Identification of microbial pathogens from deep sequencing in humans and rats

From the Mekong Delta, Vietnam

By Zeeshan Nissar

Submitted to the Board of Biology

School of Natural and Social Sciences

in partial fulfillment of the requirements

for the degree of Bachelor of Arts

Purchase College

State University of New York

May 2019

Sponsor: Dr. Stephen Harris

Table of Contents

Abstract

The emergence of enzootic zoonoses have seen worrying growth recently. This is in part through the advent of better means of travel and increased population growth, setting off a perpetual arms race of coevolution pitting host and parasite against one another resulting in the loss of billions in dollars annually and thousands of lives. In recent years, South Asia has been identified as a hotspot for emerging infectious disease due to its rapid population growth and closeness to disease causing species. In an effort to stem the rapid rise of pathogen infection and emergence of infectious diseases, we utilized samples collected from a known rodent vector, *Rattus argentiventer,* and rat traders from open air meat markets in the Mekong Delta to identify potential viruses, monitor microbial diversity, and identify any potential spill over from rats to humans. These samples were taken from 300 individuals. Of these total samples, 250 humans were split between a control set and rat traders; with the remaining 50 being rat samples. These samples were comprised of respiratory swabs and blood plasma samples. We identified bacteria from samples using next-generation sequencing and bioinformatic tools and performed diversity analyses utilizing the OneCodex database as the comparison library. It was established that there were a number of rodent borne pathogens within the human population of the tested locations, in both human controls as well as the rat traders themselves. It was ultimately concluded that based on the results, a definitive answer cannot be made on potential spill over due to the infection of human controls with rodent borne viruses. However, it can be affirmatively concluded that there are a number of disease causing bacteria, rodent in origin, within the larger population of all humans in tested regions in the Mekong Delta.

Introduction

Historical Context on Zoonotic Pathogens

Humans throughout history have always faced a seemingly endless number of infectious diseases that pass through a population with varying degrees of intensity affecting much of the social aspects of society (Allen et al. 2017). As global civilizations move towards increased population densities, the risk of encountering communicable infectious diseases rises exponentially due to poor living conditions in urban centers which have become favorable settings for insects, rodent, and other livestock vectors (Neiderud 2015). Dense populations increase interactions between humans and potential wildlife which could be vectors for disease (Neiderud 2015) and if the disease transfers from animal to human it becomes classified as a zoonotic pathogen. It is estimated that zoonotic pathogens represent 60 % of emerging infectious diseases and are the major risk to human health, as recent outbreaks of Ebola and Zika viruses support (Jones et al. 2008). Additionally, anthropogenic change is reducing natural habitats, forcing feral animals to move into human settlements to supplement diet with human food subsidies (Sharma, 2018). In cases of synanthropic avian species in urban areas, more than half the diet of foraging species consists of human derived food, an indication of increased interaction between the birds and humans (West et al. 2016).

In the past 20 years endemic and enzootic zoonosis has resulted in at least one billion cases of illness and millions of deaths (Cunningham, Daszak, and Wood 2017). The economic cost of diagnosing, treating, and rebounding from these pandemics ranges in the billions of dollars from the loss of livestock, human life, and resources consumed combating outbreaks (Narrod, Zinsstag, and Tiongco 2012). The advent of modern travel has shown the ease that pathogens can spread to different ends of the globe, showing that outbreak risks are no longer limited in scope to developing countries. As the world population increases, the inherent result of rapid population growth without proper microbial

control is self-evident, microbes will exploit new niches and adapt to new hosts. In coevolution, there is a perpetual race between the host and parasite to continually adapt and evolve with a constantly evolving environment, where the "loser" risks extinction (Valen 1973). As documented in each new era of human livestock turn over, a new emergence of zoonotic pathogens and the challenges that accompany are presented as seen in the influenza outbreaks of the past 100 years (Cunningham, Daszak, and Wood 2017).

The cycle of zoonosis emergence is predictable; humans enter into a new area and engage in some close contact with an infected species. This is believed to be the case with the Nipah virus outbreak in Bangladesh in 2004. Fruit bats, *Pteropus,* drinking from the same clay pots over night as the farmers transmitted Nipah virus directly to the workers. There were 12 total cases of which 11 fatalities were recorded (Luby et al. 2006). Often, the animal passing the microbe is not affected and is considered a reservoir species, a species that host and allow a parasite to proliferate (Ashford 2003). Often times, these species carry the infectious diseases for extended periods of time without effect to the host body (Mandl et al. 2015). No one reason has been attributed to this immune response on a whole for reservoir species but in specific cases like bats and the outbreak of the Nipah virus, bats were able to tolerate inflammation for a longer period of time by dampening the immune response (Mandl et al. 2015). The protein, NLRP3, which is found in in many species such as bats, humans and mice, barely reacts to virus inflammation in bats as compared to humans and mice (Ahn et al. 2019). The inflammation response proves more damaging then the virus infection (Kuriakose and Kanneganti 2017).

Zoonotic emergence is followed by the infection cycle, which is also predictable, and occurs in different stages. Many infectious diseases transmitted to humans are not immediately transferable to other humans. One of the earliest stages to infectious disease virulence is being able to move between individuals or another species (Wolfe, Dunavan, and Diamond 2007). This can only occur once the

pathogen gains the ability to adapt to the new host's system. Furthermore, direct contact with a reservoir species after their death has also led to zoonoses. Often, this is in the form butchering and preparation of host species in the creation of bush meat (Wolfe et al. 2004). The transmission of simian immunodeficiency virus (SIV) or simian foamy virus (SFV) has rapidly increased as the consumption of bush meat has also correlatively risen in the past two decades (Wolfe, Dunavan, and Diamond 2007). The case of transmission of SIV and its close counterpart HIV,-1 as well as HIV2 have resulted from contact with blood and body fluids. SIV in its native host species is not pathogenic but when transmitted to an external host, it can cause disease. Routinely, zoo workers handling simians and other non-human primates, have been infected due to close work with samples as well as close contact with live subjects. SIV and HIV are among the more worrying zoonosis, partly because they have the ability to transfer between individuals in new host species (Wolfe et al. 2004).

One of the best examples of the severity of zoonotic pathogens and continued greatest concerns is the Influenza virus. An example of this would be the influenza outbreaks of 1918 and the Hong Kong pandemic of 1968 where the influenza virus made the jump between animals and humans (Threats et al. 2005). The influenza virus, has been on record for a millennia where the earliest case being treated by Hippocrates (Lina 2008). There are two types which affect humans. Influenza A and B, both are responsible for seasonal flu epidemics. Influenza A , infects birds, swine and horses, all of which are common livestock (Threats et al. 2005). In 1918, much of the world's medical staff were deployed within the military, and little to no research and preparation for influenza allowed for the mortality to total roughly 50 million people (Oxford and Gill 2018). Of these 50 million, the pandemic seemed to have been unusually atypical in terms of fatality for those aged 20-40 with peak mortality being at age 28. This was the case in both the Spanish flu of 1918 and a few years prior in Russia's Influenza outbreak (Gagnon et al. 2013). The pandemic of 1918 virulence is firmly attributed to poor health in general in much of Europe and the US. Both regions were experiencing war time stress and overall poor living

conditions due to rapid migration, overcrowding and already weak immune systems due to rapid rise in tuberculosis (Lina 2008) . These same conditions led to the Hong Kong outbreak of H3N2 which was a mutated version of the 1957 H2N2 influenza virus. The new subtype of H3N2 was highly contagious compared to the previous 3 recorded pandemics due to this strain affecting the respiratory system (Mostafa and Pleschka 2018). This was attributed to the close living conditions between animals and humans in China (Tatem et al. 2011).

Next Generation Sequencing and MinION:

In the past 20 years, pathogen and microbial identification from host species has progressed greatly due to the advancement of high-throughput sequencing technologies; beginning with the first successful sequence of the human genome being published in 2006 (Bodmer 2013). The technology used to fully sequence the human genome was also used to begin sequencing pathogens allowing for rapid identification and vaccine creation. Next generation sequencing covers a broad range of sequencing technologies that allow researchers to quickly process millions of fragments and map them against known genomes and reference databases (Behjati and Tarpey 2013). Not only can next generation sequencing (NGS) be used for outbreak control, it can be used to prevent outbreaks. Disease population surveillance and investigation into possible spill-over prone areas can be used to control the amount of risk humans have through interactions with vector species (Gwinn, MacCannell, and Armstrong 2019). During the Ebola outbreak in 2015, a portable sequencing device (ONT MinION) and relevant equipment were quickly deployed to in West Africa, where a total of 147 Ebola samples were collected and sequenced in an eight-month deployment (Quick et al 2016). The MinION is a small device incasing a protein nanopore. Pores themselves sit within a polymer membrane which carries current at a specific voltage (Hadfield, 2012). Because nucleotides are inherently structured differently, when they pass through open pores on the protein membrane, they cause a characteristic voltage change, specific

to each base pair. This data can then be analyzed and a sequence created matching that of the sample that was run through the nanopore membrane (Hadfield, 2012). The sequenced ebola data was processed within a day, with sequences being ready for species identification within 15-60 minutes. This allowed researchers to remotely track the evolutionary rate, host adaptation and host response to vaccines almost in real time. Traditional sequencing required samples to be transported, often through tough train, to remote sequencing facilities. This resulted in valuable surveillance time lost on a virus which mutated between 16-27 times per site per year.

The importance of Rattus argentiventer

Rats have been an important organism in the spread of zoonotic diseases in history and this is still true today. Rodents are the largest order in mammalian society, and Rodentia are widespread throughout the world. Rodents since the 1980s have played a 20% part in the emergence of new human pathogens and are the source of spillover for 58% of viral zoonosis (Woolhouse M. and Gaunt E, 2007). They are also attributed as one of the supporting species to the spread of the bubonic plague decades ago. Their opportunistic feeding behavior led them on to ships and carriages that spread them far from the disease's origins (Threats et al. 2005). Rats are extremely adaptive to new environments and are opportunistic feeders that routinely are found in human populations of high density (Costa et al. 2017). Rats also frequently come into contact with humans and carry many well documented zoonotic pathogens. Rodents are increasingly becoming a food resource in South-Asia, partially from population growth (Gruber 2016).

It appears that Asian countries seem to be particularly vulnerable to zoonotic pathogens and outbreaks. This in part can be attributed to the nearly 1:1 population of humans to animals. Vietnam and Cambodia in particular have experienced recent emerging zoonotic pathogens. Due to a culture which promotes eating rodents as a food source, there is concern that South Asia may experience

continued zoonotic pathogen emergence (*Reuters* 2009). In Cambodia alone, 2 tons of wild rat are exported per day during peak season (Gruber 2016). In both countries, rats are considered a delicacy and eaten regularly (Trang et al. 2015). Vietnam in particular has become a hotspot for rodent-borne diseases where cases of renal syndrome, leptospirosis, typhus, bartonella, salmonella, campylobacter, Hantavirus, and plague have been identified (Van Cuong et al. 2015). In the late 90's, to counteract the destruction of rice crop and to keep the rodent population in check, the government of Vietnam created a series of polices that supported the consumption and trade of rats. This led to an increase in direct contact made with rats especially by people who butcher them. The government also did not mandate any protective clothing so the butchers have come into direct contact with the potentially infected fluids (Sandalij et al. 2016). Following the heavy rains and H1N1 outbreak shortly before, many governments and inhabitants have cut pork and pig production in reaction but supplementing protein in diets derived from avian and rodent origins.

It is because of this self-evident and constant threat of zoonotic emergence that microbial diversity be monitored in species that live within close proximity of humans. It is necessary to monitor microbial diversity and transfer to humans to stop zoonotic disease emergence at a local level before it becomes a pandemic. Few species are suited to fit these criteria other than rats. Because of the increased contact between humans and *Rattus argentiventer* in Vietnam, it was a perfect opportunity to study interspecies microbial diversity and to surveille potential spill over. *Rattus argentiventer* is the dominant rodent species in Vietnam, representing > 90% of the total rodent volume the Mekong Delta region. Vietnam estimates that roughly 3,300- 3,600 tons of live rats are captured annually. Accompanying this is 2,000 full time rat catchers which come into constant contact to rats and their biological material frequently. From January of 2017 to March 2017, Metabiota and VIZIONS collected blood and respiratory samples from three different sources. These sources included rat traders, rats themselves, and human controls who do not come into contact or eat rats. Human samples were

comprised of nose/throat swabs and blood plasma samples while rat samples were comprised of blood plasma, lung and digestive tissue. Total RNA extraction, reverse transcription, and sequencing was then performed for metagenomic analysis.

Objectives

The goal for this study is to identify potential viruses, monitor microbial diversity, and identify any potential spill over from rats to humans. This study will allow researchers to properly determine whether rodents, being the largest order of mammalian species and being wide spread, can be used as a precursor species to identify potential pathogens locally in Vietnam before they become an endemic or pandemic. This study will utilize 6 different samples and determine if, like influenza, zoonotic pathogens are regularly exchanged between animals and humans.

Materials and Methods

Sample Collections and Sequencing

Samples were received which were collected by the Vietnam Initiative on Zoonotic Infections (Vizions) in 2017. The samples were collected from 5 different cities within the An Giang Province. The samples were collected from the rodent meat markets in Châu Phú, An Pho, Tri Tôn, Tịnh Biên, and Long Xuyên. These samples were collected between January 19, 2017 to March 27, 2017 (*Table 1*). Samples were collected and separated into three categories, human control, rat traders and rice field rats. Both human samples were comprised of blood plasma and cheek swab samples. Rice field rat samples were comprised of blood plasma, intestinal tract swab and a respiratory track swab. Samples were consolidated, combined into each category of human control (HCP), human control swab (HCS), rat

trader plasma (RTP), rat trader swab (RTS), rat plasma (RP), and rat lung (RL), which were then transferred to Metabiota to be sequenced using Illumania miSeq at 175 bp fragments.

Row Labels	Female	Male	NA	Grand Total
An Phu	26	18	10	54
2/8/17	26	18		44
Control	13	9		22
Rat slaughter	13	9		22
3/10/17			10	10
Rat			10	10
Chau Phu	60	42	20	122
1/19/17	30	21		51
Rat slaughter	30	21		51
1/21/17	30	21		51
Control	30	21		51
3/6/17			20	20
Rat			20	20
Long Xuyen	6	2	5	13
3/27/17	6	2		8
Control	3	1		4
Rat slaughter	3	1		4
3/16/17			5	5
Rat			5	5
Tinh Bien	6	10	5	21
3/4/17	6	10		16
Control	3	5		8
Rat slaughter	3	5		8
3/10/17			5	5
Rat			5	5
Tri Ton	4	26	10	40
1/23/17	4	26		30
Control	2	13		15
Rat slaughter	2	13		15
3/10/17			10	10
Rat			10	10
Grand Total	102	98	50	250

Table 1. Data received from the Vietnam Initiative on Zoonotic Infections. Samples outline the number collected, date, and location of collection as well as gender of human subjects.

Bioinformatics

Samples from Metabiota after Illumina sequencing were received in FASTQ format. All bioinformatics programs were run using bash command line script. Trimmomatic v0.38-1 was used on both single end reads , using specification of "SE" (single ended) as well as "–phred33", and on unpaired end reads, using specification of "PE" (paired end) and "-phred33", to remove Illumina primers (Bolger, Lohse, and Usadel 2014). Unpaired end reads that contain forward and reverse reads were then concatenated utilizing FLASH to create a single output folder for paired end reads (Magoč and Salzberg 2011). Both uncombined reads and the larger combined reads following FLASH were kept.

The next step utilized bowtie2 v2.3.5.1 to build an index using the UniVec Core database. Bowtie was then used to map the reads to the database with the specifications "--very-sensitive-local", "--phred33", –x denoting assembled index location, and "—un-gz" denoting output of unaligned reads only (Langmead and Salzberg 2012). A shell script was created to run all 6 samples against the UniVec database. Taking the unaligned reads from bowtie against the UniVec database, a new index was built utilizing the human genome (hg19). Using the same specification for running the UnVec index against the reads, the six samples were run against the human genome. Following this, the *Rattus norvegicus* genome (rn4) was downloaded and its index assembled like the UniVec database. All samples but the human control is used in the following alignments. Using the same settings as the previous two runs, reads that have been aligned using both the UniVec core and the human genome were then aligned against the rn4 database and any alignments were discarded. The finial alignment was done using the *Mus musculus* (mm10) genome. The new index was built with the mm10 genome and the 4 samples which were aligned with the UniVec core, human genome, and the rn4 genome were aligned against the mm10 genome, discarding any alignments. Following alignment, the six separate libraries were separately uploaded to One Codex (Minot, Krumm, and Greenfield 2015).The samples were then

analyzed using the One Codex proprietary database to evaluate microbial content. Utilizing python scripting, the Jupyter notebook in One Codex was employed to create a PCA/PCoA, microbial heat maps, and metadata plots.

Microbial content data on One Codex was used to choose 6 significant zoonotic bacteria. Using NCBI, genome data was pulled in fasta format for *Yersinia pestis* (yp), *Yersinia enterocolitica* (ye), *Rickettsia akari (ra), Streptobacillus moniliformis (sm), Bartonella quintana (bq)*, and *Babesia (babesia).* Utilizing bowtie index compiling settings, each one of the bacteria were compiled within the running database. The six samples were then aligned against each bacterial genome utilizing a shell script to stream line the pipeline. The bowtie2 settings remain the same from the previous alignment with a modification to the "—un-gz" being changed to "—al-gz" instructing bowtie2 to keep aligned reads. These aligned reads were then analyzed using FastQC to determine quality and read length (Simon Andrews 2018).

Results

Following the trimming of the illumina primers and alignment against the different databases, the overall number of reads for each category decreased (table 2).

Flashed Samples						
	HCP	**HCS**	**RP**	**RL**	**RTP**	**RTS**
Initial Sample	80962862	101532987	121926150	178823546	14561562	144280596
Trimmomatic	80762842	101231848	111673610	176315788	14551229	144948095
Univec	80557597	101049058	111474301	176044257	14041138	144631727
HG19	1684211	33018693			1117334	52655927
MM10			46031746	44767066		
RN4			13872141	8488334		

Table 2: Total number of sequenced reads at each level of primer removal and alignment.

Microbial identification was ascertained through One Codex utilizing the One Codex 2018 database for comparison. Figure 1 outlines the comprehensive microbial diversity in a comparative layout between the six different samples. Looking at the microbial diversity on a genus level with a minimum of 1% abundance, 72% of the HCS sample was identified. Using the same graphic, HCP has an identification of 23%, 36% of RTP, 74% of RTS, 50% of RP, and 31% of RL. The human control plasma has the least identification at the genus level as compared to the other samples. Comparing diversity between the samples, it can be observed that the bacterial content in the human respiratory system (HCS and RTS) in both our human subjects was similar. The respiratory swab had a large group of prevotella bacteria. In the Human Control, Prevotella comprised 54% of the sample while the rat trader swab sample had a Prevotella content that comprised 56% of the total overall sample. The plasma samples in both human subjects' sample were also similar in diversity but not in read length. Microbial diversity in the rat's samples were also similar. The largest genus group in the rat samples are Mycoplasma. In the rat plasma sample, Mycoplasma comprised roughly 75% of the overall sample while in the rat respiratory swab mycoplasma comprised roughly 57% of the overall reads.

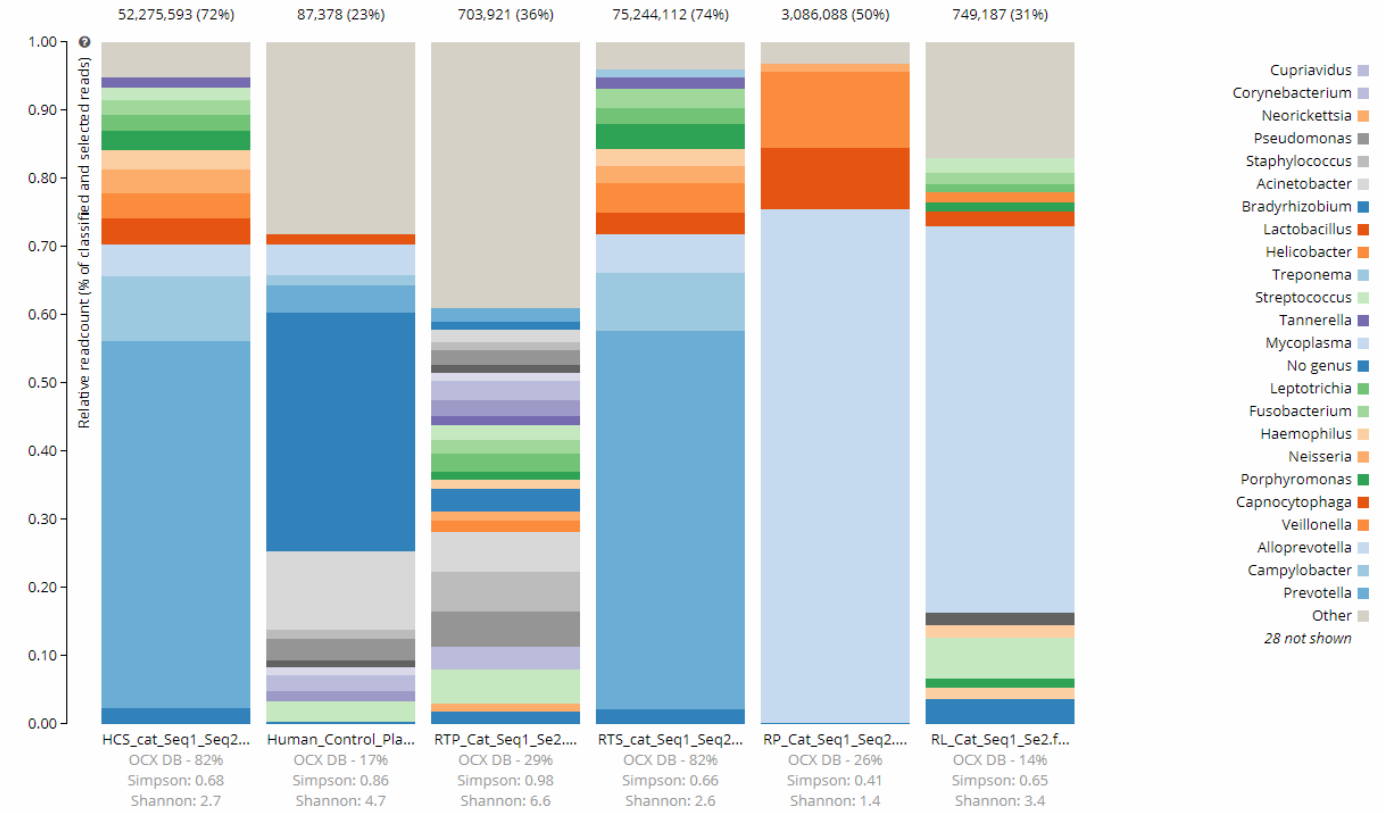

Figure 1: Comparative analysis of microbial diversity at a genus level. There is a minimum percent composition set at 1% and reads have been normalized. Figure also shows microbial diversity as compared among similar samples between species.

	Virus	Bacteria	Fungi	Arcaea	Protists	Total	Percent Identified	Simpson Diversity Index
HCS	64,175	52,190,911	20,209	160	138	88,648,002	82%	0.68
HCP	353	84,028	2,641	212	144	2,214,061	17%	0.86
RTS	180,925	75,048,795	13,981	110	301	123,664,564	82%	0.66
RTP	10,600	618,702	72,309	129	2181	6,771,551	29%	0.98
RP	1,365	3,079,306	5,211	38	168	23,581,754	26%	0.41
RL	4,130	721,069	22,535	47	1406	17,284,408	14%	0.65

Table 3 : Comparative break down of virus, bacteria, fungi, archaea, and protists utilizing the RefSeq database on One Codex. Also included is the statistical calculation of diversity utilizing the Simpson index. Index value closer to 1 represents dominance by a singular species.

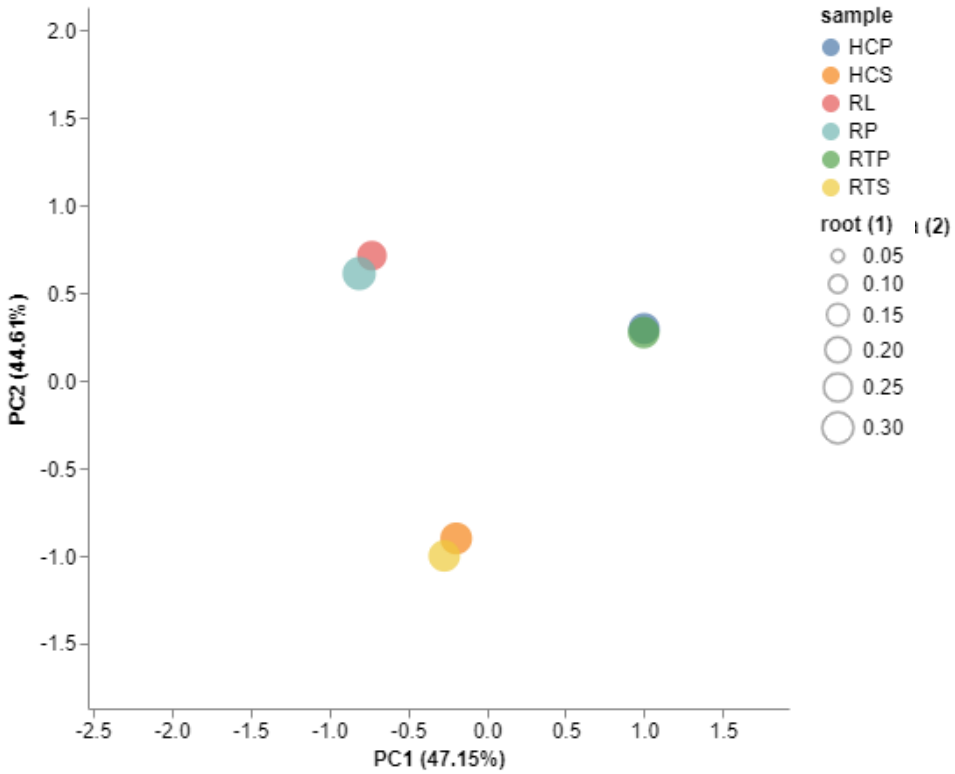

Figure 2: Unweighted principal components analysis (PCA). Plots microbial diversity between the different samples.

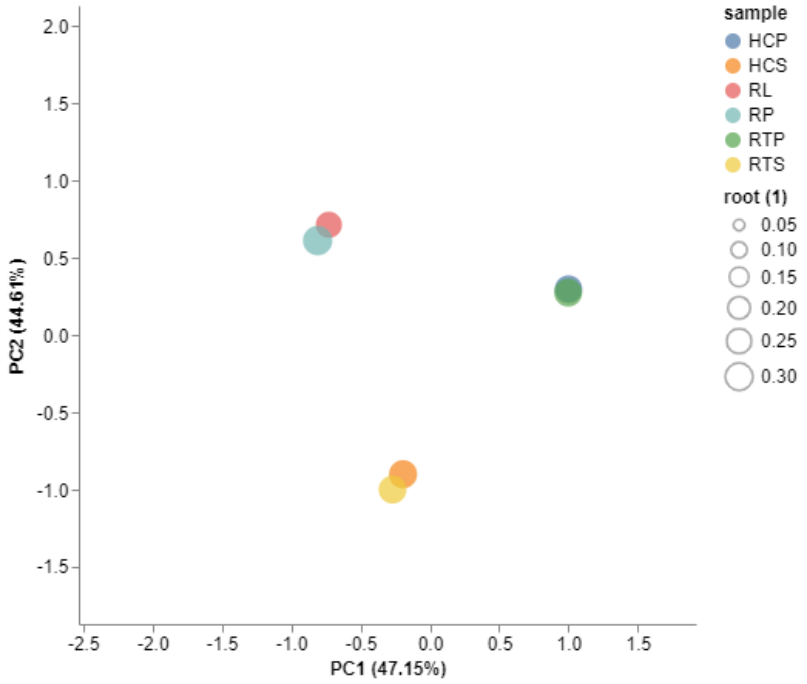

Figure 3: Weighted principal components analysis (PcoA). Weighted statistical analysis of microbial diversity. Clustering and distance from plots represent closeness in diversity.

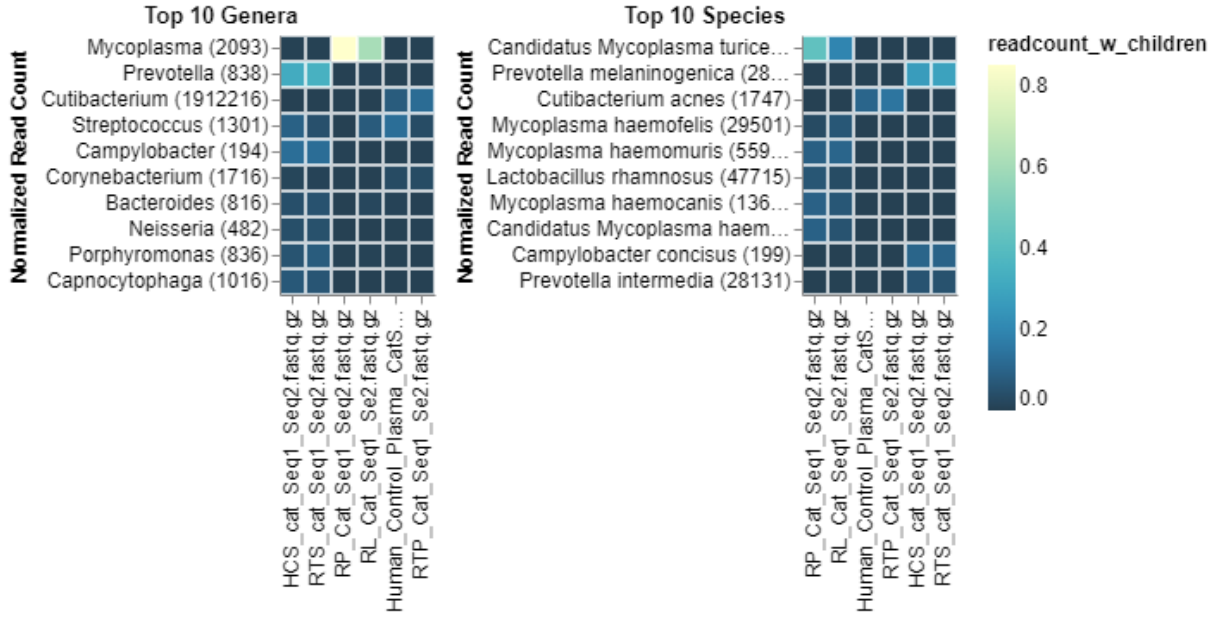

Figure 4: Heat map comparison of the top 10 genera and species based on the six principle samples.

Using the analytical data from One Codex, a *table 4* could be filled out. Using data from NCBI, a list of common rodent borne pathogens was compiled to be compared against the 6 different samples. The read counts are fairly low but diseases such rat bite fever and relapsing fever were prevalent in human samples. A small amount of hits also did come back for *Yersinia pestis*, also known as plague. There was a total of 107 reads for the rat trader swab and 68 total reads in the human control swab.

Diseases From Rodents	**Genus/Species Name**	**RTS**		**RTP**		**HCS**		**HCP**		**RL**		**RP**	
Directly Transmitted by Rodents		***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***
Hantavirus Pulmonary Syndrome (HPS)	Hantavirus												
Hemorrhagic Fever with Renal Syndrome	Bunyavirus												
Lassa Fever	Lassa virus												
Leptospirosis	Leptospira	x	254	x	3	x	194	x	6	x	23	x	23
Lymphocytic Chorio-meningitis (LCM)	Lymphocytic choriomeningitis virus												
Omsk Hemorrhagic Fever	Omsk hemorrhagic fever virus												
Plague	Yersinia pestis	x	107	x	1	x	68	x	2	x	6	x	1
Yersiniosis	Yersinia enterocolitica			x	21					x	5		
Rat-Bite Fever	Streptobacillus moniliformis	x	9349			x	4213	x	6			x	60
	Spirillum minor			x	1								
Salmonellosis	Salmonella bongori			x	1			x	7			x	1
	Salmonella enterica	x	19	x	78	x	16	x	1			x	3
Arenaviruses	Arenavirus												
Tularemia	Francisella tularensis	x	21	x	1	x	25	x	1				
Encephalomyocarditis fever virus (ECMV)	Encephalomyocarditis virus									x	225	x	15
Diseases From Rodents	**Genus/Species Name**	**RTS**		**RTP**		**HCS**		**HCP**		**RL**		**RP**	
Indirectly transmitted by Rodents		***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***	***Identifiec***	***# Reads***
Babesiosis	Babesia	x	42	x	52	x	20	x	6	x	7	x	7
Colorado Tick Fever	Coltivirus												
Cutaneous Leishmaniasis	Leishmania					x	2	x	7	x	1	x	1
Human Granulocytic Anaplasmosis	Anaplasma phagocytophilum	x	18	x	1	x	2	x	7	x	4	x	4
La Crosse Encephalitis	La Crosse Virus												
Lyme Disease	Borrelia burgdorferi	x	33										
Murine Typhus	Rickettsia typhi			x	5			x	2			x	1
Powassan Virus	Powassan virus												
Scrub Typhus	Orientia tsutsugamushi	x	98					x	5				
Rickettsialpox	Rickettsia akari					x	153	x	1				
Relapsing Fever	Rickettsia prowazekii, Bartonella quintana,	x	4594							x	292	x	1
Rocky Mountain Spotted Fever	Rickettsia rickettsii							x	8				
Sylvatic Typhus	Rickettsia prowazekii												
West Nile Virus	West Nile Virus												
Taxoplasmosis	Toxoplasma gondii	x	45	x	701	x	29			x	29	x	29

Table 4: Common and prevalent rodent borne diseases compared against data analyzed by One Codex using the One Codex database. Comparatively significant read count on rite bite fever and relapsing fever.

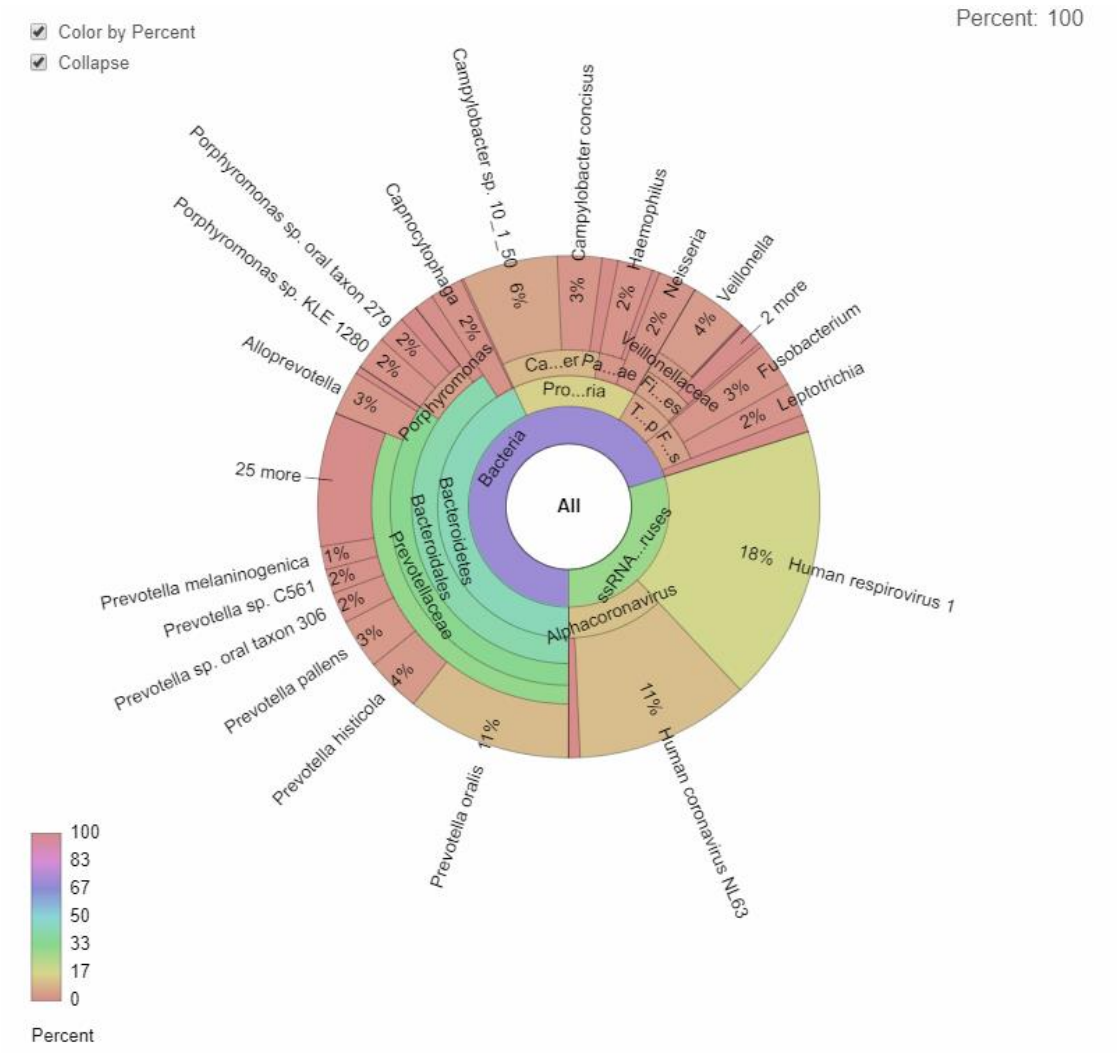

Figure 5: Rat trader swab analysis of overall sample abundance. The top three human respirovirus 1, human coronavirus NL63, and prevotella oralis are found in relative high abundance.

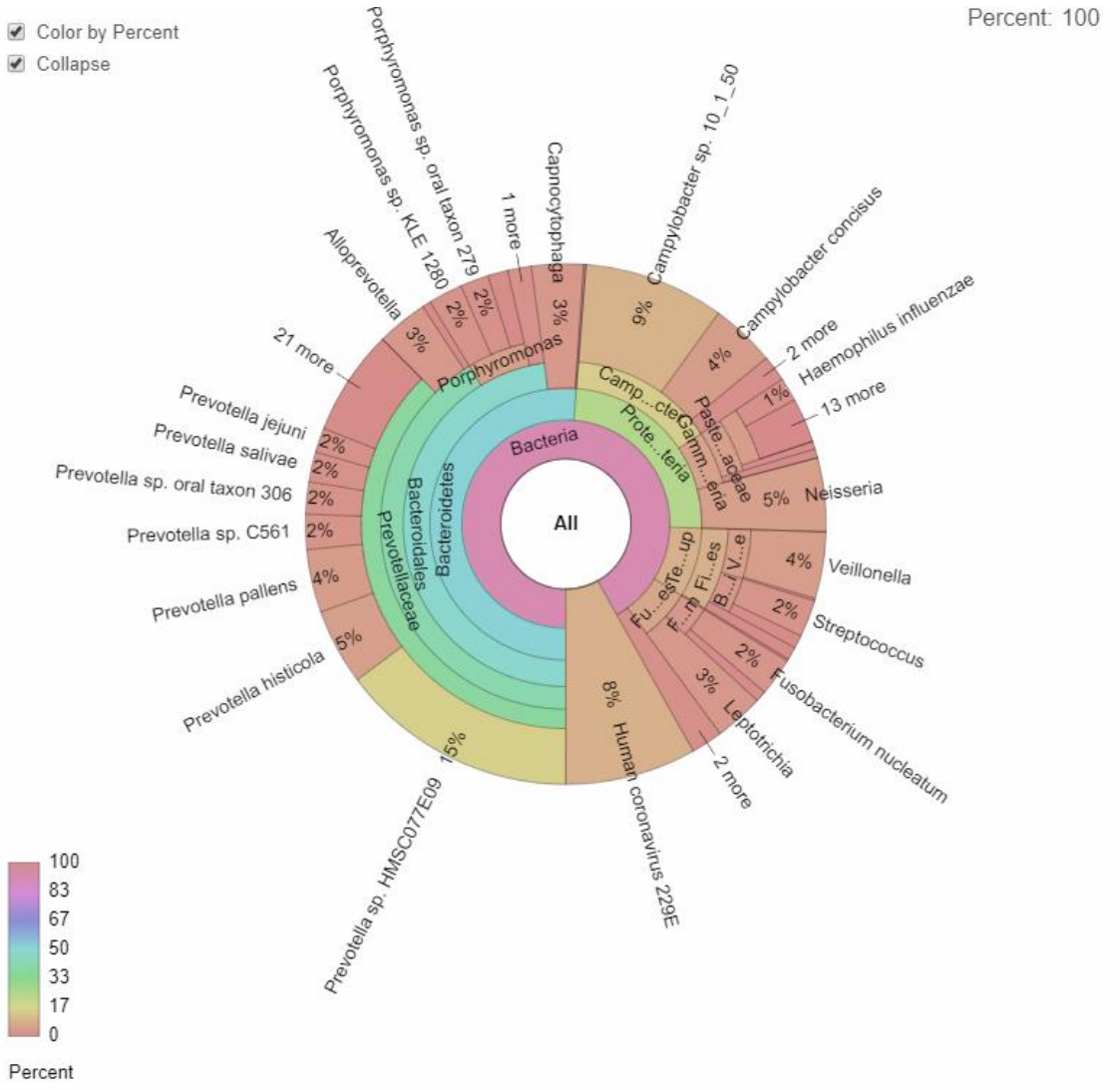

Figure 6: Human control swab sample analysis of overall abundance. The top three human respirovirus 1, human coronavirus NL63, and prevotella oralis are found in relative high abundance.

Discussion

There were three main goals for this study. First, to identify potential viruses in the system, second, to monitor the microbial diversity in each sample and third, establish whether or not there was potential spill over between the two species in review, humans and rats. Accomplishing these goals allows us to test the viability of microbial surveillance on specific species and location for assessment.

Sample collection and initial processing

Samples were collected by the VIZIONS program in Vietnam by a team over the course of 3 months from various open air butcher markets around Vietnam. Samples were divided roughly in half based on gender between humans. The rats themselves were not sexed when samples were collected. The Rat trader samples came from butchers or people who traded in rats. These were the best candidates to establish spill over and potential virus contraction. The control sample pool was also divided in half based on gender and contained samples from people with no contact with rats. This information was ultimately voluntary and cannot be verified from the participants. These samples were then pooled based on the sample grouping to create 3 pools of samples rather than 250 separate ones fromm humans. This allowed the pools to be evaluated as a whole rather on a case by case basis. The samples were then shipped to Metabiota to be sequenced using Illumania miSeq at 175 bp fragments. Illumania miSeq provided less error in reads as compared to on site sequencing such as nanopore sequencing (Tyler et al. 2018).

Bioinformatics

The samples were pooled and transferred in the large fasta files. These files contained two sets of data. One set contained fasta files that contained reads going in one directions while the other set of files contained the pooled data going in forwards and reverse as genetic data goes both ways. All the bioinformatics tools were used within the command line structure in Linux or Mac OsX. Trimmomatic

was chosen as it was able to remove illumania primer from the ends of the reads fairly quickly as compared to other primer removal tools on the market. Depending on the number of fragments, some samples such as RP, as observed in table 2, had a fairly large number of reads removed. These separate reads were then re-pooled in their main categories as the study was only looking for microbial content. This was done using FLASH which takes multiple files and looks for overlap between them. These overlapped regions are only kept once and any non-overlapping regions are also kept as they signify differences in the files. The bulk of the read length removal occurred with the use of Bowtie2. Bowtie2, an alignment program, was chosen due to its memory intensive usage as opposed to its CPU intensive use. This allowed it to be used on many desktops with the bear minimum 8 GB of needed ram (Langmead and Salzberg 2012).

Identification

Pathogen identification was possible through utilizing the raw data provided and processed by OneCodex. Table 3 lists the number of hits established for each of the taxa, where swab samples from both the human control and the rat trader, had a very large number of bacteria present. This is likely due to the respiratory system being common place for bacteria as opposed to the blood plasma. Conversely, rat blood plasma contained much more in terms of bacterial content as compared to the respiratory swab samples. As rats are the vectors in question, it stands to reason many of the viruses have aggregated within the host species. The fungi content was higher as well in the rat respiratory swab as well compared to the levels in blood plasma, this maybe a result of the feeding habits of rice field rats being reliant on plants and insects (Tristiani, Murakami, and Watanabe 2003). This pattern of increased fungal species identification carries over to the respiratory samples of both the other humans. One of the more worrying identifications comes from both respiratory swabs for humans. Both human samples contained a relatively large abundance of human respirovirus 1, human coronavirus NL63, and

prevotella oralis. The human coronavirus and respirovirus, respiratory viruses, are linked to lengthy respiratory tract infections leading to high fever and a sever coughs. In the case of corona virus, it is usually not associated with many fatal diseases but has been identified as being responsible for the outbreak of severe acute respiratory syndrome (Abdul-Rasool and Fielding 2010). It also has been seen to infect primarily children causing bronchiolitis or croup in cases of infection in the lower respiratory tract (Abdul-Rasool and Fielding 2010). Respirovirus 1, is part of the family named parainfluenza viruses (HPIV). Respirovirus , like human coronavirus NL63 , infects the lower respiratory tract and most often in children. It can also be found in a number of primates, rodents and other host species. In experimental conditions, infection have been possible in hamsters, guinea pigs, and adult ferrets (Henrickson 2003). Further, fatalities were observed in new born ferrets once exposed to the human version of the virus (Mascoli et al. 1976). Utilizing figures 1 and 4, the comparative analysis points to composition of the samples to be wholly of regular species of bacteria which are routinely found in the location of each sample. Figure 4 standardizes all 6 samples based on abundance and creates a heat map. The top 10 bacteria, in terms of abundance, across the 6 samples are regular oral or gut bacteria. Bacteria such as mycoplasma, prevotella, and cutibacterium are found in large abundance in all 6. The stand out among these 6 was by far the human control blood plasma. Either one or multiple participants had hepatitis C. Using the One Codex database, the abundance of this virus was roughly 55% compared to the whole sample.

Diversity Analysis

Using figures 2, 3 and table 3, diversity statistics can be established for the samples based on composition. Table 3 contains the Simpsons index value which calculates the diversity of a sample based on abundance and species richness (Morris et al. 2014). The closer the value is to 1, the more dominant one species is in the samples composition. Using this statistical unit to analyze the samples, the trend is

established showing blood plasma samples from both humans and rats were dominated by a few different bacteria as compared to the respiratory swabs. The largest domineering bacteria by far was in the Human Control where the blood plasma contained hepatitis C which made up 55% of the sample, which led to its Simpsons score being .86. Furthermore, utilizing python coding in the Jupyter notebook built into One Codex, figures 3 and 4 were created using the principal components analysis (PCoA). The PCoA was used to analyze closeness between sample diversity. In figure 4, the weighted analysis, samples grouped together based on the composition. The conclusion to be drawn from figure 4 is that, both the rat trader blood plasma and human control plasma, was similar in their composition. The same can be said for both human respiratory swabs. This trend was observed in both the weighted and unweighted analysis of the diversity. Furthermore, the rat microbial diversity grouped both the blood and respiratory swab together signifying no real difference in microbial content. All 6 samples grouped up into 3 separate points. All three of these points were equally spaced from one another signifying no one group of samples being more similar to another.

Pathogen transference and potential spill over

One of the central goals of this project and that of VIZIONS, was to access interspecies microbial transference and establish rice field rats as a candidate as a surveillance target to this end. Analyzing data processed by OneCodex and referencing it against known diseases from rodents, table 5 was constructed to focus efforts. The two samples of interest were Rat Traders and the rats that were within their vicinity. The traders samples contained trace amounts of *Leptospira*, *Yersenia pestis*, *Streptobacillus moniliformis, Salmonella bongori, Babesia, Bartonella quintana ,* and *Toxoplama gondiI*. All of these bacteria cause disease which can be traced back to rodents. Of the six samples, the rat trader swab contained the highest number of reads on each of the species. The second highest conglomeration of bacteria seems to be in the human controls themselves. Though these were human's

presumably that have no contact with rodents, they contained a disproportionate amount of bacteria originating from rodents. The rodents contained very little in terms of microbial reads; all being >292 number of hits. Lastly, there was a small number of reads that pointed to the presence of *Yersinia pestis*, which commonly known as plague in the human control swab. This was concluded as being a false positive because there was a negligible number pf reads present in the blood plasma sample.

Conclusion

Utilizing data collected by the VIZIONs program and analyzing samples, a number of conclusions can be made. The overall goal of this project was to determine whether or not microbial transference was occurring and to what effect. Though a sample size of 250 individuals were collected from and processed, the data determines that its inconclusive based on a number of things. Looking at reads post OneCodex analysis, they are extremely low. There are for sure enough microbial reads from known diseases to conclude that rodent borne diseases are entering the human populous. It cannot, however, be directly tied to the consumption and handling of the rodents in question. This conclusion is supported by the high level of rodent borne disease present even within the human control subjects. If human control subjects are affected, there could be external factors facilitating the transference. One of the largest number of reads were attributed to *Streptobacillus moniliformisI*. This bacterium is known to cause and present itself in the form of rate bite fever, which when untreated can be lethal. It is also an emerging infectious disease being spread in many regions (Graves and Janda 2001) .But aside from being transmitted via a bite or scratch from a rodent, it can be transmitted by contamination of foods and surfaces with the bacteria (Elliott 2007). The samples were collected within an open air meat market and in regions where unsafe handling of foods as well malnourishment is wide spread (Nguyen et al. 2017). These conditions could facilitate the transference of many of these rodents borne bacteria into humans.

Going forward

The next step forward on this project would be to further analyze the samples that were provided. To gauge whether or not the specific bacteria that cause rodent pathogens are the same ones in the humans, further alignment data would be needed to ascertain whether the same species of the disease is being spread between humans and rats. From there a definitive conclusion could be made on the mode of transference. Further, more samples need to be taken with more attention to detail. The health of subjects as well as specific data on rodent interactions should be recorded to correctly classify human controls as just that, control samples. It is important to stop infectious disease from where they begin and continuously monitor their diversity within their biomes. Continuous monitoring would allow for rapid response to spill over which could result in pandemics as with Ebola and Influenza.

Bibliography

Abdul-Rasool, Sahar, and Burtram C Fielding. 2010. “Understanding Human Coronavirus HCoV-NL63.” *The Open Virology Journal* 4 (May): 76–84. https://doi.org/10.2174/1874357901004010076.

Ahn, Matae, Danielle E. Anderson, Qian Zhang, Chee Wah Tan, Beng Lee Lim, Katarina Luko, Ming Wen, et al. 2019. “Dampened NLRP3-Mediated Inflammation in Bats and Implications for a Special Viral Reservoir Host.” *Nature Microbiology* 4 (5): 789. https://doi.org/10.1038/s41564-019-0371-3.

Allen, Toph, Kris A. Murray, Carlos Zambrana-Torrelio, Stephen S. Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin J. Olival, and Peter Daszak. 2017. “Global Hotspots and Correlates of Emerging Zoonotic Diseases.” *Nature Communications* 8 (1): 1124. https://doi.org/10.1038/s41467-017-00923-8.

Ashford, R.W. 2003. “When Is a Reservoir Not a Reservoir?” *Emerging Infectious Diseases* 9 (11): 1495–96. https://doi.org/10.3201/eid0911.030088.

Behjati, Sam, and Patrick S Tarpey. 2013. “What Is next Generation Sequencing?” *Archives of Disease in Childhood. Education and Practice Edition* 98 (6): 236–38. https://doi.org/10.1136/archdischild-2013-304340.

Bodmer, W. 2013. “Human Genome Project.” In *Brenner’s Encyclopedia of Genetics (Second Edition)*, edited by Stanley Maloy and Kelly Hughes, 552–54. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-374984-0.00746-4.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics (Oxford, England)* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Cunningham, Andrew A., Peter Daszak, and James L. N. Wood. 2017. "One Health, Emerging Infectious Diseases and Wildlife: Two Decades of Progress?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1725). https://doi.org/10.1098/rstb.2016.0167.

Elliott, Sean P. 2007. "Rat Bite Fever and Streptobacillus Moniliformis." *Clinical Microbiology Reviews* 20 (1): 13–22. https://doi.org/10.1128/CMR.00016-06.

Gagnon, Alain, Matthew S. Miller, Stacey A. Hallman, Robert Bourbeau, D. Ann Herring, David JD Earn, and Joaquín Madrenas. 2013. "Age-Specific Mortality During the 1918 Influenza Pandemic: Unravelling the Mystery of High Young Adult Mortality." *PLOS ONE* 8 (8): e69586. https://doi.org/10.1371/journal.pone.0069586.

Graves, Margot H., and J. Michael Janda. 2001. "Rat-Bite Fever (Streptobacillus Moniliformis): A Potential Emerging Disease." *International Journal of Infectious Diseases* 5 (3): 151–54. https://doi.org/10.1016/S1201-9712(01)90090-6.

Gruber, Karl. 2016. "Rodent Meat – a Sustainable Way to Feed the World?" *EMBO Reports* 17 (5): 630–33. https://doi.org/10.15252/embr.201642306.

Gwinn, Marta, Duncan MacCannell, and Gregory L. Armstrong. 2019. "Next-Generation Sequencing of Infectious Pathogens." *JAMA* 321 (9): 893–94. https://doi.org/10.1001/jama.2018.21669.

Henrickson, Kelly J. 2003. "Parainfluenza Viruses." *Clinical Microbiology Reviews* 16 (2): 242–64. https://doi.org/10.1128/cmr.16.2.242-264.2003.

Jones, Kate E., Nikkita G. Patel, Marc A. Levy, Adam Storeygard, Deborah Balk, John L. Gittleman, and Peter Daszak. 2008. "Global Trends in Emerging Infectious Diseases." *Nature* 451 (7181): 990–93. https://doi.org/10.1038/nature06536.

Kuriakose, Teneema, and Thirumala-Devi Kanneganti. 2017. "Regulation and Functions of NLRP3 Inflammasome during Influenza Virus Infection." *Molecular Immunology* 86: 56–64. https://doi.org/10.1016/j.molimm.2017.01.023.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lina, Bruno. 2008. “History of Influenza Pandemics.” In *Paleomicrobiology: Past Human Infections*, edited by Didier Raoult and Michel Drancourt, 199–211. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75855-6_12.

Luby, Stephen P., Mahmudur Rahman, M. Jahangir Hossain, Lauren S. Blum, M. Mushtaq Husain, Emily Gurley, Rasheda Khan, et al. 2006. “Foodborne Transmission of Nipah Virus, Bangladesh.” *Emerging Infectious Diseases* 12 (12): 1888–94. https://doi.org/10.3201/eid1212.060732.

Magoč, Tanja, and Steven L. Salzberg. 2011. “FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies.” *Bioinformatics* 27 (21): 2957–63. https://doi.org/10.1093/bioinformatics/btr507.

Mandl, Judith N., Rafi Ahmed, Luis B. Barreiro, Peter Daszak, Jonathan H. Epstein, Herbert W. Virgin, and Mark B. Feinberg. 2015. “Reservoir Host Immune Responses to Emerging Zoonotic Viruses.” *Cell* 160 (1): 20–35. https://doi.org/10.1016/j.cell.2014.12.003.

Mascoli, CC, TA Gower, FA Capilupo, and DP Metzgar. 1976. “Further Studies on the Neonatal Ferret Model of Infection and Immunity to and Attenuation of Human Parainfluenza Viruses.” *Developments in Biological Standardization* 33: 384–90.

Minot, Samuel S., Niklas Krumm, and Nicholas B. Greenfield. 2015. “One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification.” *BioRxiv*, September, 027607. https://doi.org/10.1101/027607.

Morris, E Kathryn, Tancredi Caruso, François Buscot, Markus Fischer, Christine Hancock, Tanja S Maier, Torsten Meiners, et al. 2014. “Choosing and Using Diversity Indices: Insights for Ecological Applications from the German Biodiversity Exploratories.” *Ecology and Evolution* 4 (18): 3514–24. https://doi.org/10.1002/ece3.1155.

Mostafa, Ahmed, and Stephan Pleschka. 2018. “Influenza H3N2 Vaccines: Recent Challenges.” *Trends in Microbiology* 26 (2): 87–89. https://doi.org/10.1016/j.tim.2017.12.003.

Narrod, Clare, Jakob Zinsstag, and Marites Tiongco. 2012. “A One Health Framework for Estimating the Economic Costs of Zoonotic Diseases on Society.” *Ecohealth* 9 (2): 150–62. https://doi.org/10.1007/s10393-012-0747-9.

Neiderud, Carl-Johan. 2015. “How Urbanization Affects the Epidemiology of Emerging Infectious Diseases.” *Infection Ecology & Epidemiology* 5 (June). https://doi.org/10.3402/iee.v5.27060.

Nguyen, Thuong V., Quang D. Pham, Quoc K. Do, Tai T. Diep, Hung C. Phan, Thang V. Ho, Hong T. Do, Lan T. Phan, and Huu N. Tran. 2017. “Cholera Returns to Southern Vietnam in an Outbreak Associated with Consuming Unsafe Water through Iced Tea: A Matched Case-Control Study.” *PLoS Neglected Tropical Diseases* 11 (4). https://doi.org/10.1371/journal.pntd.0005490.

Oxford, John S, and Douglas Gill. 2018. “Unanswered Questions about the 1918 Influenza Pandemic: Origin, Pathology, and the Virus Itself.” *The Lancet Infectious Diseases* 18 (11): e348–54. https://doi.org/10.1016/S1473-3099(18)30359-1.

Reuters. 2009. “Flu Fears, Rains Buoy Cambodia Rat Exports to Vietnam,” May 18, 2009. https://www.reuters.com/article/us-cambodia-rats-idUSTRE54H0RR20090518.

Simon Andrews. 2018. *FastQC* (version 0.11.8). Windows. English. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Threats, Institute of Medicine (US) Forum on Microbial, Stacey L. Knobler, Alison Mack, Adel Mahmoud, and Stanley M. Lemon. 2005. *The Story of Influenza*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK22148/.

Tristiani, Harsiwi, Okimasa Murakami, and Hiroyuki Watanabe. 2003. “Ranging and Nesting Behavior of the Ricefield Rat Rattus Argentiventer (Rodentia: Muridae) in West Java, Indonesia.” *Journal of Mammalogy* 84 (4): 1228–36. https://doi.org/10.1644/BEM-024.

Tyler, Andrea D., Laura Mataseje, Chantel J. Urfano, Lisa Schmidt, Kym S. Antonation, Michael R. Mulvey, and Cindi R. Corbett. 2018. “Evaluation of Oxford Nanopore’s MinION Sequencing Device for Microbial Whole Genome Sequencing Applications.” *Scientific Reports* 8 (1): 10931. https://doi.org/10.1038/s41598-018-29334-5.

Valen, Leigh Van. 1973. “A New Evolutionary Law.” *Evolution* 27 (1): 27–35. https://doi.org/10.1111/j.1558-5646.1973.tb05914.x.

West, Elena H., William R. Henry, William Goldenberg, and M. Zachariah Peery. 2016. “Influence of Food Subsidies on the Foraging Ecology of a Synanthropic Species in Protected Areas.” *Ecosphere* 7 (10): e01532. https://doi.org/10.1002/ecs2.1532.

Wolfe, Nathan D., Claire Panosian Dunavan, and Jared Diamond. 2007. “Origins of Major Human Infectious Diseases.” *Nature* 447 (7142): 279–83. https://doi.org/10.1038/nature05775.

Wolfe, Nathan D, William M Switzer, Jean K Carr, Vinod B Bhullar, Vedapuri Shanmugam, Ubald Tamoufe, A Tassy Prosser, et al. 2004. “Naturally Acquired Simian Retrovirus Infections in Central African Hunters.” *The Lancet* 363 (9413): 932–37. https://doi.org/10.1016/S0140-6736(04)15787-5.

Acknowledgements

I would first like to thank the number of organizations such the Vietnam Initiative on Zoonotic Infections and Metabiota for both collecting and sequencing the samples. Without the hard work of the ground team in Vietnam collecting samples, this project would not be possible.

I would also like to thank Dr. Harris for giving me the opportunity to complete a worthwhile project. This research project, more than any other, allowed me to realize my dream as a researcher and to finally make a real impact. You were able to teach me new techniques and skills that will help me throughout my career. For that I cannot thank Dr. Harris enough.

Lastly, I would like to thank my friends for supporting me in the rigorous trails of being a scientist, my family for allowing me the opportunity to grow, the biology faculty for guidance as well as purpose, and lastly the Harris Lab for creating a space that we could call our own

.