes/k2>) was used with Kraken 2, which was originally built from NCBI RefSeq archaeal, bacterial, viral, plasmid, human, UniVec_Core, protozoan, and fungal sequences. A Kraken 2 detection threshold of 3 reads was used for bacterial species (selected based on F1 score optimization), while no threshold was used for viruses. New pathogens identified by Kraken 2 but not included in the clinical panel were further validated using BLAST (J. Ye et al., 2006), MASH (Ondov et al., 2016) and metAnnotate (Petrenko et al., 2015), focusing on samples associated with the largest estimated abundance for each pathogen.

The normalized abundance of each taxon was calculated as the number of reads per million (RPMs). Relative abundance heatmaps were generated using R v4.2.1 and the pheatmap package. For display, $\log_{10}(\text{RPM} + 1)$ values were used to avoid $\log(0)$ errors. Receiver operator curves were also generated in R and the area under the curve was computed using the pROC package. Pathogen abundance jitter plots and top species plots were generated using ggplot2 in R (Wickham, 2011).

Viral load was estimated from RNA-seq data following the method of Graf et al (Graf et al., 2016). The number of detected reads for a virus was divided by the total number of reads in the sample and the size of the respective viral genome in kilobases, and then multiplied by 1 million to generate an RPKM value (reads per kilobase of reference sequence per million total sequencing reads).

2.1.6 Detecting beta-lactamase genes using RNA-seq

For the samples that were positive for *H. influenzae* based on culture tests, sequencing reads classified as non-human by Kraken 2 were extracted using `extract_kraken_reads.py` and assembled into contigs using the rnaSPAdes v3.15.4 with default parameters (Bankevich

et al., 2012). Using CARD resistance gene identifier (RGI) software v6.0.1 (Alcock et al., 2023) and default database, the contigs were analyzed with the ‘main’ function of the RGI tool with the ‘low-quality’ and ‘include-nudge’ parameters. The results were filtered to keep “strict” or “perfect” hits to beta-lactamase genes, genes acting on antibiotics belonging to the penam drug class, and hits with at least 10.0% sequence coverage to the reference gene.

2.1.7 Viral genome assembly and phylogenetic analysis

RefSeq genomes for all viruses of interest were downloaded from NCBI. Non-human reads were mapped to viral genomes using BMap v38.86 (Bushnell, 2014) to create .bam files. The consensus sequence for each sorted mapping result was produced using samtools v1.16.1 with the ‘-a’ option. A python script was used to calculate whole genome coverage relative to the RefSeq viral genome. Genome coverage was considered complete if at least 99.5%. FastANI v1.32 was used to calculate the average nucleotide identity to the closest reference genome for each genome assembled.

Complete viral genomes were queried against the complete NCBI non-redundant nucleotide database using BLAST (J. Ye et al., 2006). Up to 35 top matching sequences were downloaded and aligned to the assembled genome using the MUSCLE algorithm (Edgar, 2004). The multiple genome alignment was used to generate a phylogenetic tree with FastTree v2.1.10 (Price et al., 2010), and FigTree v1.4.4 was used for tree visualization.

2.1.8 Host response gene expression analysis

Host transcript abundance quantification was performed using Salmon v1.7.0 (Patro et al., 2017) with the Human Gencode v39 reference transcriptome. Differential gene expression analysis was performed using DESeq2 and tximport in R (Love et al., 2014). Related statistical analyses are described in the following section. Heatmaps were produced in R using pheatmap, v1.0.12 jitter plots using ggplot2 v3.3.6, and volcano plots using the EnhancedVolcano package v1.14.0.

2.1.9 Statistical analysis

Differentially expressed genes (DEGs) were detected by comparing samples positive for viruses only versus samples positive for bacteria only based on culture or qRT-PCR testing. In the design formula for the ‘DESeqDataSetFromTximport’ function, potential confounding variables was controlled for including “batch number”, “sex”, and “age (scaled)”. Log2 fold changes and adjusted p-values (q-values) were calculated for all genes, and a significance threshold of $q \leq 0.05$ was used to identify DEGs. Function enrichment analysis of genes with significantly increased expression in the viral and bacterial groups was performed using EnrichR (accessed June, 2023) (E. Y. Chen et al., 2013) with the GO Biological Process 2021 ontology and an FDR threshold of 0.05.

2.2 Results

Nasopharyngeal swabs collected from 221 pediatric sinusitis patients underwent RNA sequencing, qRT-PCR, and culturing in parallel. The resulting data was analyzed for this thesis. Analysis included pathogen detection and quantification, assembly of respiratory viruses, detection of beta-lactamase genes, and differential analysis of host response. Study overview is depicted in Figure 1.

2.2.1 Cohort characteristics

A subset of 221 pediatric patients presenting with symptoms of acute sinusitis from a previous study (Shaikh et al., 2023) (Feb 2016 to Mar 2022) were selected for NP RNA-seq (Figure 1, Table 1). Further details are provided in the Methods and in Shaikh et al. (Shaikh et al., 2023). One naris was sampled using a NP swab and this was used for viral qRT-PCR, bacterial culture, and RNA-sequencing (Lopez et al., 2019); 171 (77 %) and 169 (76%) of the children tested positive for at least one bacteria or virus, respectively. Parents assessed symptom severity daily during the 10 days following diagnosis.

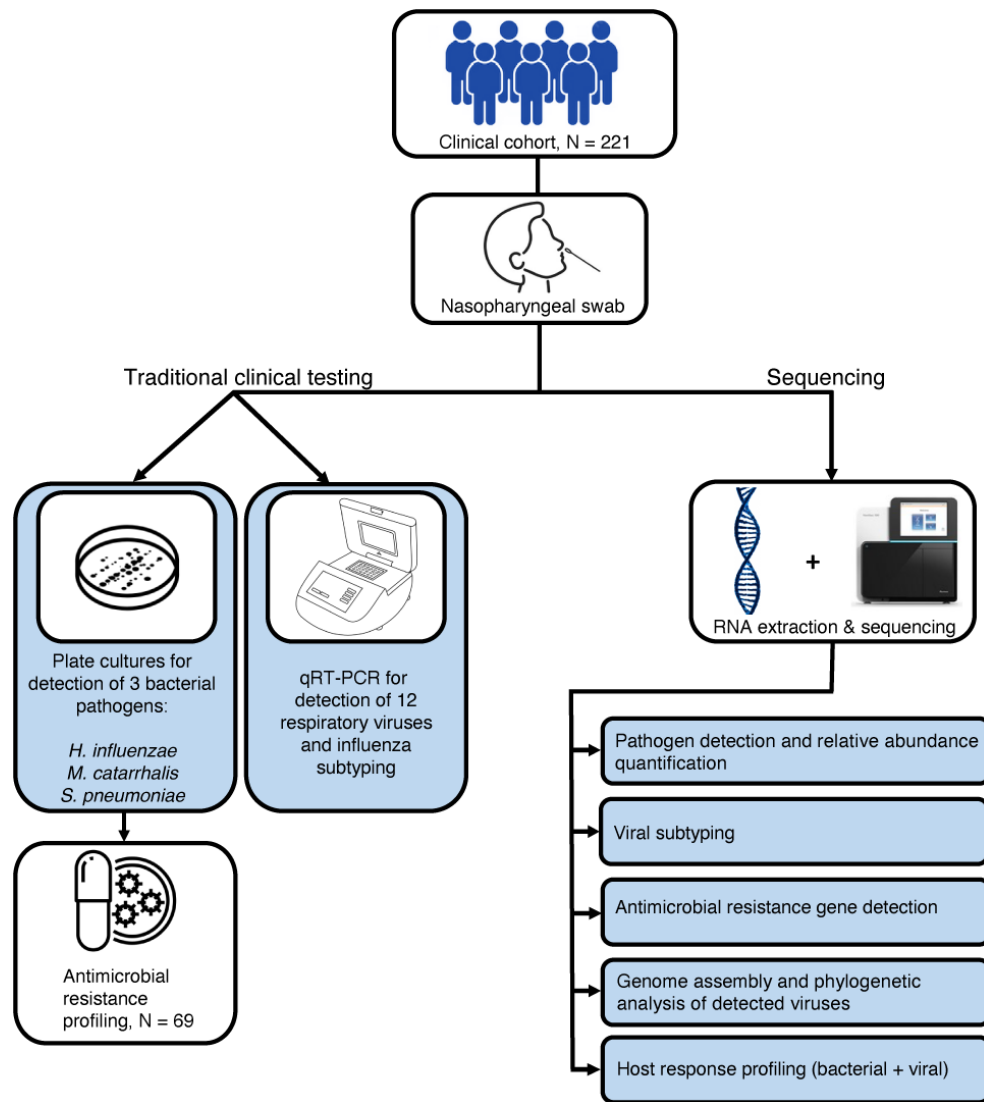


Figure 1. Overview of study design. The study cohort was comprised of 221 children with acute sinusitis who underwent collection of NP swabs. Culture was used to detect three bacterial species (*Haemophilus influenzae*, *Streptococcus pneumoniae*, *Moraxella catarrhalis*) and qRT-PCR was used to detect 12 viruses of clinical relevance. *Haemophilus influenzae* isolates were tested for beta-lactamase production (N=69). Parallel to this, RNA extraction from NP swabs and sequencing was also done to conduct metatranscriptomic analysis using a bioinformatics approach. Using the sequencing data, several analyses were performed: pathogen detection and quantification, assembly of detected respiratory viruses, detection of beta-lactamase genes, and transcriptomic analysis of host responses.

2.2.2 Bacterial pathogen detection by metatranscriptomic analysis of NP samples

To identify potential bacterial and viral pathogens in the 221 samples, high throughput sequencing was performed of total RNA derived from NP swabs. First, the abundance of three bacterial pathogens of interest was quantified – *S. pneumoniae* (SPN), *M. catarrhalis* (MCAT), and *H. influenzae* (HFLU) – as these pathogens are commonly isolated in children with bacterial sinusitis (Wald et al., 1981). Note that the use of the term “pathogen” does not imply that these organisms are necessarily the causative agents of sinusitis infections. After quality filtering, taxonomic classification of the sequencing reads using Kraken 2 was performed (Wood et al., 2019). The relative abundance of the three bacterial pathogens (shown in Figure 2A) was calculated based on the normalized abundance of reads (reads per million, RPM) that mapped to each species. One or more of these three bacterial pathogens were detected in a total of 177 patients (80%). Two or more bacterial pathogens were detected in 89 (40%) patients, and 25 (11%) of patients had all three bacterial pathogens detected. On an individual basis, SPN was detected in 73 (33%), MCAT in 137 (62%), and HFLU in 81 (37%) of patient samples. Tables S1 and S2 contain the clinical culture and RNA-seq based results for bacterial detection for each patient.

Next, the extent that the calculated abundance of these bacterial pathogens from RNA-seq agreed with their presence/absence based on culture was examined. For all three pathogens, a significant increase in RNA-seq abundance in those with a positive culture was detected demonstrating concordance between metatranscriptomic data and culture (Figure 2B). Some pathogen-negative samples based on culture had an RNA-seq pathogen abundance greater or equal to the mean abundance seen in positive samples. Next, the ability of the RNA-seq data to predict the culture-based test results for each pathogen was assessed, and Receiver Operator Curves were generated by varying the detection threshold (Figure 2C). HFLU infections could be detected with the highest accuracy by RNA-seq with an area under the curve (AUC) of 0.95, SPN with an AUC of 0.89, and MCAT with an AUC of 0.82. Using a threshold of 3 reads per million, HFLU was detected with a sens/spec of 94%/90%, SPN with 81%/89% and MCAT with 85%/64% (Table 2).

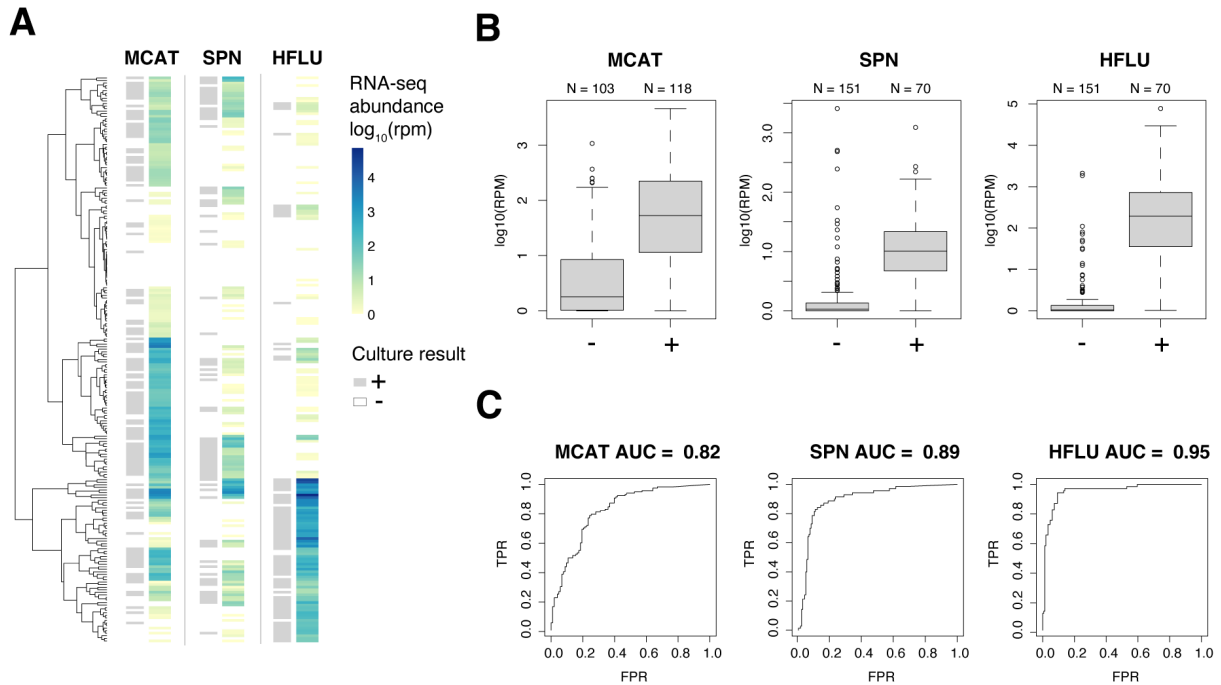


Figure 2. Metatranscriptomic detection of bacterial pathogens in NP samples from children with clinically diagnosed acute sinusitis.

(A) Heatmap showing the detected abundance of three bacterial pathogens (*H. influenzae*, *M. catarrhalis*, *S. pneumoniae*) in patient metatranscriptomes. For each bacterium, the culture based test result (positive – grey, negative – white) is shown on the left of the column, and the estimated RNA-seq abundance is depicted on the right of the column as a color gradient (absent – white, low – yellow, high – dark blue). Each row in the heatmap and tip in the hierarchical tree corresponds to an individual patient sample.

(B) Boxplots depicting pathogen abundance in positive (+) versus negative (-) samples (labeled on X axis) defined based on culture. The boxes show the interquartile range and median line, and the whiskers show the variability extending to the furthest data points within 1.5 times above and below the interquartile range. Outliers outside of these ranges are shown as data points.

(C) ROC curves illustrating specificity and sensitivity of metatranscriptomic pathogen detection with area under the curve (AUC) values displayed above.

2.2.3 Beta-lactamase gene detection in HFLU positive samples

Next, it was examined whether metatranscriptomics could identify potential resistance genes associated with HFLU. Culture-based tests for beta-lactamase were performed for all HFLU positive samples, and these were used as the reference standard to analyze the accuracy of RNA seq based detection. All non-human reads from samples that were clinically positive for HFLU (N=69) were assembled and used the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2023) to detect beta-lactamase genes with at least 10% coverage (Figure 3). Beta-lactamase genes were detected in 74.1% (20/27) of the samples associated with resistant HFLU, and in 33% (13/42) of the samples associated with non-resistance HFLU, which reflects a significant (2.1-fold) increase in detected beta-lactamase genes in the resistant samples ($p = 0.002$, Fisher exact test). The imperfect concordance between RNA-seq based and culture-based beta-lactamase detection reflects the known challenges in detecting AMR genes using metagenomic approaches (de Abreu et al., 2021). The complete list of genes and the portion of the reference genome detected for each hit can be found in tables S3-S5.

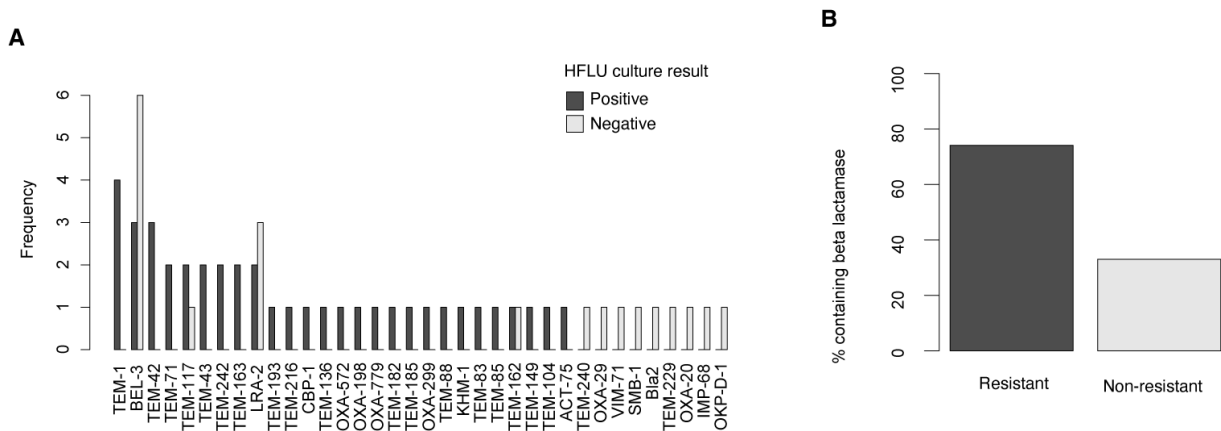


Figure 3. Detected beta-lactamase genes by CARD in resistant versus non-resistant HFLU samples.

(A) Frequency histogram of genes detected across all HFLU-positive and negative samples (based on culture tests).

(B) Percent of resistant and non-resistant samples with detected beta lactamase genes.

2.2.4 Metatranscriptomic detection and analysis of respiratory viruses

To examine the ability of metatranscriptomics to detect viral infections, respiratory viruses identified using qRT-PCR were focused on first. Viruses tested for included Influenza A (INFA), Influenza B (INFB), influenza C (INFC), human metapneumovirus (MPV), human rhinovirus (HRV, which tested for rhinovirus types A, B, and C), parainfluenza virus 1 (PIV1), parainfluenza virus 2 (PIV2), parainfluenza virus 3 (PIV3), parainfluenza virus 4 (PIV4), respiratory syncytial virus (RSV, types A and B), human adenovirus (ADV), and enterovirus D68 (EVD68). One or more viruses were detected by metatranscriptomics in 175 patients (79%), two or more in 101 patients (46%), and three or more in 36 patients (16%). HRV was detected most frequently (45%), followed by MPV (14%) and INFA (13%).

Next, the extent that the RNA-seq based predictions matched viral presence/absence based on the qRT-PCR was examined. As shown visually in Figure 4A, the relative abundance of viruses detected by metatranscriptomics was in strong agreement with the results of qRT-PCR based tests, with lower qRT-PCR cycle threshold (Ct) values corresponding to higher RPM values in RNA-seq. A significant inverse correlation ($r = 0.75$, $p = 1.3 \times 10^{-46}$) was detected between Ct values and viral load calculated as $\log_{10}(\text{reads per kilobase million, rpkm})$ (Graf et al., 2016) (Figure 4B). Samples containing viruses detected by qRT-PCR but not by RNA-seq had significantly higher cycle thresholds (mean = 34.7) compared to true positives (mean = 23.2; t-test p-value = 5.5×10^{-5}), which has been reported in previous RNA-seq studies (Thorburn et al., 2015). For all viruses except for INFC (which only had 8 positive samples), an increase in metatranscriptomic abundance in those with a positive qRT-PCR result was detected (Figure 4C).

The accuracy of viral detection was calculated by using the results of the qRT-PCR tests as the ground truth. Due to the uniqueness of viral sequences, it was found that a very low threshold (≥ 1 RPM) was sufficient to distinguish virus-positive from negative samples. Using this threshold, the sensitivity and specificity of metatranscriptomic pathogen detection was calculated for each of the 12 viruses as shown in Table 2. Nine out

of the 12 viruses were detected with 90-100% sensitivity and specificity, while INFC, HRV, and ADV were detected with lower accuracy. Overall, the 12 viruses were detected with an average sensitivity/specificity of 86%/92%. These accuracies are consistent with other studies performing sequencing-based pathogen detection using NP samples (Graf et al., 2016; Thorburn et al., 2015).

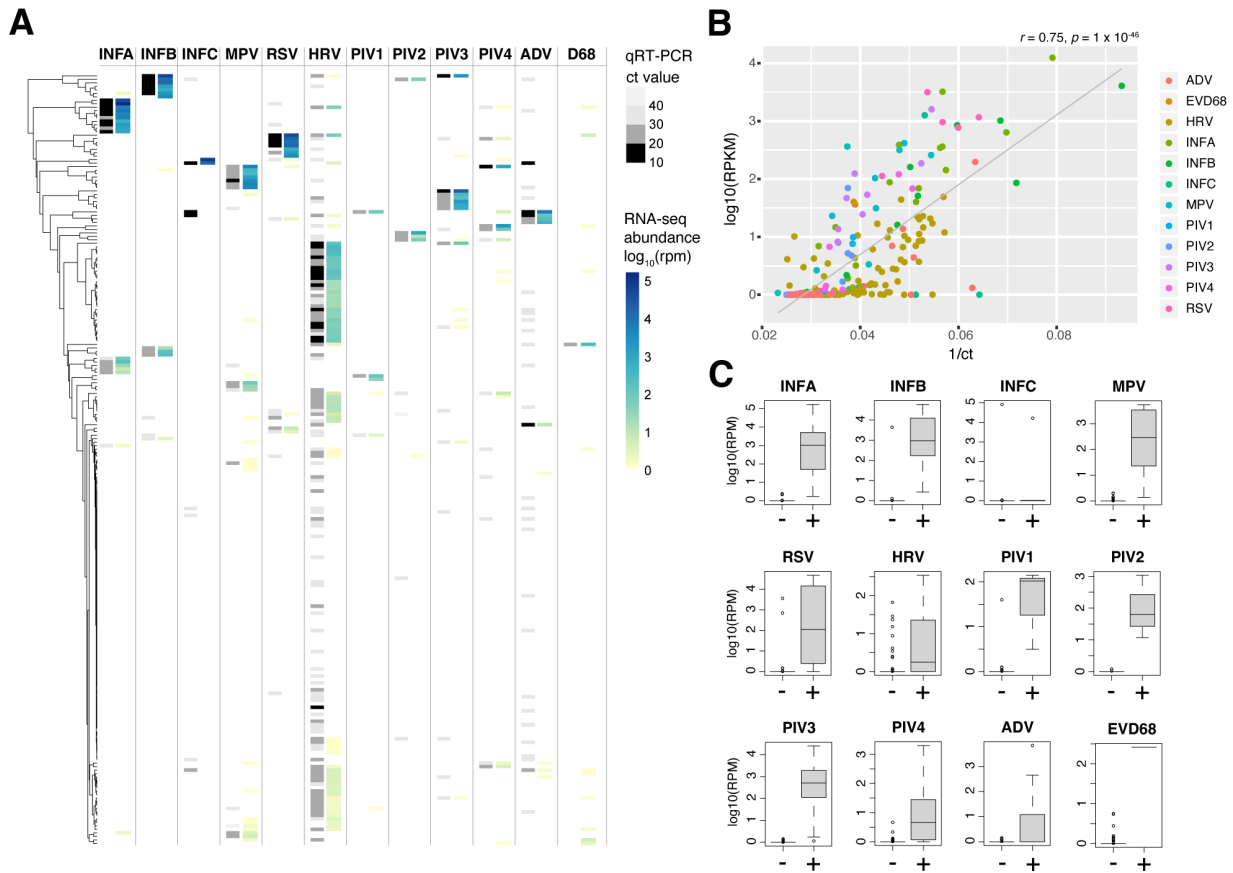


Figure 4. Detection of common respiratory viruses in NP metatranscriptomes. (A) Abundance heatmap for viruses detected in NP metatranscriptomes for 221 patients. For each virus, the qRT-PCR result is shown on the left of the column as a color gradient (negative – white, high to low cycle threshold values – light gray to black), and the estimated RNA-seq abundance is depicted on the right of the column as a color gradient (absent – white, low – yellow, high – dark blue). Each row in the heatmap and tip in the hierarchical tree corresponds to an individual patient sample. (B) qRT-PCR abundance (1/ cycle threshold) versus metatranscriptomic viral load (\log_{10} of the RPKM). The estimated viral load from RNA-seq is significantly correlated with 1/Ct value from qRT-PCR. (C) Metatranscriptomic abundance of respiratory viruses in negative (-) versus positive (+) samples (labelled on X axis) defined by qRT-PCR test result. The boxes show the interquartile range and median line, and the whiskers show the variability extending to the furthest data points within 1.5 times above and below the interquartile range. Outliers outside of these ranges are shown as data points.

2.2.5 RNA-seq uncovers additional pathogens and alternate explanations of disease etiology

By sequencing total RNA within a sample, metatranscriptomics has the potential to detect additional pathogens beyond those tested by culture or qRT-PCR. Therefore the RNA-seq dataset was screened for additional pathogens previously associated with URTIs and/or sinusitis infections, as well non-URTI pathogens and opportunistic pathogens, and further validated the identified species using additional bioinformatic approaches (see Methods). Across the 221 patient samples, 22 additional pathogens were detected that were not tested for clinically, including 11 bacteria and 11 viruses (Figure 5). These species were then ranked in terms of their maximum relative abundance within a sample (Figure 5).

Newly identified bacterial pathogens includes fourteen species listed in Figure 5. The most notable identifications include *Mycobacterium pneumoniae* and *Chlamydia pneumoniae*, which were not included in the clinical panel but have been previously implicated in pediatric sinusitis and URTIs (Blasi, 1996; Waites & Atkinson, 2009). In addition, opportunistic pathogens including *Fusobacterium nucleatum*, *Moraxella* spp., and others, were also detected (Figure 5), but some of these likely have a commensal role in the nasopharynx. Interestingly, periodontitis-associated bacteria were also detected including *Treponema medium*, *Prevotella intermedia*, and *Tannerella forsythia* (Pérez-Chaparro et al., 2014), in a few (N = 1 to 4) samples, and all three co-occurring in the same patient. Follow-up investigation of this patient revealed that they were admitted to an emergency room one year after the NP swab sample was taken with a severe tooth infection, highlighting the potential of NP RNA-seq to detect subclinical infection.

Newly identified viral pathogens with the highest abundance include four human coronaviruses known to cause upper respiratory infections (NL63, OC43, HKU1, and 229E). Parechovirus A and cardiovirus B (saffold virus) were detected, which have been associated with respiratory illness in children (Olijve et al., 2018; Zoll et al., 2009), as well as other viruses that are not typically associated with respiratory infections including mamastrovirus 9, enteroviruses A and B, human gammaherpes virus 5, human betaherpes virus 5, and sequences related to murine leukemia virus (Figure 5).

Of the 19 samples that had no pathogen detected by culture or qRT-PCR, 11 contained identified pathogens based on RNA-seq profiling. Three of the 11 samples (circled in Figure 5) contained known pathogens detected at high abundance that were not included in the clinical pathogen panel: the coronaviruses NL63 and 229E, and the bacterium, *Chlamydia pneumoniae*. Eight of the 11 samples had pathogens detected by RNA-seq but not by qRT-PCR or culture, including Influenza B (N = 1), parainfluenza virus 1 (N = 1), SPN (N = 1), MCAT (N = 4), and HFLU (N = 1).

Ultimately, these additional detected pathogens highlight the ability of RNA-seq to provide a more complete picture of the microbiome and virome present in acute sinusitis samples and suggests an expanded panel of viruses and bacterial pathogens to be used in future clinical workflows.

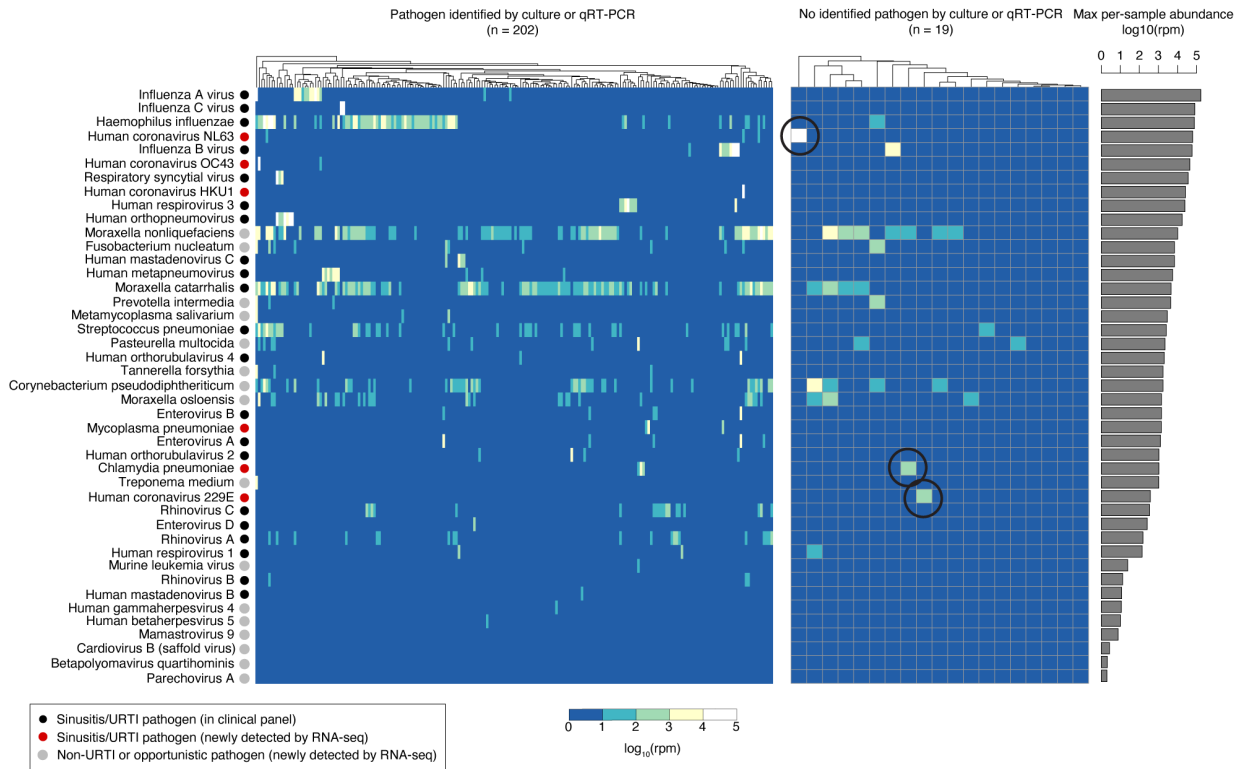


Figure 5. Metatranscriptomics of NP samples from children with acute sinusitis identified organisms not detected by qRT-PCR or culture. The organisms included in the heatmap are a subset of the full set of organisms detected by RNA-seq that exceed minimum abundance thresholds and include human pathogenic bacteria and viruses (see table S6 for full dataset). The organisms are sorted vertically based on their maximum relative abundance within a sample (across 221 samples). The heatmap displays the relative abundance of each organism in each sample as estimated by Kraken 2. The left heatmap includes samples with clinically identified pathogens by qRT-PCR or culture (N = 202), and the right heatmap includes 19 samples without a pathogen detected by qRT-PCR or culture. For the latter samples, several samples contain additional organisms identified by metatranscriptomics that are plausible causes of sinusitis. The barplot on the right depicts the total number of samples containing each detected pathogen.

2.2.6 Viral genome assembly and subtyping from host-derived metatranscriptomes

Read-based taxonomic classifications provide an estimate of microbial species present in each sample. However, de novo genome assembly methods may be used to assemble longer fragments including genomes of full-length RNA viruses, which can validate read-based predictions and reveal additional information. By aligning the RNA-seq reads to reference genomes of identified viruses, genomes were assembled with partial or complete coverage for a total of 205 viruses across 163 samples, including 25 different human pathogenic viruses (Figure 6A). In addition to the 12 viral groups from the clinical panel (Figure 4), genomes were assembled for 9 additional respiratory viruses (e.g., coronaviruses) not tested for clinically. Enterovirus A and B, WU polyomavirus, and mamastrovirus 9 genomes were also assembled, which are typically implicated in other illnesses such as gastroenteritis. A total of 31 (15%) were 100% complete, while 60 (30%) had completeness of at least 90% or more (table S7). All assembled viral genomes were phylogenetically verified by sequence comparison to related genomes in NCBI through BLAST, with average nucleotide identities (ANIs) ranging from 95- 100%.

To explore the use of assembled genomes for viral subtyping, I focused on the predictions for Influenza A and B, since these were subtyped clinically using qRT-PCR. The subtyping results using assembled influenza genomes showed excellent agreement with the clinical results, with Influenza A subtypes H1N1 and H3N2 having 100% (15/15) agreement and Influenza B subtypes Yamagata and Victoria having 82% agreement (9/11) with qRT-PCR results (table S8). Then, several cases of interest were focused on, performing a deeper genomic and phylogenetic analysis of newly assembled genomes. Three examples of assembled viral genomes are shown in Figure 6B, including a genome of a novel COV OC43 strain, an RSV B genome, and an enterovirus D68 genome. All three of these genomes have distinct mutation profiles from other strains in the NCBI database (Figure 6B), and clustered as unique strains in phylogenetic analysis (Figure 6C). All three of the genomes also showed broad sequencing coverage across the genome, with the exception of the RSV B genome from sample 1141, which showed a lack of coverage spanning the glycoprotein G

gene. Interestingly, a previous study also identified G protein deletion mutant RSV strains in pediatric pneumonia patients from South Africa (Venter et al., 2011).

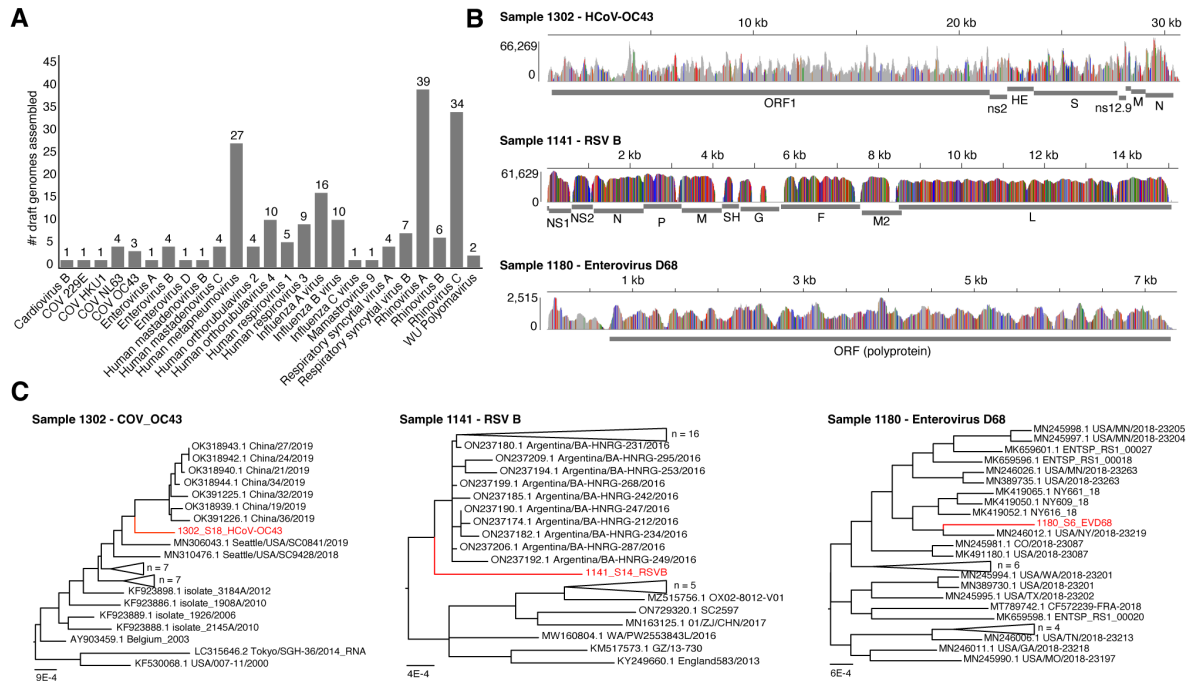


Figure 6. Assembled genomes of viruses from children with clinically diagnosed acute sinusitis.

(A) Bar graph depicting the number of assembled genomes for various species of respiratory viruses across the full dataset (N = 205 total viruses assembled from 163 samples).

(B) Read pileups for three selected samples showing sequencing reads mapped to reference genomes of coronavirus (COV) OC43, RSV, and enterovirus D68 with SNP profiles as colored lines.

(C) Phylogenetic analysis of three assembled viral genomes and their top 25 closest matching complete genomes from BLAST. Each newly assembled virus (red) is a unique strain that clusters as a distinct branch within its phylogenetic tree.

2.2.7 Host-response expression profiles distinguish bacterial from viral infections

Although RNA-seq analysis was capable of detecting pathogens directly from reads, most reads within RNA-seq samples were host (human) derived, ranging from 64.7-99.9%, which enables host-response profiling to potentially identify host biomarkers and immune responses associated with disease etiology (Butler et al., 2021; Cheemarla et al., 2023; Sims et al., 2022; Wesolowska-Andersen et al., 2017). To identify differentially expressed genes (DEGs) associated with bacterial versus viral infections, the host gene expression profiles of patients with bacterial pathogens were compared to those with viral pathogens based on clinical diagnostic testing (Figure 7A). Due to the presence of many ($N = 138$) complex samples containing a mixture of viral and bacterial pathogens, the initial comparison was simplified by compared samples with only bacterial pathogens ($N = 33$) to those with only viral pathogens ($N = 31$), but subsequently analyzed all 221 samples. A total of 821 significant DEGs were detected with $q < 0.001$, of which 548 genes had increased expression in bacterial-positive patients and 273 genes had increased expression in viral-positive patients (Figure 6A, table S9). These genes were termed as “bacterial upDEGs” and “viral upDEGs”.

Based on function enrichment analysis, bacterial upDEGs were significantly associated with neutrophil regulation, regulation of inflammatory response, response to lipopolysaccharide, and response to molecule of bacterial origin (Figure 7B), which are consistent with an immune response to bacterial infection. The identified bacterial upDEGs include genes previously shown to be markers of bacterial infection: for example, PTGS2 (6-fold increase in bacterial-positive patients, $q = 3.1 \times 10^{-7}$), S100A9 (4-fold increase, $q = 4.2 \times 10^{-6}$), PLAUR (5-fold increase, $q = 7.3 \times 10^{-6}$), TNFAIP3 (4-fold increase, $q = 1.3 \times 10^{-5}$), IL1A (6-fold increase, $q = 1.0 \times 10^{-4}$), IL1B (6-fold increase, $q = 4.0 \times 10^{-5}$), CXCL2 (4-fold increase, $q = 1.3 \times 10^{-5}$), and NFKBIA (4-fold increase, $q = 1.8 \times 10^{-5}$) (Figure 7D).

Viral upDEGs were found to be significantly associated with cytokine signaling, defense response to virus, T cell receptor signaling, and inflammatory response (Figure 7C), which

are related to viral immune response pathways. Consistent with this, the identified viral upDEGs include genes shown to be markers of viral infection in previous studies: for example, CXCL11 which was increased 33-fold in virus-positive patients ($q = 4.9 \times 10^{-23}$), CXCL10 (15-fold increase, $q = 2.6 \times 10^{-15}$), CCL8 (23-fold increase, $q = 2.3 \times 10^{-6}$), PRF1 (4-fold increase, $q = 3.8 \times 10^{-9}$) and IFI27 (2-fold increase, $q = 8.5 \times 10^{-7}$), which represent putative biomarkers of viral infection in the analysis (Figure 7D).

In general, representative viral and bacterial upDEGs had lower expression levels for samples in which no bacteria or virus was detected by qRT-PCR/culture, and higher expression levels for samples containing both a virus and bacterial pathogen (Figure 7D). Interestingly, there are several exceptions to this pattern including four samples that had a strong antiviral response despite there being no virus detected by qRT-PCR/culture. Deeper investigation of these samples by RNA seq revealed that three of them contained respiratory viruses (two coronaviruses and Influenza B) (Figure 5B) that were not detected by the qRT-PCR tests. Other exceptions include two samples which had no bacterial pathogen detected by culture/qRT-PCR but had a strong antibacterial response. One of these samples (sample 1303) had a bacterial pathogen (MCAT) identified in high abundance by RNA-seq. These results suggest that host-response profiling may provide an indication of viral or bacterial infection when traditional tests fail to detect a pathogen.

2.2.8 Magnitude of host responses correlates with viral and bacterial pathogen abundance

If the identified viral and bacterial upDEGs are genuine biomarkers of viral and bacterial infections, respectively, then their levels of expression should correlate with the abundance of viral and bacterial pathogens estimated from RNA-seq. To test this hypothesis, the total bacterial pathogen abundance was calculated as the sum of the relative abundance of the pathogens SPN, HFLU, and MCAT. The, all the samples were binned into ten groups, with group 1 having the lowest bacterial pathogen abundance, and group 10 having the highest. This analysis was then repeated for viral pathogens, summing the total abundance of 12 viral pathogens as well as the coronaviruses that were clearly present based on RNA-seq data, but missing from the clinical test.

As shown in Figure 8A, with increasing abundance of bacterial sinusitis pathogens (MCAT, SPN, HFLU), there is a clear increase in expression levels of bacterial upDEGs. To quantify this pattern, for each sample the “magnitude” of the bacterial and viral host response was calculated as the average expression level (Z-score) of the bacterial and viral upDEGs. As shown in Figure 8B, the magnitude of bacterial host response correlated significantly with bacterial pathogen abundance (Pearson $r = 0.50$, two-tailed $p = 1.6 \times 10^{-15}$). The same pattern was also seen for viruses: that is, the abundance of viral pathogens also correlated significantly with the magnitude of viral host response (Pearson $r = 0.33$, two-tailed $p = 5.8 \times 10^{-7}$) (Figure 8C,D). Both the bacterial and viral host responses however did not correlate with other clinical features including the duration of cold symptoms and symptom severity (Figure 8A). Although these pathogen-host-response correlations are a general pattern, not all samples display this trend. For example, several samples with high bacterial pathogen abundance lack a strong bacterial host response. In addition, one outlier (marked * in Figure 8A) shows an individual with a low detected bacterial pathogen abundance but a strong bacterial host response. This could indicate an immune response to an unknown bacterial species.

In addition to the association between host-response and pathogen abundance, host-response correlations with other clinical metadata was also tested for. A weaker but sig-

nificant ($r = 0.33$, $p = 6.6 \times 10^{-7}$) host-response pattern was detected between a subset of genes and patient symptom severity scores (PRSS) at the time of diagnosis. A total of 45 genes were differentially expressed as a function of PRSS, which subdivided into 2 expression clusters (Figure 9). Cluster 1 was positively correlated with PRSS and includes the following genes: METTL7B, MMP3, PRF1, GNLY, MMP1, FPR3, GIMAP6, OLFML2B, DESI1, IL12RB2. Function enrichment analysis revealed that cluster 1 was associated with a response to infection (cellular defense response, natural killer cell mediated immunity, and cellular response to cytokine stimulus). Other pathways such as proteolysis and pyroptosis are also involved in innate host immune response by eliminating and degrading infected cells (Dardevet, 2016; Yu et al., 2021).

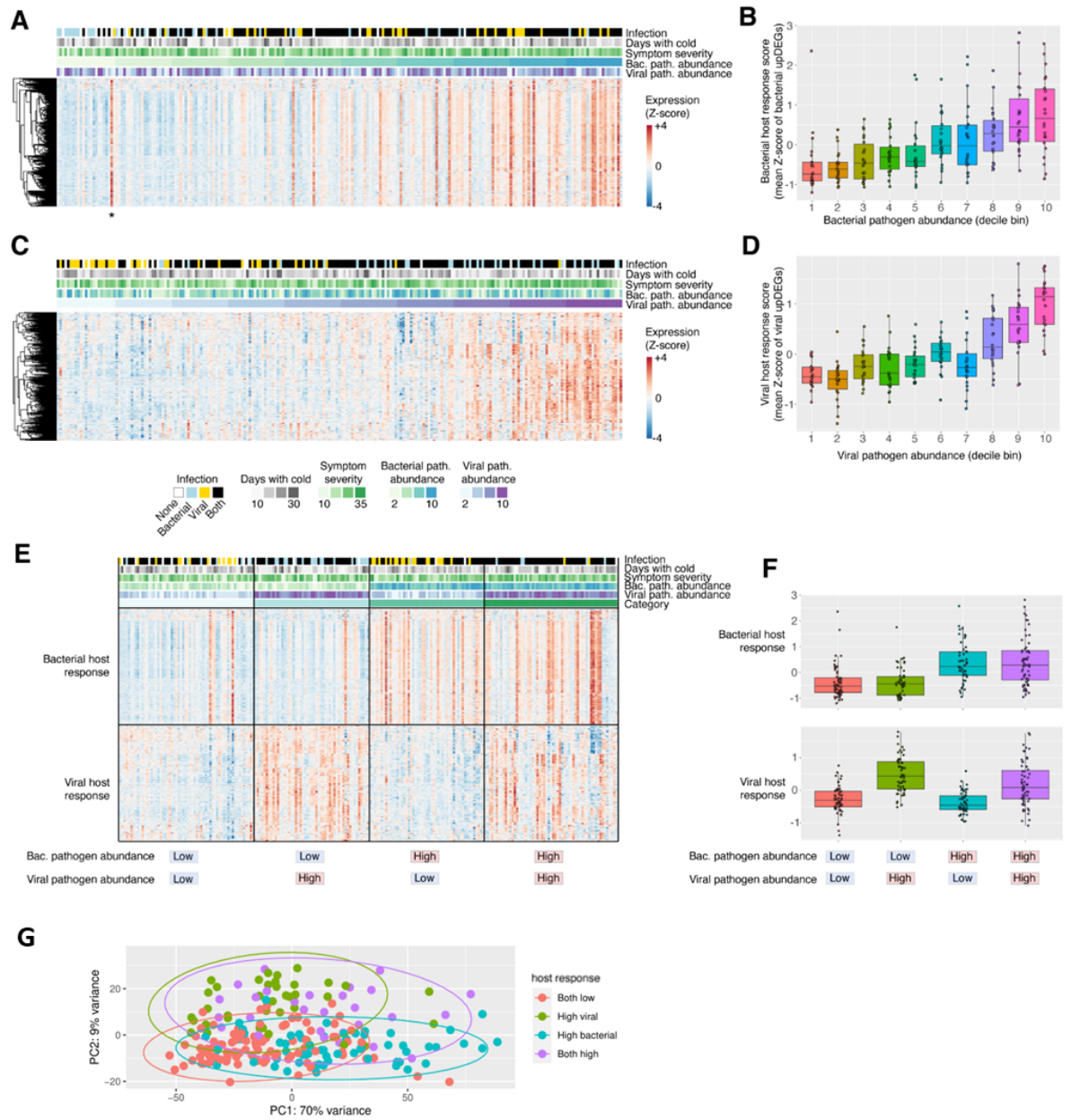


Figure 8. Host-response correlates with relative abundance of bacterial and viral pathogens.

- (A) Expression heatmap of bacterial upDEGs (bacterial host response genes), with samples (columns) sorted by total metatranscriptomic bacterial pathogen abundance. The associated metadata for all samples is also plotted above the heatmap. * Also shown is an outlier sample associated with a strong bacterial host response but with low detected abundance of MCAT, HFLU, or SPN.
- (B) Bacterial host response score versus metatranscriptomic bacterial pathogen abundance. The bacterial host response score was calculated as the mean expression level (Z-scores) of all the bacterial upDEG genes.
- (C) Expression heatmap of viral upDEGs (viral host response genes), with samples (columns) sorted by metatranscriptomic viral pathogen abundance.
- (D) Viral host response score versus metatranscriptomic viral pathogen abundance. The viral host response score was calculated as the mean expression level (Z-scores) of all the viral upDEG genes.
- (E) Heatmap of bacterial and viral host responses (upDEGs), where samples (columns) have been sorted into four groups based on high or low bacterial/viral pathogen abundance, with high considered as a 60th percentile or greater relative abundance. In general, samples with low bacterial and viral abundance tend to lack a bacterial/viral host response, whereas samples containing bacteria, viruses, or both displayed the appropriate response.
- (F) Jitter plots of the bacterial and viral host response scores across four categories of samples. Bacterial and viral host response scores were calculated by averaging the expression level Z-scores of all bacterial and viral upDEGs, respectively.
- (G) PCA plot of the 821 bacterial and viral up-regulated genes where samples have been sorted into four groups based on high or low bacterial/viral pathogen abundance.

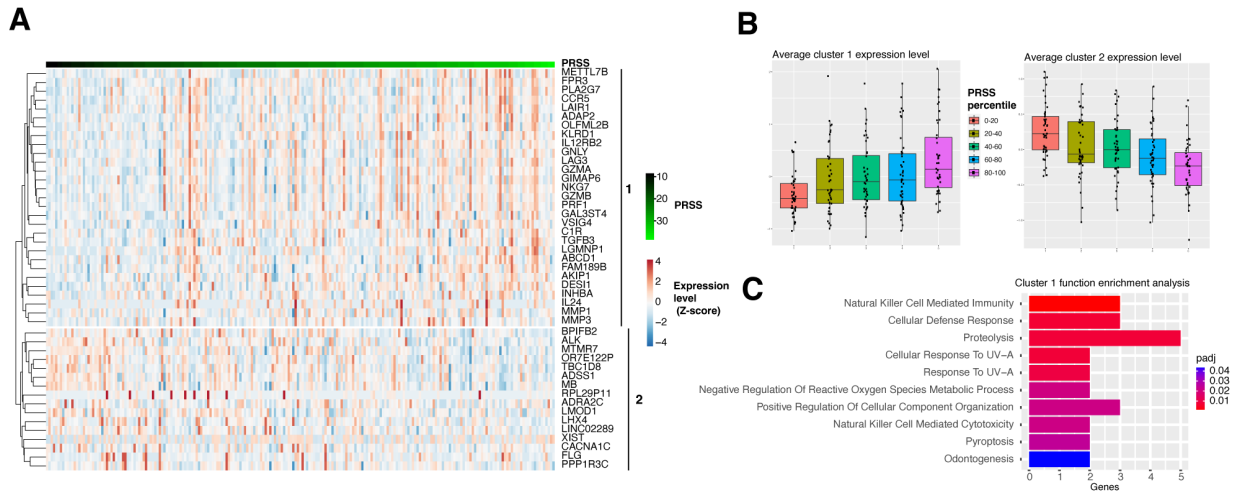


Figure 9. Differential host response expression analysis based on patients' symptom severity score (PRSS) at time of sample collection.

(A) Heatmap displays the DEGs with $q < 0.05$. A total of 45 genes were differentially expressed and are divided into 2 clusters on the heatmap based on their expression patterns.

(B) Average expression level (Z-scores) of genes in cluster 1 and cluster 2 for five PRSS percentile categories.

(C) Significantly enriched ($q < 0.05$) GO Biological Process 2021 database pathways results using EnrichR of gene cluster 1.

2.2.9 RNA-seq classifies patients into distinct groups with unique pathogen-host response profiles

After examining host responses to bacterial and viral infections individually, it was considered how bacterial and viral relative abundance together impact host responses within patients. To investigate this, the RNA-seq abundance was used to bin samples into four groups: those with low bacterial / low viral pathogen abundance (N = 60, 27%), high viral / low bacterial pathogen abundance (N = 51, 23%), high bacterial / low viral pathogen abundance (N = 51, 23%), and high bacterial / high viral pathogen abundance (N = 59, 27%). Here, the thresholds of “high” and “low” pathogen abundance based on RNA-seq estimated levels (≥ 60 th percentile) and not the presence/absence classification obtained from qRT-PCR and culture-based testing.

The four groups of patients display distinct host response signatures (Figure 8E,F). As expected, samples with low bacterial and low viral pathogen abundance tend to have weak bacterial and antiviral responses (Figure 8E). Samples with high viral abundance but low bacterial abundance display a strong antiviral pattern and a weak bacterial response. Samples with high bacterial pathogen abundance but low viral pathogen abundance are associated with a strong bacterial host response, and samples with high bacterial and viral pathogen abundance show both host responses. Again, there are several outliers that are exception to these general trends. The viral host response for individuals with both bacterial and viral pathogens was lower than the viral only group ($p = 0.01$), and the bacterial host response for individuals with both bacterial and viral pathogens was not significantly different from the bacterial-only group ($p = 0.82$). Finally, whether the bacterial and viral host-response magnitude alone could predict samples with high pathogen abundance was tested, with pathogen abundance defined as described above using RNA-seq measurements. The bacterial host response magnitude predicted high-bacterial samples with an area under the receiver operator curve (AUROC) of 0.79, and the viral host response magnitude predicted high-virus samples with an AUROC of 0.80. If sensitivity is desired over specificity, high-bacterial samples could be predicted with a sensitivity/specificity of 80%/68% using host-response information alone. Ultimately, these analyses suggest that host-response information alone may have diagnostic value in differentiating between viral and bacterial sinus infections, especially when the relative abundance (pathogen load) is high.

Chapter 3

Discussion

In this study, metatranscriptomic analysis of 221 NP samples from children with clinically diagnosed acute sinusitis was performed. Prior to this work, there has been a lack of research evaluating the use and applications of RNA-seq profiling in this clinical context. This study provides several research contributions. First, it highlights the ability of RNA-sequencing of clinical samples to accurately identify bacterial and viral pathogens associated with sinusitis infections and URTIs. Second, it provides an original dataset to assist with the development of future bioinformatic approaches for infectious disease profiling, including hundreds of assembled viral pathogen genomes contributing to ongoing pathogen genomic surveillance efforts. Third, it identifies host-response signatures of bacterial and viral infections in sinusitis, which could serve as the basis for the development of biomarker assays to be used in future clinical workflows that optimize delivery of care. Using RNA-seq, an overall sensitivity of 87% and specificity of 81% was achieved in reproducing the clinical results for detection of three bacterial species that are mostly commonly implicated in sinusitis (Wald et al., 1981). RNA-seq also demonstrated a significant ability to detect viral pathogens that were also detected by the qRT-PCR panel (average sens/spec of 86%/92%), as well as predict viral load (Ct value). These accuracies are comparable to results obtained by previous studies using NGS for pathogen detection in NP samples (Graf et al., 2016; Thorburn et al., 2015). It is important to note, however, that the accuracies differed by pathogen, with metatranscriptomic detection of HFLU and SPN

performing very well, and MCAT performing relatively poorly. The weaker specificity for MCAT may indicate a potential issue with taxonomic profiling accuracy for this species, or issues with the culture-based test itself (i.e., potential false positive cultures due to growth of non-target organisms).

For clinical decision making regarding antibiotic treatment, a key goal of sequencing-based approaches is to not only detect the pathogen of interest but also its antimicrobial genes, which can be especially challenging in mixed metagenomic samples. As proof of principle, beta-lactamase resistance in HFLU isolates was focused on, which represents a key clinical issue (Eldere et al., 2014; Tristram et al., 2007). As done previously for pediatric nose and ear samples (Lobb et al., 2023) CARD was used (Alcock et al., 2023) to identify beta lactamases in RNA-seq data. This RNA-seq workflow was able to correctly detect beta-lactamase genes in 74% of the resistant HFLU isolates, with a specificity of 67%. Additionally, beta lactam resistance SNPs in the *Haemophilus influenzae* PBP3 gene were also detected in several samples, which may represent an additional resistance mechanism that was detected by RNA-seq profiling but not covered by clinical AMR testing. A difficulty in metatranscriptomic detection of AMR, however, is the abundance of AMR genes present that may be contributed by non-target organisms. Due to the fragmented nature of metatranscriptomic data, it may be difficult or impossible to associate AMR gene fragments with their parent organisms using current approaches. These factors may lead to a tendency of sequencing approaches to overpredict AMR potential. Despite this, significant increase was found in beta-lactamase genes in resistant versus non-resistant HFLU samples, suggesting that the approach used in this study is able to distinguish HFLU resistance with a significant degree of specificity.

Finally, through *de novo* assembly methods, 205 viral pathogens genomes were assembled with varying degrees of completeness. Assembled genomes confirm read-based predictions and provide added information that cannot be obtained from short sequencing reads or qRT-PCR-based methods. For example, phylogenetic analyses of some of these viruses (e.g., COV OC43, RSV B, enterovirus D68) revealed unique differences from closely related genomes in the database, suggesting that they represent distinct strains. A potentially relevant mutation (absence of large segments of the G gene) was identified in an RSV

B strain similar to previous reports (Venter et al., 2011). Further analysis of RSV genomes from patient samples is needed to determine the frequency of G deletion mutants, which could be important information to consider for RSV vaccine design.

An advantage of metatranscriptomic RNA-seq over culture or qRT-PCR is the ability to perform a broad and untargeted analysis to detect any species whose genome is available in the reference database, which theoretically improves sensitivity of pathogen detection and discovery. Out of 221 pediatric sinusitis patients tested, 19 did not have any bacterial or viral pathogen detected by culture-based or qRT-PCR testing. RNA-seq identified plausible pathogens for acute sinusitis in 11 of these 19 samples including cases of Influenza B and PIV1 that were missed by qRT-PCR. Not surprisingly, several new pathogenic bacteria and viruses were also detected in these samples and were verified by genome assembly and phylogenetics. These included two coronaviruses (NL63 and 229E), as well as the bacterium, *Chlamydia pneumoniae*. Other identified organisms included commensal organisms of the nasal microbiome and opportunistic pathogens that may or may not play a direct role in sinusitis (e.g., different species of *Moraxella* and *Corynebacterium*). Clarifying the role of these and other species in sinusitis etiology is a challenging goal for future work.

One of the most exciting aspects of this study is the identified host-response gene expression patterns associated with bacterial and viral sinusitis infections. Since the pathogen composition of the patient cohort was complex including a large number of samples containing both bacterial and viral pathogens based on culture/qRT-PCR, the initial comparison was simplified by looking at virus-positive only samples versus bacteria-positive only samples. This enabled the detection of virus associated and bacteria associated host DEGs (“viral host response” and “bacterial host response”) that formed the basis of subsequent analyses. Function enrichment analysis of these two gene sets revealed that they were significantly associated with expected pathways: e.g., viral upDEGs were associated with “T Cell Activation”, “cytokine-mediated signaling” and “defense response to viruses”, while bacterial upDEGs were associated with “inflammatory response” and “response to lipopolysaccharide”. Remarkably, the magnitude of these host responses correlated significantly with the total abundance of bacterial or viral pathogens detected in the samples.

Importantly, this correlation between pathogen abundance and host-response magnitude was only identified for a limited subset of bacterial species (those previously identified as sinusitis pathogens, MCAT, SPN, HFLU) and respiratory viruses, and the correlation was absent when examining other species detected in the data that may reflect commensal organisms. This finding indicates that the relative abundance of specific bacterial and viral species within the nasopharynx is a determinant of the strength of the host immune response. This is consistent with immunology since the expression of host antiviral and antibacterial pathways are dependent on the levels of viral (e.g., dsRNA) and bacterial pathogen-associated molecular patterns (e.g., lipopolysaccharide) sensed by the host immune system. Previous studies have also reported a correlation between antiviral host responses in RNA-seq and viral load (Cheemarla et al., 2021; Landry & Foxman, 2018; Saravia-Butler et al., 2022). However, this study is unique by analyzing the interplay between a complex mixture of bacterial and viral pathogens and their impact on the host transcriptomic response.

Although traditional methods (culture and qRT-PCR) provided a simple classification of the samples based on detected presence/absence of a pre-defined set of pathogens, metatranscriptomic data enabled a more holistic classification based on pathogen abundance and host-response information (Figure 8). When taking both pathogen abundance and host-response information into consideration, the samples could be subdivided into four main groups: those with a “low” abundance of bacterial or viral pathogens which tend to lack a host-response, and those with a “high” abundance of bacterial pathogens, respiratory viruses, or both, which tend to show the expected host responses. Interestingly, the observed correlation between pathogen abundance and host-response is not perfect; there are several outlier samples which exhibited a strong host response pattern and yet lack a detected pathogen, and other samples which contained a high pathogen abundance but lack a detectable host response. For the former category, it is possible that those samples contained other pathogens that were not included in the pathogen panel, which may include opportunistic infections by commensal organisms for example. For the latter category, these cases could indicate delayed host-responses in patients at the time of sampling, shedding of viral RNA at a post-infection time point which may be associated

with a reduced host-response, or simply an imperfect correlation between host-responses and pathogen abundance. Nevertheless, future research focusing on host-responses of patients with infectious disease and factors that account for discrepancies between detected pathogen abundance, could clarify mechanistic understanding of disease etiology.

There are several limitations of this study that could account for variation in the results obtained. First, the classification into viral and bacterial infection was inferred based on the presence/absence of bacterial and viral pathogens, but some of these organisms may be present as commensals and their presence alone does not necessitate an infection (Henriques-Normark & Normark, 2010; Karalus & Campagnari, 2000). Second, the enrollment criteria for this study recruited patients experiencing symptoms for at least 6 days when sampled. Since peak shedding of some viruses can occur within 48 hours of symptom onset, the chosen sampling time may have led to a reduced sensitivity of viral detection as well as lower coverage for genomes assembled. Variation in the timing of bacterial infections could also impact sensitivity of bacterial detection by RNA-seq. Third, the sensitivity for pathogen detection by RNA-seq is dependent on the depth of sequencing. Deeper sequencing may have been necessary to detect viruses, for example, that were false negatives by RNA-seq but were detected using qRT PCR. DNA viruses in particular (e.g., adenoviruses) may have been more prone to weak detection due to the use of RNA-seq over DNA-seq. Future studies that employ both metatranscriptomic and metagenomic sequencing with repeated time-series sampling of patients may overcome some of the limitations described above. Nevertheless, the current study provides a starting framework for exploring the use of high-throughput sequencing of patient samples to uncover etiology and host response in pediatric sinusitis and other upper respiratory infections.

Conclusion

In conclusion, diagnosis of acute sinusitis poses a significant challenge in pediatric medicine, being a major driver of antibiotic prescriptions. The accurate discrimination between bacterial and viral causes is imperative to avoid unnecessary antibiotic administration, yet current diagnostic methods face considerable limitations. However, untargeted RNA se-

quencing of clinical samples from pediatric patients with suspected acute sinusitis presents a promising alternative. The work done in this thesis, involving RNA-seq analysis of nasopharyngeal specimens from 221 clinically diagnosed AS pediatric cases, yielded encouraging results. Metatranscriptomic pathogen detection exhibited significant agreement with traditional culture and qRT-PCR methods, demonstrating sensitivities and specificities of 87%/81% for bacteria and 86%/92% for viruses, respectively. Furthermore, the analysis unveiled pathogens that were not identified by conventional methods and provided insights into host-response signatures distinguishing between bacterial and viral infections. Notably, the identification of novel strains of coronaviruses, respiratory syncytial virus, and enterovirus D68 underscores the potential of RNA sequencing in comprehensive pathogen characterization. By correlating pathogen abundance with host gene expression, the findings offer valuable insights into optimizing patient care strategies for acute sinusitis. In summary, this work highlights the utility of untargeted metatranscriptomics in elucidating the etiology of AS and emphasizes its potential for advancing personalized patient care in this clinical context.

Future work

This work has shown that metatranscriptomics has the potential to be used as a diagnostic tool for acute pediatric sinusitis. Future studies can focus on combining both RNA and DNA sequencing to increase the accuracy of detection and sample at different time points during infection. Future studies can also build on this work by selecting a subset of the differentially expressed genes reported here to create a predictive model or clinical test to differentiate between bacterial and viral infection.

References

- Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., . . . McArthur, A. G. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *51*, D690–D699. <https://doi.org/10.1093/nar/gkac920>
- Andrews, S. (2010, January 1). *Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data*. Retrieved March 29, 2023, from <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Anzai, Y., & Paladin, A. (2010). Diagnosis and management of acute and chronic sinusitis in children. In L. S. Medina, K. E. Applegate, & C. C. Blackmore (Eds.), *Evidence-based imaging in pediatrics: Optimizing imaging in pediatric patient care* (pp. 141–159). Springer. https://doi.org/10.1007/978-1-4419-0922-0_11
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing [Publisher: Mary Ann Liebert, Inc., publishers]. *Journal of Computational Biology*, *19*(5), 455–477. <https://doi.org/10.1089/cmb.2012.0021>
- Bhattacharya, S., Rosenberg, A. F., Peterson, D. R., Grzesik, K., Baran, A. M., Ashton, J. M., Gill, S. R., Corbett, A. M., Holden-Wiltse, J., Topham, D. J., Walsh, E. E., Mariani, T. J., & Falsey, A. R. (2017). Transcriptomic biomarkers to discriminate bacterial from nonbacterial infection in adults hospitalized with respiratory illness

- [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, 7(1), 6548. <https://doi.org/10.1038/s41598-017-06738-3>
- Blasi, F. (1996). Clinical features of chlamydia pneumoniae acute respiratory infection. *Clinical Microbiology and Infection*, 1, S14–S18. <https://doi.org/10.1111/j.1469-0691.1996.tb00585.x>
- Boolchandani, M., D’Souza, A. W., & Dantas, G. (2019). Sequencing-based methods and resources to study antimicrobial resistance [Number: 6 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, 20(6), 356–370. <https://doi.org/10.1038/s41576-019-0108-4>
- Borchers, A. T., Chang, C., Gershwin, M. E., & Gershwin, L. J. (2013). Respiratory syncytial virus—a comprehensive review. *Clinical Reviews in Allergy & Immunology*, 45(3), 331–379. <https://doi.org/10.1007/s12016-013-8368-9>
- Bosch, A. A. T. M., Biesbroek, G., Trzcinski, K., Sanders, E. A. M., & Bogaert, D. (2013). Viral and bacterial interactions in the upper respiratory tract. *PLoS Pathogens*, 9(1), e1003057. <https://doi.org/10.1371/journal.ppat.1003057>
- Branche, A. R., & Falsey, A. R. (2016). Parainfluenza virus infection [Publisher: Thieme Medical Publishers]. *Seminars in Respiratory and Critical Care Medicine*, 37(4), 538–554. <https://doi.org/10.1055/s-0036-1584798>
- Brealey, J. C., Sly, P. D., Young, P. R., & Chappell, K. J. (2015). Viral bacterial co-infection of the respiratory tract during early childhood. *FEMS Microbiology Letters*, 362(10), fnv062. <https://doi.org/10.1093/femsle/fnv062>
- Brook, I. (2011). Microbiology of sinusitis [Publisher: American Thoracic Society - PATS]. *Proceedings of the American Thoracic Society*, 8(1), 90–100. <https://doi.org/10.1513/pats.201006-038RN>
- Brook, I. (2013). Acute sinusitis in children. *Pediatric Clinics of North America*, 60(2), 409–424. <https://doi.org/10.1016/j.pcl.2012.12.002>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND [Number: 1 Publisher: Nature Publishing Group]. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Bushnell, B. (2014, March 17). *BBMap: A fast, accurate, splice-aware aligner* (LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). Retrieved March 29, 2023, from <https://www.osti.gov/biblio/1241166>

- Butler, D., Mozsary, C., Meydan, C., Foox, J., Rosiene, J., Shaiber, A., Danko, D., Afshinnekoo, E., MacKay, M., Sedlazeck, F. J., Ivanov, N. A., Sierra, M., Pohle, D., Zietz, M., Gisladdottir, U., Ramlall, V., Sholle, E. T., Schenck, E. J., Westover, C. D., . . . Mason, C. E. (2021). Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 1660. <https://doi.org/10.1038/s41467-021-21361-7>
- Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: Opportunities and challenges [Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *17*(5), 257–271. <https://doi.org/10.1038/nrg.2016.10>
- Caldeweyher, E. (2021). Kallisto: A command-line interface to simplify computational modelling and the generation of atomic features. *Journal of Open Source Software*, *6*(60), 3050. <https://doi.org/10.21105/joss.03050>
- Castañeda-Mogollón, D., Kamaliddin, C., Oberding, L., Liu, Y., Mohon, A. N., Faridi, R. M., Khan, F., & Pillai, D. R. (2021). A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity. *Journal of Clinical Virology*, *145*, 105025. <https://doi.org/10.1016/j.jcv.2021.105025>
- Charlton, C. L., Babady, E., Ginocchio, C. C., Hatchette, T. F., Jerris, R. C., Li, Y., Loeffelholz, M., McCarter, Y. S., Miller, M. B., Novak-Weekley, S., Schuetz, A. N., Tang, Y.-W., Widen, R., & Drews, S. J. (2018). Practical guidance for clinical microbiology laboratories: Viruses causing acute respiratory tract infections [Publisher: American Society for Microbiology]. *Clinical Microbiology Reviews*, *32*(1), e00042–18. <https://doi.org/10.1128/CMR.00042-18>
- Cheemarla, N. R., Hanron, A., Fauver, J. R., Bishai, J., Watkins, T. A., Brito, A. F., Zhao, D., Alpert, T., Vogels, C. B. F., Ko, A. I., Schulz, W. L., Landry, M. L., Grubaugh, N. D., Dijk, D. v., & Foxman, E. F. (2023). Nasal host response-based screening for undiagnosed respiratory viruses: A pathogen surveillance and detection study [Publisher: Elsevier]. *The Lancet Microbe*, *4*(1), e38–e46. [https://doi.org/10.1016/S2666-5247\(22\)00296-8](https://doi.org/10.1016/S2666-5247(22)00296-8)
- Cheemarla, N. R., Watkins, T. A., Mihaylova, V. T., Wang, B., Zhao, D., Wang, G., Landry, M. L., & Foxman, E. F. (2021). Dynamic innate immune response determines sus-

- ceptibility to SARS-CoV-2 infection and early replication kinetics. *The Journal of Experimental Medicine*, 218(8), e20210583. <https://doi.org/10.1084/jem.20210583>
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1), 128. <https://doi.org/10.1186/1471-2105-14-128>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics [Number: 6 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, 20(6), 341–355. <https://doi.org/10.1038/s41576-019-0113-7>
- Choudhary, M. L., Anand, S. P., Tikhe, S. A., Walimbe, A. M., Potdar, V. A., Chadha, M. S., & Mishra, A. C. (2016). Comparison of the conventional multiplex RT–PCR, real time RT–PCR and luminex xTAG® RVP fast assay for the detection of respiratory viruses [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.24299>]. *Journal of Medical Virology*, 88(1), 51–57. <https://doi.org/10.1002/jmv.24299>
- Dardevet, D. (2016, January 1). *THE MOLECULAR NUTRITION OF AMINO ACIDS AND PROTEINS. nutrition series*.
- Dasaraju, P. V., & Liu, C. (1996). Infections of the respiratory system. In S. Baron (Ed.), *Medical microbiology* (4th). University of Texas Medical Branch at Galveston. Retrieved December 3, 2022, from <http://www.ncbi.nlm.nih.gov/books/NBK8142/>
- de Abreu, V. A. C., Perdigão, J., & Almeida, S. (2021). Metagenomic approaches to analyze antimicrobial resistance: An overview. *Frontiers in Genetics*, 11. Retrieved October 2, 2023, from <https://www.frontiersin.org/articles/10.3389/fgene.2020.575592>
- DeMuri, G. P., & Wald, E. R. (2012). Acute bacterial sinusitis in children [Publisher: Massachusetts Medical Society eprint: <https://doi.org/10.1056/NEJMcp1106638>]. *New England Journal of Medicine*, 367(12), 1128–1134. <https://doi.org/10.1056/NEJMcp1106638>
- de Vries, J. J. C., Brown, J. R., Couto, N., Beer, M., Le Mercier, P., Sidorov, I., Papa, A., Fischer, N., Oude Munnink, B. B., Rodriguez, C., Zaheri, M., Sayiner, A., Hönemann, M., Pérez-Cataluña, A., Carbo, E. C., Bachofen, C., Kubacki, J., Schmitz, D., Tsioka, K., ... López-Labrador, F. X. (2021). Recommendations for

- the introduction of metagenomic next-generation sequencing in clinical virology, part II: Bioinformatic analysis and reporting. *Journal of Clinical Virology*, 138, 104812. <https://doi.org/10.1016/j.jcv.2021.104812>
- de Vries, J. J. C., Brown, J. R., Fischer, N., Sidorov, I. A., Morfopoulou, S., Huang, J., Munnink, B. B. O., Sayiner, A., Bulgurcu, A., Rodriguez, C., Gricourt, G., Keyaerts, E., Beller, L., Bachofen, C., Kubacki, J., Cordey, S., Laubscher, F., Schmitz, D., Beer, M., . . . Claas, E. C. J. (2021). Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *Journal of Clinical Virology*, 141, 104908. <https://doi.org/10.1016/j.jcv.2021.104908>
- Dulanto Chiang, A., & Dekker, J. P. (2020). From the pipeline to the bedside: Advances and challenges in clinical metagenomics. *The Journal of Infectious Diseases*, 221, S331–S340. <https://doi.org/10.1093/infdis/jiz151>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Eldere, J. V., Slack, M. P. E., Ladhani, S., & Cripps, A. W. (2014). Non-typeable haemophilus influenzae, an under-recognised pathogen [Publisher: Elsevier]. *The Lancet Infectious Diseases*, 14(12), 1281–1292. [https://doi.org/10.1016/S1473-3099\(14\)70734-0](https://doi.org/10.1016/S1473-3099(14)70734-0)
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Florensa, A. F., Kaas, R. S., Clausen, P. T. L. C., Aytan-Aktug, D., & Aarestrup, F. M. (2022). ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics*, 8(1), 000748. <https://doi.org/10.1099/mgen.0.000748>
- Flynn, M. F., Kelly, M., & Dooley, J. S. G. (2021). Nasopharyngeal swabs vs. nasal aspirates for respiratory virus detection: A systematic review. *Pathogens*, 10(11), 1515. <https://doi.org/10.3390/pathogens10111515>
- Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, 9(1), 207–216. <https://doi.org/10.1038/ismej.2014.106>

- Graf, E. H., Simmon, K. E., Tardif, K. D., Hymas, W., Flygare, S., Eilbeck, K., Yandell, M., & Schlager, R. (2016). Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: A systematic comparison to a commercial PCR panel [Publisher: American Society for Microbiology]. *Journal of Clinical Microbiology*, *54*(4), 1000–1007. <https://doi.org/10.1128/JCM.03060-15>
- Henriques-Normark, B., & Normark, S. (2010). Commensal pathogens, with a focus on streptococcus pneumoniae, and interactions with the human host. *Experimental Cell Research*, *316*(8), 1408–1414. <https://doi.org/10.1016/j.yexcr.2010.03.003>
- Hersh, A. L., Jackson, M. A., Hicks, L. A., & American Academy of Pediatrics Committee on Infectious Diseases. (2013). Principles of judicious antibiotic prescribing for upper respiratory tract infections in pediatrics. *Pediatrics*, *132*(6), 1146–1154. <https://doi.org/10.1542/peds.2013-3260>
- Holm-Hansen, C. C., Midgley, S. E., & Fischer, T. K. (2016). Global emergence of enterovirus d68: A systematic review. *The Lancet Infectious Diseases*, *16*(5), e64–e75. [https://doi.org/10.1016/S1473-3099\(15\)00543-5](https://doi.org/10.1016/S1473-3099(15)00543-5)
- Jacobs, S. E., Lamson, D. M., St. George, K., & Walsh, T. J. (2013). Human rhinoviruses [Publisher: American Society for Microbiology]. *Clinical Microbiology Reviews*, *26*(1), 135–162. <https://doi.org/10.1128/CMR.00077-12>
- Javanian, M., Barary, M., Ghebrehewet, S., Koppolu, V., Vasigala, V., & Ebrahimpour, S. (2021). A brief review of influenza virus infection. *Journal of Medical Virology*, *93*(8), 4638–4646. <https://doi.org/10.1002/jmv.26990>
- Kang, S. H., Cheong, H. J., Song, J. Y., Noh, J. Y., Jeon, J. H., Choi, M. J., Lee, J., Seo, Y. B., Lee, J.-S., Wie, S.-H., Jeong, H. W., Kim, Y. K., Park, K. H., Kim, S. W., Jeong, E. J., Lee, S. H., Choi, W. S., & Kim, W. J. (2016). Analysis of risk factors for severe acute respiratory infection and pneumonia and among adult patients with acute respiratory illness during 2011-2014 influenza seasons in Korea [Publisher: The Korean Society of Infectious Diseases and Korean Society for Chemotherapy]. *Infection & Chemotherapy*, *48*(4), 294–301. <https://doi.org/10.3947/ic.2016.48.4.294>
- Karalus, R., & Campagnari, A. (2000). *Moraxella catarrhalis*: A review of an important human mucosal pathogen. *Microbes and Infection*, *2*(5), 547–559. [https://doi.org/10.1016/S1286-4579\(00\)00314-2](https://doi.org/10.1016/S1286-4579(00)00314-2)

- Khanal, S., Ghimire, P., & Dhamoon, A. S. (2018). The repertoire of adenovirus in human disease: The innocuous to the deadly [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Biomedicines*, *6*(1), 30. <https://doi.org/10.3390/biomedicines6010030>
- Khomich, O. A., Kochetkov, S. N., Bartosch, B., & Ivanov, A. V. (2018). Redox biology of respiratory viral infections [Number: 8 Publisher: Multidisciplinary Digital Publishing Institute]. *Viruses*, *10*(8), 392. <https://doi.org/10.3390/v10080392>
- Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Research*, *26*(12), 1721–1729. <https://doi.org/10.1101/gr.210641.116>
- Knipe, D. M., & Howley, P. (2013, June 25). *Fields virology (knipe, fields virology)-2 volume set* (Sixth edition). LWW.
- Ko, E. R., Yang, W. E., McClain, M. T., Woods, C. W., Ginsburg, G. S., & Tsalik, E. L. (2015). What was old is new again: Using the host response to diagnose infectious disease [Publisher: Taylor & Francis eprint: <https://doi.org/10.1586/14737159.2015.1059278>]. *Expert Review of Molecular Diagnostics*, *15*(9), 1143–1158. <https://doi.org/10.1586/14737159.2015.1059278>
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. <https://doi.org/10.1101/pdb.top084970>
- Landry, M. L., & Foxman, E. F. (2018). Antiviral response in the nasopharynx identifies patients with respiratory virus infection. *The Journal of Infectious Diseases*, *217*(6), 897–905. <https://doi.org/10.1093/infdis/jix648>
- Leung, A. K., Hon, K. L., & Chu, W. C. (2020). Acute bacterial sinusitis in children: An updated review. *Drugs in Context*, *9*, 2020–9–3. <https://doi.org/10.7573/dic.2020-9-3>
- Li, C.-X., Li, W., Zhou, J., Zhang, B., Feng, Y., Xu, C.-P., Lu, Y.-Y., Holmes, E. C., & Shi, M. (2020). High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory infection [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *10*(1), 3963. <https://doi.org/10.1038/s41598-020-60992-6>

- Lobb, B., Lee, M. C., McElheny, C. L., Doi, Y., Yahner, K., Hoberman, A., Martin, J. M., Hirota, J. A., Doxey, A. C., & Shaikh, N. (2023). Genomic classification and antimicrobial resistance profiling of streptococcus pneumoniae and haemophilus influenza isolates associated with paediatric otitis media and upper respiratory infection. *BMC infectious diseases*, *23*(1), 596. <https://doi.org/10.1186/s12879-023-08560-x>
- Lopez, S. M. C., Martin, J. M., Johnson, M., Kurs-Lasky, M., Horne, W. T., Marshall, C. W., Cooper, V. S., Williams, J. V., & Shaikh, N. (2019). A method of processing nasopharyngeal swabs to enable multiple testing. *Pediatric Research*, *86*(5), 651–654. <https://doi.org/10.1038/s41390-019-0498-1>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Lu, B., Yan, Y., Dong, L., Han, L., Liu, Y., Yu, J., Chen, J., Yi, D., Zhang, M., Deng, X., Wang, C., Wang, R., Wang, D., Wei, H., Liu, D., & Yi, C. (2021). Integrated characterization of SARS-CoV-2 genome, microbiome, antibiotic resistance and host response from single throat swabs [Publisher: Nature Publishing Group]. *Cell Discovery*, *7*(1), 1–10. <https://doi.org/10.1038/s41421-021-00248-3>
- Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *7*(1), 11257. <https://doi.org/10.1038/ncomms11257>
- Mohammad, H. A., Madi, N. M., & Al-Nakib, W. (2020). Analysis of viral diversity in stool samples from infants and children with acute gastroenteritis in kuwait using metagenomics approach. *Virology Journal*, *17*(1), 10. <https://doi.org/10.1186/s12985-020-1287-5>
- Mostafa, H. H., Fissel, J. A., Fanelli, B., Bergman, Y., Gniazdowski, V., Dadlani, M., Carroll, K. C., Colwell, R. R., & Simner, P. J. (2020). Metagenomic next-generation sequencing of nasopharyngeal specimens collected from confirmed and suspect COVID-19 patients [Publisher: American Society for Microbiology]. *mBio*, *11*(6), e01969–20. <https://doi.org/10.1128/mBio.01969-20>
- Mulcahy-O’Grady, H., & Workentine, M. L. (2016). The challenge and potential of metagenomics in the clinic. *Frontiers in Immunology*, *7*, 29. <https://doi.org/10.3389/fimmu.2016.00029>

- Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., . . . Nakaya, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach [Publisher: Public Library of Science]. *PLOS ONE*, *4*(1), e4219. <https://doi.org/10.1371/journal.pone.0004219>
- Olijve, L., Jennings, L., & Walls, T. (2018). Human parechovirus: An increasingly recognized cause of sepsis-like illness in young infants. *Clinical Microbiology Reviews*, *31*(1), e00047–17. <https://doi.org/10.1128/CMR.00047-17>
- Omics of antimicrobials and antimicrobial resistance [Publisher: Taylor & Francis]. (2019). *Expert Opinion on Drug Discovery*, *14*(5), 455–468. <https://doi.org/10.1080/17460441.2019.1588880>
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. <https://doi.org/10.1186/s13059-016-0997-x>
- Panda, S., Mohakud, N. K., Pena, L., & Kumar, S. (2014). Human metapneumovirus: Review of an important respiratory pathogen. *International Journal of Infectious Diseases*, *25*, 45–52. <https://doi.org/10.1016/j.ijid.2014.03.1394>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression [Number: 4 Publisher: Nature Publishing Group]. *Nature Methods*, *14*(4), 417–419. <https://doi.org/10.1038/nmeth.4197>
- Pérez-Chaparro, P. J., Gonçalves, C., Figueiredo, L. C., Faveri, M., Lobão, E., Tamashiro, N., Duarte, P., & Feres, M. (2014). Newly identified pathogens associated with periodontitis: A systematic review. *Journal of Dental Research*, *93*(9), 846–858. <https://doi.org/10.1177/0022034514542468>
- Petrenko, P., Lobb, B., Kurtz, D. A., Neufeld, J. D., & Doxey, A. C. (2015). MetAnnotate: Function-specific taxonomic profiling and comparison of metagenomes. *BMC Biology*, *13*(1), 92. <https://doi.org/10.1186/s12915-015-0195-4>
- Piantadosi, A., Mukerji, S. S., Ye, S., Leone, M. J., Freimark, L. M., Park, D., Adams, G., Lemieux, J., Kanjilal, S., Solomon, I. H., Ahmed, A. A., Goldstein, R., Ganesh, V.,

- Ostrem, B., Cummins, K. C., Thon, J. M., Kinsella, C. M., Rosenberg, E., Frosch, M. P., ... Sabeti, P. (2021). Enhanced virus detection and metagenomic sequencing in patients with meningitis and encephalitis [Publisher: American Society for Microbiology]. *mBio*, *12*(4), e01143–21. <https://doi.org/10.1128/mBio.01143-21>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Quintanilla-Dieck, L., & Lam, D. J. (2018). Chronic rhinosinusitis in children. *Current Treatment Options in Pediatrics*, *4*(4), 413–424. <https://doi.org/10.1007/s40746-018-0142-z>
- Rajagopala, S. V., Bakhoun, N. G., Pakala, S. B., Shilts, M. H., Rosas-Salazar, C., Mai, A., Boone, H. H., McHenry, R., Yooseph, S., Halasa, N., & Das, S. R. (2021). Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. *Cell Reports Methods*, *1*(6), 100091. <https://doi.org/10.1016/j.crmeth.2021.100091>
- Rajapakse, N., & Dixit, D. (2021). Human and novel coronavirus infections in children: A review. *Paediatrics and International Child Health*, *41*(1), 36–55. <https://doi.org/10.1080/20469047.2020.1781356>
- Ramchandrar, N., Burns, J., Coufal, N. G., Pennock, A., Briggs, B., Stinnett, R., Bradley, J., Arnold, J., Liu, G. Y., Pring, M., Upasani, V. V., Rickert, K., Dimmock, D., Chiu, C., Farnaes, L., & Cannavino, C. (2021). Use of metagenomic next-generation sequencing to identify pathogens in pediatric osteoarticular infections. *Open Forum Infectious Diseases*, *8*(7), ofab346. <https://doi.org/10.1093/ofid/ofab346>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Ross, M. H., Zick, B. L., & Tsalik, E. L. (2019). Host-based diagnostics for acute respiratory infections. *Clinical Therapeutics*, *41*(10), 1923–1938. <https://doi.org/10.1016/j.clinthera.2019.06.007>
- Saravia-Butler, A. M., Schisler, J. C., Taylor, D., Beheshti, A., Butler, D., Meydan, C., Foox, J., Hernandez, K., Mozsary, C., Mason, C. E., & Meller, R. (2022). Host transcriptional responses in nasal swabs identify potential SARS-CoV-2 infection in

- PCR negative patients. *iScience*, 25(11), 105310. <https://doi.org/10.1016/j.isci.2022.105310>
- Schlaberg, R., Queen, K., Simmon, K., Tardif, K., Stockmann, C., Flygare, S., Kennedy, B., Voelkerding, K., Bramley, A., Zhang, J., Eilbeck, K., Yandell, M., Jain, S., Pavia, A. T., Tong, S., & Ampofo, K. (2017). Viral pathogen detection by metagenomics and pan-viral group polymerase chain reaction in children with pneumonia lacking identifiable etiology. *The Journal of Infectious Diseases*, 215(9), 1407–1415. <https://doi.org/10.1093/infdis/jix148>
- Schomacker, H., Schaap-Nutt, A., Collins, P. L., & Schmidt, A. C. (2012). Pathogenesis of acute respiratory illness caused by human parainfluenza viruses. *Current Opinion in Virology*, 2(3), 294–299. <https://doi.org/10.1016/j.coviro.2012.02.001>
- Scotta, M. C., Chakr, V. C. B. G., de Moura, A., Becker, R. G., de Souza, A. P. D., Jones, M. H., Pinto, L. A., Sarria, E. E., Pitrez, P. M., Stein, R. T., & Mattiello, R. (2016). Respiratory viral coinfection and disease severity in children: A systematic review and meta-analysis. *Journal of Clinical Virology*, 80, 45–56. <https://doi.org/10.1016/j.jcv.2016.04.019>
- Shaikh, N., Hoberman, A., Shope, T. R., Jeong, J.-H., Kurs-Lasky, M., Martin, J. M., Bhatnagar, S., Muniz, G. B., Block, S. L., Andrasko, M., Lee, M. C., Rajakumar, K., & Wald, E. R. (2023). Identifying children likely to benefit from antibiotics for acute sinusitis: A randomized clinical trial. *JAMA*, 330(4), 349–358. <https://doi.org/10.1001/jama.2023.10854>
- Sims, J. T., Poorbaugh, J., Chang, C.-Y., Holzer, T. R., Zhang, L., Engle, S. M., Beasley, S., Doman, T. N., Naughton, L., Higgs, R. E., Kallewaard, N., & Benschop, R. J. (2022). Relationship between gene expression patterns from nasopharyngeal swabs and serum biomarkers in patients hospitalized with COVID-19, following treatment with the neutralizing monoclonal antibody bamlanivimab. *Journal of Translational Medicine*, 20(1), 134. <https://doi.org/10.1186/s12967-022-03345-3>
- Sinha, M., Jupe, J., Mack, H., Coleman, T. P., Lawrence, S. M., & Fraley, S. I. (2018). Emerging technologies for molecular diagnosis of sepsis. *Clinical Microbiology Reviews*, 31(2), e00089–17. <https://doi.org/10.1128/CMR.00089-17>
- Tada, A., & Hanada, N. (2010). Opportunistic respiratory pathogens in the oral cavity of the elderly. *FEMS Immunology & Medical Microbiology*, 60(1), 1–17. <https://doi.org/10.1111/j.1574-695X.2010.00709.x>

- Tan, L. V., Hong, N. T. T., Ngoc, N. M., Thanh, T. T., Lam, V. T., Nguyet, L. A., Nhu, L. N. T., Ny, N. T. H., Minh, N. N. Q., Man, D. N. H., Hang, V. T. T., Khanh, P. N. Q., Xuan, T. C., Phong, N. T., Tu, T. N. H., Hien, T. T., Hung, L. M., Truong, N. T., Yen, L. M., ... Chau, N. V. V. (2020). SARS-CoV-2 and co-infections detection in nasopharyngeal throat swabs of COVID-19 patients by metagenomics [Publisher: Elsevier]. *Journal of Infection*, *81*(2), e175–e177. <https://doi.org/10.1016/j.jinf.2020.06.033>
- Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., & Murcia, P. R. (2015). The use of next generation sequencing in the diagnosis and typing of respiratory infections. *Journal of Clinical Virology*, *69*, 96–100. <https://doi.org/10.1016/j.jcv.2015.06.082>
- Toma, R., Duval, N., Shen, N., Torres, P. J., Camacho, F. R., Chen, J., Ogundijo, O., Banavar, G., & Vuyisich, M. (2022). Pathogen detection and characterization from throat swabs using unbiased metatranscriptomic analyses. *International Journal of Infectious Diseases*, *122*, 260–265. <https://doi.org/10.1016/j.ijid.2022.05.062>
- Tristram, S., Jacobs, M. R., & Appelbaum, P. C. (2007). Antimicrobial resistance in haemophilus influenzae [Publisher: American Society for Microbiology]. *Clinical Microbiology Reviews*, *20*(2), 368–389. <https://doi.org/10.1128/CMR.00040-06>
- Troy, N. M., & Bosco, A. (2016). Respiratory viral infections and host responses; insights from genomics. *Respiratory Research*, *17*(1), 156. <https://doi.org/10.1186/s12931-016-0474-9>
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlan2 for enhanced metagenomic taxonomic profiling [Number: 10 Publisher: Nature Publishing Group]. *Nature Methods*, *12*(10), 902–903. <https://doi.org/10.1038/nmeth.3589>
- Venter, M., van Niekerk, S., Rakgantso, A., & Bent, N. (2011). Identification of deletion mutant respiratory syncytial virus strains lacking most of the g protein in immunocompromised children with pneumonia in south africa. *Journal of Virology*, *85*(16), 8453–8457. <https://doi.org/10.1128/JVI.00674-11>
- Vollmers, J., Wiegand, S., & Kaster, A.-K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! [Publisher: Public Library of Science]. *PLOS ONE*, *12*(1), e0169662. <https://doi.org/10.1371/journal.pone.0169662>

- Waites, K. B., & Atkinson, T. P. (2009). The role of mycoplasma in upper respiratory infections. *Current Infectious Disease Reports*, *11*(3), 198–206. <https://doi.org/10.1007/s11908-009-0030-6>
- Wald, E. R., Milmoie, G. J., Bowen, A., Ledesma-Medina, J., Salamon, N., & Bluestone, C. D. (1981). Acute maxillary sinusitis in children. *The New England Journal of Medicine*, *304*(13), 749–754. <https://doi.org/10.1056/NEJM198103263041302>
- Wang, Y., Zhu, N., Li, Y., Lu, R., Wang, H., Liu, G., Zou, X., Xie, Z., & Tan, W. (2016). Metagenomic analysis of viral genetic diversity in respiratory samples from children with severe acute respiratory infection in china. *Clinical Microbiology and Infection*, *22*(5), 458.e1–458.e9. <https://doi.org/10.1016/j.cmi.2016.01.006>
- Waskito, L. A., Rezkitha, Y. A. A., Vilaichone, R.-k., Wibawa, I. D. N., Mustika, S., Sugihartono, T., & Miftahussurur, M. (2022). Antimicrobial resistance profile by metagenomic and metatranscriptomic approach in clinical practice: Opportunity and challenge [Number: 5 Publisher: Multidisciplinary Digital Publishing Institute]. *Antibiotics*, *11*(5), 654. <https://doi.org/10.3390/antibiotics11050654>
- Wesolowska-Andersen, A., Everman, J. L., Davidson, R., Rios, C., Herrin, R., Eng, C., Janssen, W. J., Liu, A. H., Oh, S. S., Kumar, R., Fingerlin, T. E., Rodriguez-Santana, J., Burchard, E. G., & Seibold, M. A. (2017). Dual RNA-seq reveals viral infections in asthmatic children without respiratory illness which are associated with changes in the airway transcriptome. *Genome Biology*, *18*(1), 12. <https://doi.org/10.1186/s13059-016-1140-8>
- Wickham, H. (2011). Ggplot2 [<https://onlinelibrary.wiley.com>]. *WIREs Computational Statistics*, *3*(2), 180–185. <https://doi.org/10.1002/wics.147>
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, *20*(1), 257. <https://doi.org/10.1186/s13059-019-1891-0>
- Xie, F., Duan, Z., Zeng, W., Xie, S., Xie, M., Fu, H., Ye, Q., Xu, T., & Xie, L. (2021). Clinical metagenomics assessments improve diagnosis and outcomes in community-acquired pneumonia. *BMC Infectious Diseases*, *21*(1), 352. <https://doi.org/10.1186/s12879-021-06039-1>
- Xu, L., Zhu, Y., Ren, L., Xu, B., Liu, C., Xie, Z., & Shen, K. (2017). Characterization of the nasopharyngeal viral microbiome from children with community-acquired pneumonia but negative for luminex xTAG respiratory viral panel assay detec-

- tion [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.24895>]. *Journal of Medical Virology*, 89(12), 2098–2107. <https://doi.org/10.1002/jmv.24895>
- Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: Improvements for better sequence analysis. *Nucleic Acids Research*, 34, W6–W9. <https://doi.org/10.1093/nar/gkl164>
- Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification [Publisher: Elsevier]. *Cell*, 178(4), 779–794. <https://doi.org/10.1016/j.cell.2019.07.010>
- Yin, X., Jiang, X.-T., Chai, B., Li, L., Yang, Y., Cole, J. R., Tiedje, J. M., & Zhang, T. (2018). ARGs-OAP v2.0 with an expanded SARG database and hidden markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, 34(13), 2263–2270. <https://doi.org/10.1093/bioinformatics/bty053>
- Yu, P., Zhang, X., Liu, N., Tang, L., Peng, C., & Chen, X. (2021). Pyroptosis: Mechanisms and diseases [Number: 1 Publisher: Nature Publishing Group]. *Signal Transduction and Targeted Therapy*, 6(1), 1–21. <https://doi.org/10.1038/s41392-021-00507-5>
- Zhang, N., Wang, L., Deng, X., Liang, R., Su, M., He, C., Hu, L., Su, Y., Ren, J., Yu, F., Du, L., & Jiang, S. (2020). Recent advances in the detection of respiratory virus infection in humans [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.25674>]. *Journal of Medical Virology*, 92(4), 408–417. <https://doi.org/10.1002/jmv.25674>
- Zhu, X., Ge, Y., Wu, T., Zhao, K., Chen, Y., Wu, B., Zhu, F., Zhu, B., & Cui, L. (2020). Co-infection with respiratory pathogens among COVID-2019 cases. *Virus Research*, 285, 198005. <https://doi.org/10.1016/j.virusres.2020.198005>
- Zoll, J., Hulshof, S. E., Lanke, K., Lunel, F. V., Melchers, W. J. G., Ven, E. S.-v. d., Roivainen, M., Galama, J. M. D., & Kuppeveld, F. J. M. v. (2009). Saffold virus, a human theiler's-like cardiovirus, is ubiquitous and causes infection early in life [Publisher: PLOS]. *PLoS Pathogens*, 5(5). <https://doi.org/10.1371/journal.ppat.1000416>

Appendix

Supplementary Data

Supplementary data can be found at <https://github.com/nuabumaz/Masters-thesis/>