AFRL-PH-WP-THE-2023-0001

COMPARING THE PERFORMANCE OF A TARGETED PULL-DOWN ASSAY TO SHOTGUN SEQUENCING FOR IMPROVING RESPIRATORY INFECTIOUS DISEASE SURVEILLANCE

Monica Christian

Oak Ridge Institute of Science Education

Report Date

November 2023

Final Report

Controlled by: Department of the Air Force, DCPH-D
CUI Category: Unclassifed
Distribution/Dissemination Control:
POC: Mary Allen, mary.allen.2@us.af.mil
Date of Determination: 10 October 2023

DCPH-Dayton
2510 Fifth St., Bldg. 840

Wright-Patterson AFB, OH 45433-7913

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-PH-WP-THE-2023-0001 has been reviewed and is approved for publication in accordance with assigned distribution statement.

AGANS.RICHARD.THOMAS.1513855440 Digitally signed by AGANS.RICHARD.THOMAS.1513855440 Date: 2023.11.28 14:12:56 -05'00'

RICHARD T. AGANS, PhD
Senior Research Biology Scientist
DCPH-D

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED 2021-2023	
1 NOV 23	DTIC	START DATE Sept 2021	END DATE JUN 23

4. TITLE AND SUBTITLE

COMPARING THE PERFORMANCE OF A TARGETED PULL-DOWN ASSAY TO SHOTGUN SEQUENCING FOR IMPROVING RESPIRATORY INFECTIOUS DISEASE SURVEILLANCE

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)

Christian, Monica, R, Chapleau, Richard, R, Starr, Clarise, R

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

DCPH-Dayton
2510 Fifth St., Bldg. 840

Wright-Patterson AFB, OH 45433-7913

8. PERFORMING ORGANIZATION REPORT NUMBER

AFRL-PH-WP-THE-2023-0001

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DCPH-Dayton 2510 Fifth St., Bldg. 840 Wright-Patterson AFB, OH 45433-7913	DCPH-D	AFRL-PH-WP-THE-2023-0001

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Clearance number is AFRL-2023-5052. Approved for public release: distribution is unlimited. 10 October 2023. Other requests shall be referred to DCPH-D/OEC, 2510 Fifth St., Bldg. 840, Wright-Patterson AFB, OH 45433-7913

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Current surveillance focuses on well characterized pathogens such as influenza. Since 2000, there have been multiple outbreaks of respiratory disease. These outbreaks have demonstrated the need for robust and collaborative global efforts to identify, monitor, and contain novel respiratory viruses.

This study aims to improve the ability of public health agencies to monitor and respond to respiratory disease outbreaks. Using five respiratory pathogens, this study compares a molecular capture technology from Twist Biosciences to the shotgun sequencing approach of whole transcriptome amplification (WTA, Qiagen) with the goal of determining which method is most effective using cost, usability, and sequencing quality metrics for evaluation.

Twist, though more expensive, had a 92.1% positive identification of targets on successful sequencing runs with greater depth and breadth of coverage. WTA failed to sequence and identify targets except Human Adenovirus 7, proving that Twist is more reliable and efficient in this study.

15. SUBJECT TERMS

Next-generation sequencing, respiratory virus, hybrid capture, bioinformatics, novel pathogen, pathogen detection, molecular capture, pathogen identification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	Unclassified	64

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER
Dr. Richard Agans	(312) 798-2971

Contents

LIST OF TABLES

LIST OF FIGURES

1 SUMMARY

Current surveillance focuses on well characterized pathogens such as influenza. Since 2000, there have been multiple outbreaks of respiratory disease. These outbreaks have demonstrated the need for robust and collaborative global efforts to identify, monitor, and contain novel respiratory viruses.

This study aims to improve the ability of public health agencies to monitor and respond to respiratory disease outbreaks. Using five respiratory pathogens, this study compares a molecular capture technology from Twist Biosciences to the shotgun sequencing approach of whole transcriptome amplification (WTA, Qiagen) with the goal of determining which method is most effective using cost, usability, and sequencing quality metrics for evaluation.

Twist, though more expensive, had a 92.1% positive identification of targets on successful sequencing runs with greater depth and breadth of coverage. WTA failed to sequence and identify targets except Human Adenovirus 7, proving that Twist is more reliable and efficient in this study.

2 INTRODUCTION

2.1 Background

2.1.1 History of Large-Scale Respiratory Virus Outbreaks

There have been numerous outbreaks of respiratory viruses with dramatic and deleterious effects on human history. Even recent history dating back to as close as the 20th century reveals the pervasive effects these viruses have on humankind, and the effect that globalization has on infectious disease spread. The influenza A H1N1 pandemic of 1918-1919 coincided with the United States' entry into the First World War. When the chlorine-gas filled dust of the war had settled, an estimated 16 million people were dead from combat. Conversely, when H1N1 finally disappeared, more than 50 million had perished, 25% of the U.S. population had been infected, and the average U.S. life expectancy dropped by 12 years (27). This novel virus was indiscriminate; unlike traditional seasonal flu, the young and fit were just as likely to become infected as the elderly and/or immunocompromised. Additionally, those who were infected presented with extraordinarily severe symptoms, conceivably due to immunological naivete and increased inflammatory response to the novel pathogen.

While diseases such as typhus (28), cholera (30), and polio (29) reached epidemic levels, there was not another respiratory pandemic until 1957 (35). Global use of planes, trains, and automobiles all contributed to the rapid spread of a new H2N2 strain of influenza A (36). First identified in April of 1957 in Hong Kong, this strain of influenza reached all corners of the globe by October of the same year (36). The resulting 1957-58 pandemic cost British taxpayers more than 10 million pounds and resulted in a significant economic stall due to the closing of businesses (36). All told, the death toll was around 1 million worldwide.

H2N2 took over as the prevailing seasonal flu until 1968, when H3N2 began circulating (25). A product of antigenic shift, H3N2 contained two new genes from an influenza A virus found in birds (35) while retaining the neuraminidase from the previous pandemic (31). H3N2 was largely spread through the movements of different militaries around the world, also aided by the global use of transportation (31). Also like the previous pandemic, the number of deaths is estimated to be between 1 and 3 million (31).

In addition to influenza A outbreaks, variants of the human coronavirus have also influenced the recent course of human history. For example, both Severe Acute Respiratory Syndrome (SARS) and Middle Eastern Respiratory Syndrome (MERS) appeared in the early 21st century and have shaped the global pandemic response. The first suspected case of SARS was in November of 2002 and was classified as an atypical pneumonia (74). It wasn't until March 12, 2003 that a global alert was issued by the World Health Organization (WHO) for cases of severe and atypical pneumonia and the full sequence of the virus was released April 14 (71). By early July, the WHO declared that the outbreak was contained (71). Global morbidity was more than 8000, with a 9.6% case fatality rate (77). MERS was identified in 2012 and, to date, there have been slightly more than 2200 cases reported worldwide (5,7). Devastatingly, the case fatality rate of MERS is 35% (5).

These deadly influenza and coronavirus outbreaks have demonstrated the need for robust and collaborative global efforts to identify, monitor, and contain novel respiratory viruses.

2.1.2 **Respiratory outbreak control through surveillance and testing**

Rigorous surveillance systems are in place for monitoring some infectious diseases such as influenza, food borne illnesses, and bioterrorism threats (25). The Global Influenza Surveillance and Response System (GISRS) was formed in 1952 by the WHO. It consists of more than 150 laboratories in 114 countries (24). Additionally, the United States Centers for Disease Control and Prevention (CDC) sequence more than 7000 influenza samples annually (10), contributing these sequences to the Global Initiative on Sharing Avian Influenza Data (GISAID). Sequences submitted to GISAID are used by GISRS to help determine which strains should be included in the yearly influenza vaccine (11).

The CDC also operates the National Notifiable Diseases Surveillance System (NNDSS) (23). The NNDSS utilizes case surveillance to gather not only disease information, but community demographics and clinical outcomes. Approximately 3000 public health departments submit data to the CDC through the NNDSS to be used to monitor and track 120 different conditions. Such conditions include infectious diseases such as *Haemophillus influenzae* and vancomycin-resistant *Staphylococcus Aureus*, as well as non-infectious conditions like lead poisoning and cancer (23).

However, despite these surveillance systems, respiratory infections continue to emerge and cause considerable disruption to the global population. The 2009 H1N1 Influenza outbreak resulted in more than 60 million cases in the United States, with a fatality rate of 8-12 times that of seasonal flu (6). MERS is still a potential threat, with smaller outbreaks still occurring in the Middle East (7). The greatest example of the failure to identify and track respiratory pathogens with pandemic potential is the COVID-19 pandemic that began in late 2019 and has resulted in more than 624 million cases and 6.5 million deaths at the time of this paper's publication (2).

2.2 Approach

2.2.1 Molecular testing for respiratory viruses

There are three primary methods for identifying respiratory viruses through molecular techniques: polymerase chain reaction (PCR), antigen testing, and genomic sequencing. PCR and antigen testing are the only methods widely-available for clinical diagnosis of respiratory infectious diseases (80), both of which require previous knowledge of the genetic makeup and/or molecular epitopes for each particular pathogen. Genomic sequencing is becoming increasingly more valuable for pathogen surveillance, including tracking of site-specific outbreaks and drug resistance markers (80).

PCR testing is based on amplifying and detecting unique segments of the viral genome. Amplification is done through use of molecularly engineered DNA polymerases and 18-30 bp primers (81) situated on the 5' and 3' sides of the target oligomer. Detection is done visually via electrophoresis or by using fluorescent probes designed to bind to the target amplicon. As such, PCR based identification requires equipment and personnel that are commonly found in molecular biology laboratories; however there are two primary limitations with employing PCR testing in a surveillance capacity.

The first limitation is that PCR tests are susceptible to genetic mutations interfering with the abilities of primers and/or probes to hybridize with their targets. The recent COVID-19 pandemic serves as an example of this limitation. Initially, the CDC released an assay targeting three locations on the viral genome (N1, N2, and N3). In order for a sample to be confirmed as "positive" for the causative virus (SARS-CoV-2), all three viral targets were required to amplify with a crossover threshold (Ct) less than 40, meaning, the fluorescent signal rose above the level of background noise (threshold) in under 40 PCR cycles. As early as April 2020, one common

variant was being reported from samples that prevented N3 from amplifying: an A28095T substitution that later became the classic signature of the Alpha (B.1.1.7) variant (9). This PCR-dropout required the third viral target to be re-evaluated and removed almost immediately after the assay was released.

A second limitation with PCR-based surveillance programs is that they will often miss new and emergent pathogens. For example, a retrospective study by Chapleau et al. (2020) testing respiratory samples from early in the COVID-19 pandemic, demonstrated that pathogens can be present in otherwise clinically negative patient samples weeks preceding an official outbreak (8). More than 7000 samples from 86 medical treatment facilities were analyzed using the CDC's SARS-CoV-2 qPCR detection assay (without the N3 target previously described as non-functional). The authors found evidence of SARS-CoV-2 positive samples more than 11 days prior to the officially reported detection. They also found that approximately 2% of samples from early in the pandemic were misdiagnosed as negative for the virus, even after the CDC had expanded the clinical case definition guidelines for testing (50,51). This study clearly demonstrated a significant gap in the ability to surveil for rapidly emerging respiratory viruses.

In contrast to PCR's reliance upon a pathogen's genetic information, antigen testing is a type of immunoassay that detects proteins of interest (52). The applications for antigen detection systems are very broad; ranging from at-home pregnancy tests to complex fluorescence-linked antibody assays for pathogens in stool (52). Antigen testing is based on antibody-antigen interactions, where at least one of the antibodies involved is conjugated with a reporter molecule. The reporter molecule is then detected by an analyzer, or as is the case in point of care lateral flow assays, by the human eye. While antigen testing is not as sensitive as PCR, it is often cheaper per test and requires less skilled labor to perform the testing than for performing PCR

testing (52, 53). Most antigen testing is considered “low complexity” by the College of American Pathologists and thus can be performed by individuals with minimal training (57).

Antigen testing is still considered the “gold standard” for detection of many pathogens, including rabies virus (54) and *Giardia lamblia* (56), though it is also at the mercy of viral mutation. As with PCR testing, the major limitations of antigen testing are susceptibility to antigenic drift (mutations causing the epitopes to no longer be recognized by the assay) and inability to detect novel pathogens. Newer variants have escaped detection by at-home antigen tests at an alarming frequency. For example, variants of SARS-CoV-2 containing a D399N missense mutation evaded detection on the Quidel SARS Antigen FIA while still being detectable on other platforms (10).

Sequencing of pathogens has the potential to address the major limitations of PCR and antigen testing. Similarly to PCR, genetic sequencing amplifies the nucleic acids present in a patient sample and assesses for the presence of molecular signatures of specific pathogens. As a read-out technology, sequencing is compatible with a wide range of different sample detection approaches. One approach to circumvent mutations interfering with detection is the use of less stringent primers than PCR that could amplify a broad-spectrum of pathogens. Alternatively, highly specific primers could also be included for the targeted enrichment of selected pathogens. Finally, targeted hybridization capture technologies could enrich the genetic material of pathogens in the sample. Once the sample is prepared, the genetic material is run through an analyzer and the sequences of genetic code obtained from the sample are compared to a database of genomic sequences. The compatibility of sequencing with numerous sample preparation methods including random amplification and targeted capture makes this technology an ideal solution to the problems of mutational variation and emergent pathogens.

For example, sequencing of samples from symptomatic patients that were diagnosed as negative for respiratory viruses by conventional molecular techniques led to the discovery of SARS-CoV-2 (1) and the development of diagnostic testing for the novel pathogen (2)(3). Therefore, it is reasonable to assume that other pathogens or variants of known pathogens could be discovered and tracked through surveillance of symptomatic, negative-testing patient samples. Rapid discovery and accurate tracking of these data can provide critical information to public health organizations about potential outbreaks, as well as possibly predict which pathogens could potentially cause other pandemics. Furthermore, sequencing-based tests may be favorable for wide-spread infectious disease surveillance as a single test kit and analysis device could perform the equivalent of the hundreds of PCR tests simultaneously. Such an innovation would fundamentally alter public health surveillance and prevent pandemics at a reduced overall cost compared to the current testing system.

2.2.2 Goals and objectives of this study

The primary goal driving this study is to improve the ability of the global public health community to monitor and respond rapidly to the next pandemic and prevent deaths. To achieve this goal, this study focuses on evaluating a newly released method of capturing large quantities of molecular sequence data from respiratory viruses potentially in a patient sample and comparing its performance to the "gold standard" for broad-spectrum sequencing. Specifically, this study compares a targeted molecular capture technology released by Twist Biosciences to the shotgun sequencing approach of whole transcriptome amplification using random primers manufactured by Qiagen. The objective was to compare ease of use, per sample cost, and technical sequencing ability. Ease of use was determined qualitatively based upon time at the

bench and requirements for advanced training and/or equipment, costs were solely determined based upon consumables and not personnel labor, and technical performance was determined using quality control metrics inherent to sequencing (e.g., depth and breadth of coverage).

2.3 Scope

2.3.1 Natural diversity in respiratory pathogens

Another important challenge to on-going public health surveillance is that symptoms of viral respiratory infections in adults remain relatively similar across different viruses. However, the viruses vary greatly in pathogenicity, physical attributes, and biochemistry (37). The respiratory viruses selected for this endeavor (Table 1) were chosen for not only their impact on adult human health, but also the variations in their nucleic acid type and arrangement. This will allow for the testing of nucleic acid configurations most commonly found in respiratory viruses of particular concern in adult populations. As this is a pilot study, it was important to focus on a few select targets to refine the process.

Table 1: Viruses and genomic details selected for this study

Virus	Nucleic Acid Type	Genomic Arrangement
Adenovirus 7 (hAdv)	dsDNA	Continuous
Coronavirus OC43 (OC43)	(+) ssRNA	Continuous
Enterovirus D68 (D68)	(+) ssRNA	Continuous
Influenza A	(-) ssRNA	Segmented
Influenza B	(-) ssRNA	Segmented

Pathogens such as respiratory syncytial virus and human metapneumovirus are responsible for a number of critical pediatric infections and present a significant burden on healthcare resources, however, they were not included in this study at this time as the focus was on pathogens that typically infect healthy adult populations. Additionally, most respiratory viruses do not have a dsRNA genome and therefore an example pathogen (i.e., rotavirus) was not included in the current evaluation.

2.3.2 Influenza A

Influenza A belongs to the *Alphainfluenzavirus* genus of the *Orthomyxoviridae.* It is an enveloped, negative-strand RNA virus. The genome is segmented, meaning that each portion of the genome is packaged in separate ribonucleoprotein complexes (84) rather than a single continuous strand of nucleic acids. This arrangement is not dissimilar to that of chromosomes and provides ample opportunity for genetic exchange among individuals belonging to the same genus. Influenza A is further characterized by two critical viral envelope glycoproteins; hemagglutinin (HA) and neuraminidase (NA). There are 198 possible distinct combinations of these two proteins, however, only a select number have been responsible for sustained human transmission (22)(23). As described previously and in Table 2, Influenza A has been responsible for four pandemics and epidemics (35), including the 1918 “Spanish Flu” pandemic responsible for more than 50 million deaths (27).

Table 2: Notable pandemics and epidemics caused by Influenza A

Year	Subtype	Estimated Deaths
1918 (“Spanish Flu”)	H1N1	>50 million
1957-58	H2N2	1 million
1968-69	H3N2	1-3 million
2009	H1N1	105-395 thousand

With respect to a global molecular surveillance program, the segmented genome of Influenza allows for the possibility of antigenic shift. This shift occurs when the Influenza A virus acquires HA and NA genomic segments, which are the primary targets for current molecular surveillance approaches (57) from another subtype. Antigenic shift that happens in cells co-infected with human and animal influenza A can provide a pandemic-level pathogen due to the immunological naivete of the host population (22).

From a molecular perspective, the influenza A strains that cause the most common human infections have HA proteins which preferentially bind to α2-6-linked sialic acids. In contrast, Influenza A viruses infecting avian sources have HA proteins known to bind to α2-3–linked sialic acids, however, mutations in the binding pocket can allow for binding to α2-6 linkages as well, thus conferring the ability to infect humans hosts (41,58). The human upper respiratory tract epithelium is rich in α2-6-linked sialic acids, whereas α2-3 linkages are more abundant in the lower respiratory tract, type II pneumocytes, and alveolar macrophages (41).

This difference in infection location between HA strains possibly explains the difference in severity and morbidity of the H1N1 pandemics when compared to others. Since the strain differences could lead to vastly different infection severity, an on-going molecular surveillance program for Influenza A is critical for ensuring public health.

2.3.3 **Influenza B**

Influenza B is the sole member of *Betainfluenzavirus* genus of the *Orthomyxoviridae.* Structurally and biochemically it is similar to Influenza A, also being an enveloped virus with a segmented RNA genome. The outer envelope is also covered in HA and NA glycoproteins responsible for receptor binding and invasion of hose cells (22,23). While influenza A and B are from different *Orthomyxoviridae* genuses, the family similarities permit both to infect human hosts.

The differences in Influenza B may be why it is not known to cause pandemics on the scale and/or severity of those caused by Influenza A. The rate of mutation for Influenza B is much slower. Influenza B polymerase exhibits a lower rate of mutation, higher fidelity, and there is a significant transmission bottleneck within hosts (39). Additionally, it is only known to infect humans & seals. Because mutations are less common and there are fewer chances for antigenic shift with only two known hosts, the potential for influenza B to cause a pandemic is virtually non-existent (39)

There are two main strains of Influenza B, B/Victoria/2/1988 and B/Yamagata/16/1988, and historically both strains co-circulate every flu season (40). The HA protein of Influenza B binds to both α2-3 and α2-6 linked sialic acids (41). Heavy glycosylation of the virus often restricts infection to the upper airways, with variants with less glycosylation producing alveolar

lung injury similar to that of pandemic Influenza A strains (42). Because of their frequency of circulation and their ability to bind sialic acids in all parts of the human airway, both strains of Influenza B are included in the quadrivalent vaccine (40).

Though it is predicted that Influenza B lacks pandemic potential, it does contribute to morbidity numbers during flu seasons. From 2010-2011, Influenza B was responsible for 36% of pediatric flu deaths, yet only comprised 26% of influenza infections (42). Even more alarming is that 49% of these children had no preexisting conditions that would predispose them to a severe case of respiratory illness (42). Interestingly, the B/Yamagata/16/1988 lineage has seemingly disappeared from circulation with no sequence-confirmed community acquired cases since March 2020 (40, 55). The reason for the seasonal disappearance is unknown, but may have been related to the coronavirus pandemic infection control measures (55). Regardless of pandemic potential, it remains important to maintain annual surveillance of Influenza B strains for maintaining public health.

2.3.4 **Adenovirus 7**

Adenoviruses are linear, double stranded DNA viruses enclosed in an icosahedral protein capsid (43). Members of the *Adenoviridae* family infect a wide range of vertebrate hosts (44). Adenoviruses capable of infecting humans belong to the genera *Mastadenovirus*, of which there are 88 distinct viruses that belong to seven subgroups; A-G (45). Aside from respiratory illness, Adenoviruses are also indicated in gastrointestinal disease, conjunctivitis, tonsillitis, and meningitis (46). Adenovirus 7 (hAdv7), a member of the B subgroup, is responsible for a number of larger and deadly outbreaks within certain populations (47). While most hAdv7 cases are mild, the virus spreads rapidly in congested environments and warranted introduction of a vaccine for use in the U.S. military in 1971 (48).

Most adenovirus cases are no more serious than a common cold, however, certain subtypes such as hAdv7 and hAdv3 have been implicated in deadly outbreaks. For example, in January-March of 2017 an outbreak of adenovirus in a rehabilitation facility resulted in three deaths and four hospitalizations (61). While the total numbers are low, this represented a 5% rate of hospitalization and a 4% fatality rate among the infected. The close confines and atypically severe presentation of symptoms lead to a delay in identification of the causative agent (61). Patients can shed Adenovirus for weeks post infection(62). The virus is also rugged and can survive on dry surfaces for up to 7 days (62). These factors, in addition to the potential for high fatality rate, make Adenovirus 7 an important subject of surveillance.

2.3.5 **Enterovirus D68**

Enteroviruses belong to the family *Picornaviridae*, which also includes notable members such as respiratory syncytial virus; poliovirus; and hand, foot, and mouth disease (63). Structurally, enteroviruses have a small, positive sense, single stranded RNA genome enclosed in a non-enveloped icosahedral protein capsid (63). Most enteroviruses infect the gastrointestinal tract, however, enterovirus D68 (EV-D68) exhibits more similarities with rhinoviruses, another genus in the *Picorniviridae* family, and has been found in respiratory sampling indicating tissue tropism for the upper respiratory tract (64-65).

Once internalized from the upper respiratory tract, Enteroviruses are capable of crossing the blood-brain barrier (63). The majority of enterovirus infections are mild (66), however EV-D68 is implicated as a causative agent of Acute Flaccid Myelitis (AFM) (63). In 2014, an outbreak in the United States and Canada resulted in 1153 confirmed cases of EV-D68, of which 10.4% developed AFM and other neurological symptoms (67). Sequencing analysis of enterovirus samples from around the world showed divergence into four distinct clades (A-D)

with almost all of the samples from the 2014 outbreak belonging to the B1 clade (67). Comparatively speaking, AFM caused by EV-D68 is rare (66). The concern and need for surveillance lies in the ability of a respiratory virus to invade and infect mammalian nervous tissue. The potentially devastating clinical outcomes of infection with EV-D68 warrant strain-specific surveillance efforts.

2.3.6 Coronavirus OC43

All coronaviruses are enveloped, single-stranded, positive-sense RNA (73). The viral envelope is covered with spike proteins that participate in receptor binding and cellular entry (71). Prior to the 2003 SARS pandemic, coronaviruses were a bigger issue for the veterinary and livestock farming communities than for human health. Outbreaks of bovine and porcine coronavirus result in massive herd losses through death of young stock and culling of infected individuals (78,79). Human coronavirus OC43 (hCoV-OC43) and bovine coronavirus diverged from a common ancestor around the year 1890 (71).

Human coronavirus OC43 usually manifests as a “common cold”, with more severe cases presenting in children, elderly adults, and immunocompromised individuals (73). Nonetheless, it is still a significant burden on healthcare resources (73). Even more important is that hCoV-OC43 is an acceptable alternative model for SARS and SARS-CoV-2 research, as well as representing historical coronavirus zoonoses with a well-established lineage (72). Working with either SARS or SARS-CoV-2 requires Biosafety Level (BSL)-3 facilities, meaning they require extra precautions such as unidirectional airflow and specialized decontamination systems (72). In contrast, hCoV-OC43 can be worked with under BSL-2 conditions available in most microbiology and molecular biology labs, provided they are equipped with class IIa biological

safety cabinets. Its utility as a model pathogen, history of zoonotic transmission, and close relation to previous pandemic coronaviruses warrant use of hCoV-OC43 for this study.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Nucleic acid preparation

Phosphate buffered saline (PBS) (Gibco) was sterilized using a 0.22µm fitted syringe filter. Dilutions of American Type Tissue Culture (ATCC) frozen live-virus stocks were made in the sterile PBS at concentrations of 5000, 200, and 10 plaque forming units (pfu)/10µl. The dilutions were repeated so that each library preparation technique could be repeated three times per viral dilution. Calculations for the dilutions were done using the specific lot numbers found in each virus' Certificate of Analysis (Table 3). Per the library preparation kit manufacturers' literature, these three levels are expected to span the limits of detection for both WTA and the Twist Comprehensive Viral panel (6,7). Nucleic acid extraction was done via Maxwell RSC Viral Total Nucleic Acid Purification Kit (Promega). The resulting nucleic acids were eluted in 50 µl nuclease-free water (Promega). This automated, bead-based nucleic acid extraction technique was used for the entirety of the retrospective COVID-19 study (8) referenced in the background. Extracted nucleic acid was frozen once at -80°C and held at this temperature until library prep.

Table 3: ATCC lot numbers

Virus	ATCC	Lot #
Influenza A (H3N2)	VR-544	57899071
Influenza B	VR-1931	70020870
Coronavirus OC43	VR-1558	70034234
Adenovirus 7	VR-7	61527421
Enterovirus D68	VR-1824	63188595

3.2 Twist Total Comprehensive Viral Panel

All steps were completed per manufacturer's recommendations (31), except where otherwise specified.

3.2.1 Production of indexed libraries

Each dilution aliquot underwent cDNA conversion using Random Primer 6 and ProtoScript First Strand cDNA Synthesis Kit (New England Biosciences). All samples underwent this step, as in the case of an unknown illness we would not know what nucleic acid arrangement the pathogen would possess. Second strand synthesis was completed using the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module (New England Biosciences, Ipswitch, MA). Directional second strand synthesis is used when preserving strand orientation is important when analyzing the subsequent sequences. Non-directional second strand

synthesis is less expensive and has fewer steps. Directionality was not critical to analyzing the sequencing data from the viral targets, so the non-directional kit was used.

Using reagents from the Twist library preparation kit, the resulting DNA was purified with magnetic beads to remove free primers and fragmented DNA. Enzymatic fragmentation, end repair, and dA-tailing were completed in a single step. The dA-tailed DNA then underwent adapter ligation and a 0.8x magnetic bead clean. This cleaning step ensured that only fragments > 200bp were selected to proceed to the next step. The adapted library was then indexed and amplified using Twist Universal Adapter System UDI Primers from plate C. Another magnetic bead cleaning step (1x) removed any excess primers and ensured that fragments were larger than 200bp.

3.2.2 **Pull-down selection**

Samples were pooled into 8-plex hybridization reactions, with the exception of samples Blank 4 and 5. Hybridization pools were air-dried at room temperature under cover to prevent contamination. Dried pools were resuspended in blocker solution and universal blockers to decrease non-specific binding of sample DNA by hybridization probes (12). Probe solution and sample were heated independently to 105°C, then equilibrated at room temperature. Probe solution and hybridization enhancer reagent were added to the resuspended pools and incubated at 70°C for 16 hours. Recovery of probe-bound DNA fragments began by washing the magnetic streptavidin beads provided in the kit. The DNA samples were added directly from their tubes in the thermocycler to the pre-washed beads. Samples were incubated on a rocker at room temperature for 30 minutes, then washed, and eluted with nuclease-free water. This bead slurry was then placed on ice. NEB Q5 Hot Start High Fidelity and primers provided in the kit were

used to amplify the selected DNA fragments. A final bead clean at 1x was completed and the samples stored at -20°C until sequencing.

3.3 **Qiagen Whole Transcriptome Amplification (WTA)**

All steps were completed per manufacturer's recommendations, except where otherwise specified (69).

3.3.1 **Amplification of purified RNA**

Purified nucleic acid was denatured to remove any secondary structures, and gDNA removed using the provided gDNA Wipeout buffer. While the exact contents of the gDNA buffer are proprietary, it is reasonable to assume it is a combinations of DNAses that preferentially target gDNA. Sample then underwent reverse transcription using Quantiscript RT enzyme mix and random primer mix. RT-PCR was done without the oligo dT primers, to avoid disproportionately amplifying human mRNA rather than viral RNA. Because the initial samples (ATCC stock virus diluted in filter-sterilized PBS) are acellular, the amount of mRNA from the viral template may escape detection. Samples then were ligated, meaning, the individual amplicons produced in the RT step were ligated end-to-end in random order. This provides the template for the next step, multiple strand displacement amplification (MDA). MDA utilizes a highly sensitive polymerase and random hexamer primers to amplify cDNA. For this kit, REPLI-g SensiPhi DNA polymerase is used for the MDA. This step is an isothermal reaction that results in high molecular weight, branched dsDNA ranging in length from 2kb to 70kb.

3.3.2 **Nextera XT library prep**

Samples were tagmented using the Nextera transposome. The transposome enzymatically fragments the genome into 400-1200bp pieces and ligates a universal adapter in one step. After

neutralization, the sample is indexed with two unique indices according to the Illumina Index Adapters Pooling Guide (20). These indexes are added to the adapters using limited cycles of PCR. Samples were cleaned using the Illumina Purification Beads provided and library size verified to be approximately 300-400bp using the Applied Biosciences Tape Station and D1000 reagents.

3.4 **Sequencing**

All samples were sequenced on a single Illumina MiSeq V3 600 cycle cartridge. 150 cycle, paired-end reads were selected as run parameters. Libraries were normalized to 4nM and denatured using 1N NaOH. Loading of pooled libraries and initiation of sequencing were completed per manufacturer's protocol. TruSeq compatible index sequences were manually entered per sample. Sequencing was successful with 20% PhiX for cluster generation.

3.5 **Bioinformatics**

3.5.1 **Set up and Retrieval of Data**

Paired-end FASTQ files were trimmed for adapters and quality controlled (QC) using TrimmomaticPE (16). Bases with Q scores lower than 30 were removed from both the 3' and 5' ends. Sequences shorter than 36 bp after quality trimming were discarded, as they were likely the result of primer-only reads (13). Discarded reads were saved in a separate file, in case less stringent QC parameters were needed. Trimmed files were visually QC'd using FastQC (14). Files were copied, with one copy remaining as a FASTQ file, and the other converted to FASTA using SeqTK (15). This is due to differing file type requirements for the programs used in the proceeding steps. Scripts are provided in Appendix C and D.

3.5.2 **NCBI Blastn**

Refseq Viral database files were retrieved from https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/ downloaded 5 JUL 22 using wget. A local database was created using the makeblastdb application (16). Local, automated BLASTn search was completed on all trimmed files using massblast.sh (Appendix A). An E value of 0.001 was used to ensure that variant sequences could be matched to their closest genetic neighbor (16). Each paired read file was treated as an individual search, with the results concatenated at a later step. Output for BLASTn was set at one result per read, with the identity, percent match, E value, and bitscore included. The resulting text file was analyzed in LibreOffice Calc and ranked by bitscore, then E value. BLASTn output files were also evaluated for number of unique hits per viral identity using hitcount.sh (Appendix B). Paired end files were manually combined to get a total hit count of the sample. These hit counts were then used to calculate Limits of Blank and Limits of Detection for each library preparation method (49).

3.5.3 **Database curation**

The BLASTn results were used to form a smaller, more manageable database that was used for alignment. Using a local instance of a curated database greatly decreased the processing power and storage required. This also eliminated the "coin flip" phenomenon of incorrectly calling equal scoring reads in conserved regions of closely related sequences. Additionally, Bowtie2 requires a local database in order to align and score reads. Read identities with a bitscore > 90% of bitscore max were selected to make this secondary database as a new file. Database files were retrieved as needed from https://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral using wget. Bowtie2 (17) was retrieved and installed from its source and indexed databases were created using the bowtie2-build application. If there were multiple database entries (hAdv7,

Influenza A, etc) then they were concatenated prior to indexing and database creation per requirements of the program (17).

3.5.4 **Bowtie2 Alignment**

The paired end files were aligned against the curated database using end-to-end alignment. Sequences were considered "aligned" if they passed the minimum score threshold, defined as -0.6 + -0.6L, with L corresponding to read length. Unaligned sequences were discarded and aligned reads were output in a SAM file (17).

3.6 **Sequencing analysis**

3.6.1 **Calculation of detection limits**

Limits of Blank and Detection (LOB and LOD, respectively) were calculated using Blastn results (70). LOB was determined by calculating the total number of reads in the PBS blank samples per library preparation method, followed by averaging the number of reads. The following equation was used for the final LOB calculation (70):

LOB = Average of the blank + 1.645 * Standard Deviation of the blank

LOD was calculated using Blastn results from the lowest successful concentrations of each virus. The following equation was used for the final LOD calculation (70):

LOD = LOB + 1.645 * Standard Deviation of the Low Analyte Concentration Samples

Using 1.645 standard deviations in the equation allows for capture of the central 90% of the sample population in a standard normal population distribution (85)

3.6.2 **Calculation of Depth and Breadth of Coverage**

Depth and breadth of coverage were calculated for each sample file using Samtools coverage function (19). Breadth of coverage is represented as percent of genome covered for a particular database entry at 1x depth. Depth of coverage is presented as average coverage at over the length of the genome.

3.7 **Additional Quantitative PCR (qPCR)**

qPCR testing using an ribonuclease P (RNAseP) primer and probe mixture (Integrated DNA Technologies)(90) and SuperScript™ III One-Step RT-PCR System with Platinum™ *Taq* DNA Polymerase (ThermoFisher) was performed on the following samples using manufacturer's recommendations:

1. Extracted nucleic acids from hAdv-7, Influenza A, Influenza B, hCoV-OC43, and Enterovirus D68.
2. Post-WTA samples before Nextera XT library prep for hAdv-7, Influenza A, Influenza B, hCoV-OC43, Enterovirus D68, and Blank.
3. Pool of indexed WTA samples prior to sequencing
4. Post amplification, Pre pull-down Twist samples for hAdv-7, hCoV-OC43, and Enterovirus D68.
5. Post pull-down pool of indexed Twist samples

The original Twist samples for Influenza A and Influenza B were unavailable for additional testing. The filter sterilized PBS used to make dilutions was also unavailable.

This testing was completed in order to explain high levels of background in certain sample types and to identify sources of contamination found during the bioinformatic analysis. Samples were run for 45 cycles on a QuantStudio 5 Dx and analysis was completed using the in vitro diagnostic (IVD) software partition.

4 RESULTS AND DISCUSSION

4.1 Cost, time, and ease of use

Cost and ease of use are important metrics that are often absent in evaluations of methods. Easier methods require less training, but often are more expensive and can be less customizable to the task you wish to accomplish. Consequently, highly complex methods can be difficult to repeat and come with higher time and labor costs. In comparing the Twist library preparation method with WTA, there is a stark difference in the time required to complete each method and in the cost per sample.

The kits themselves are both easy to use. and are designed to be used by individuals with experience in basic laboratory techniques. Both kits require use of standard equipment (e.g. thermocyclers, magnets, and lo-bind nuclease-free plastic consumables) commonly found in molecular labs or readily available from laboratory suppliers (31, 68).

Cost analysis calculations were done assuming 96 samples for each kit and a minimum 1 million reads per sample on sequencing. The price per sample goes up if there are fewer samples, as the sequencing reagents are single use only. Assuming that all samples sequence on the first try, and that there are no repeats, the Twist kit is 26% more expensive per sample than WTA. However, there was a failure to generate meaningful data of the first 45 WTA samples. The same WTA library pool was repeated using MiSeq reagents of a different lot number, and also failed to generate sequences that passed the Q30 filter. This failure is unlike the initial Twist failure, as inability to generate clusters is a sequencer-based issue rather than a potential problem with the wet lab process and reads not passing the quality filter. In order to further address the sequencing failures, one replicate of each target at the high and medium concentrations, as well as one blank, were prepared for sequencing as previously described.

As part of the cost analysis, the higher rate of failures inherent to the WTA process must be considered. The final 10 samples were run with 100% success, so the failure rate of WTA overall was 81% (45/55), meaning a conservative estimate of doubling the costs per sample is reasonable for the cost analysis.

Table 4: Summary of cost and ease of use comparisons

	Twist	**WTA**
Time	3 days	1 day
Cost/sample	$108.20	$85.10
Ease of use	Easy	Easy
Failed runs	7/45	45/55[a]
Failed to ID	3/38[c]	8/10[b]

[a] An additional 10 samples were prepared and sequenced in order to have some data for analysis. These numbers do not include blanks.

[b]These are the 10 samples that successfully sequenced after the initial failures

[c]Twist successfully identified at least one replicate for each target at each concentration.

4.2 Database description

The RefSeq Viral database from NCBI has complete genomes from more than 11,600 viruses at a total size of 465,259 kb. In comparison, the cumulative curated databases have 10 viruses totaling 146.7 kb (Table 5).

Table 5: Characteristics of the curated database and RefSeq reference database

	Curated db	
Sample	No. of Entries	Size
Adenovirus 7	5	55kb
Coronavirus OC43	1	9.6kb
Enterovirus D68	1	2.5kb
Influenza A H3N2	2	9.1kb
Influenza B	1	4.8kb

4.3 Sequencing analysis

4.3.1 Limits of Blank

PBS blanks sequenced using the Twist kit averaged less than 2 hits per blank, with only 5 organisms represented across blanks (Figure 1). Based upon an LOB of 2 reads, the minimum theoretical LOD for the Twist method is 3 reads (Figure 2). The WTA blank had a total of

40,055 reads on more than 230 organisms (Figure 1). There were an average of 168.3 hits per organism, with an LOB calculated at 2959 (Figure 2)(70).

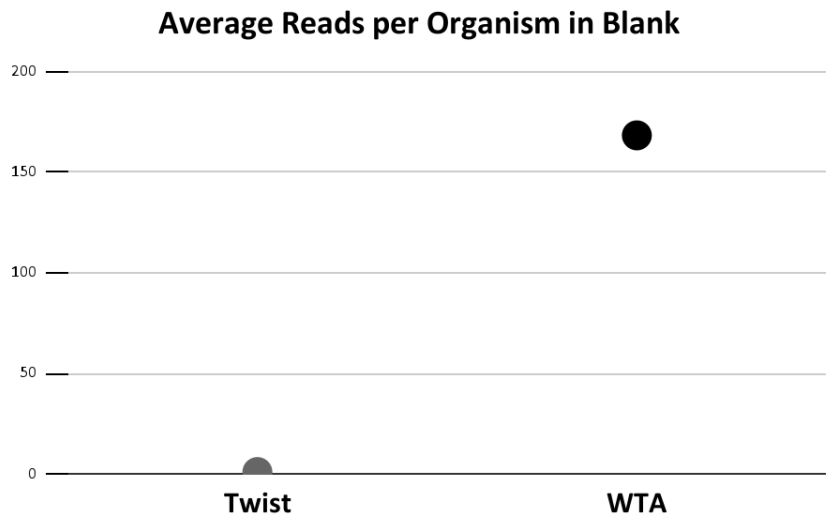

Figure 1 : Comparison of average reads per organism between Twist and WTA.

Twist averaged 1.2 ± 0.4 reads and WTA averaged 168.3 ± 1692.4 reads. Values are presented as reads per unique identification.

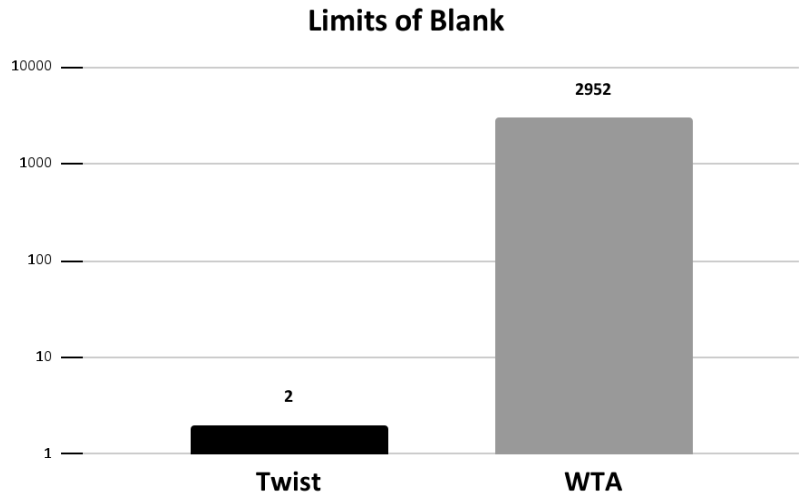

Figure 2 Limit of blank comparison between Twist and WTA

. The LOB from Twist (dark gray bar) was over 3 log lower than the LOB from WTA (black bar). Values presented as total reads.

4.3.2 **Limits of Detection**

As mentioned previously, the samples from the WTA preparation failed to produce the correct identification for spiked samples. The exception was hAdv7 (Figure 3), which had a lowest read count of 14397 reads from 200 pfu, or 72 reads per pfu. In contrast, the worst performing Twist assay was 120 reads from 10 pfu or 12 reads per pfu (influenza B). The best performing targets using Twist were influenza A and enterovirus D68 at 11 reads or 1.1 reads per pfu. Considering the theoretical lowest read count determined above to determine a positive detection was 3 reads, it is possible that this method could detect 2 pfu in a 200 µl sample (10 pfu/mL)(70).

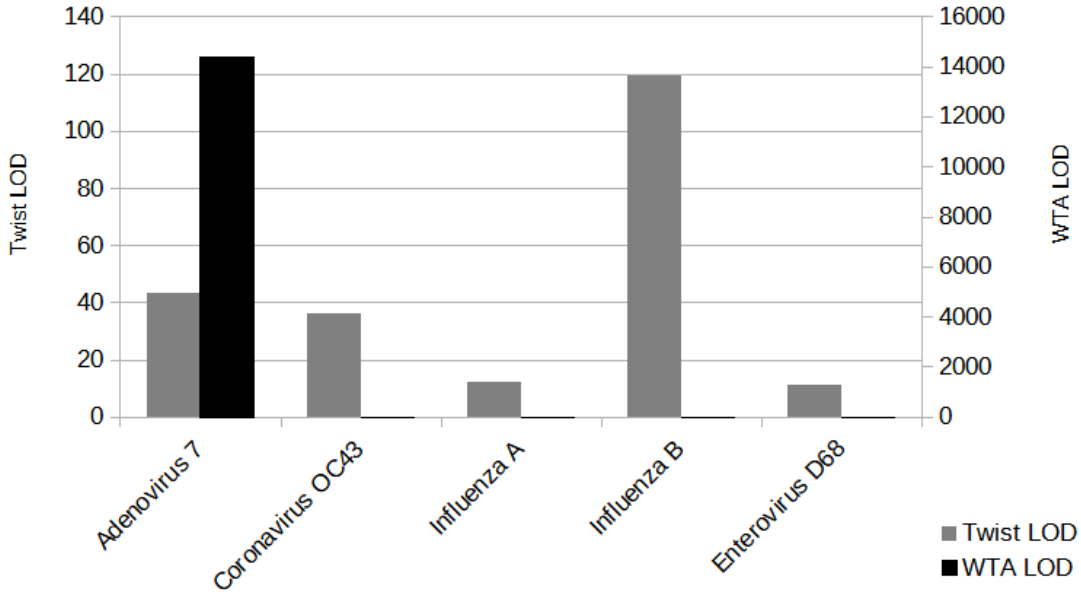

Figure 3 Comparison between LODs of Twist and WTA.

WTA (solid black bars, right Y-axis) was able to detect Adenovirus 7, but had a more than 3-log worse LOD as compared to Twist (solid gray bars, left Y-axis). LODs were unobtainable for WTA in all other circumstances. LOD values are in read counts.

4.3.3 **Depth and breadth of coverage**

At 5,000 pfu WTA was able to cover 43.1% of the hAdv7 genome, but failed to produce even 1% of coverage for any other target (Figure 4). Comparatively, Twist was able to cover over 70% of the genomes for hCoV OC43 and Influenza B, and more than 40% for Enterovirus D68. The two worst performing targets for Twist were hAdv and Influenza A

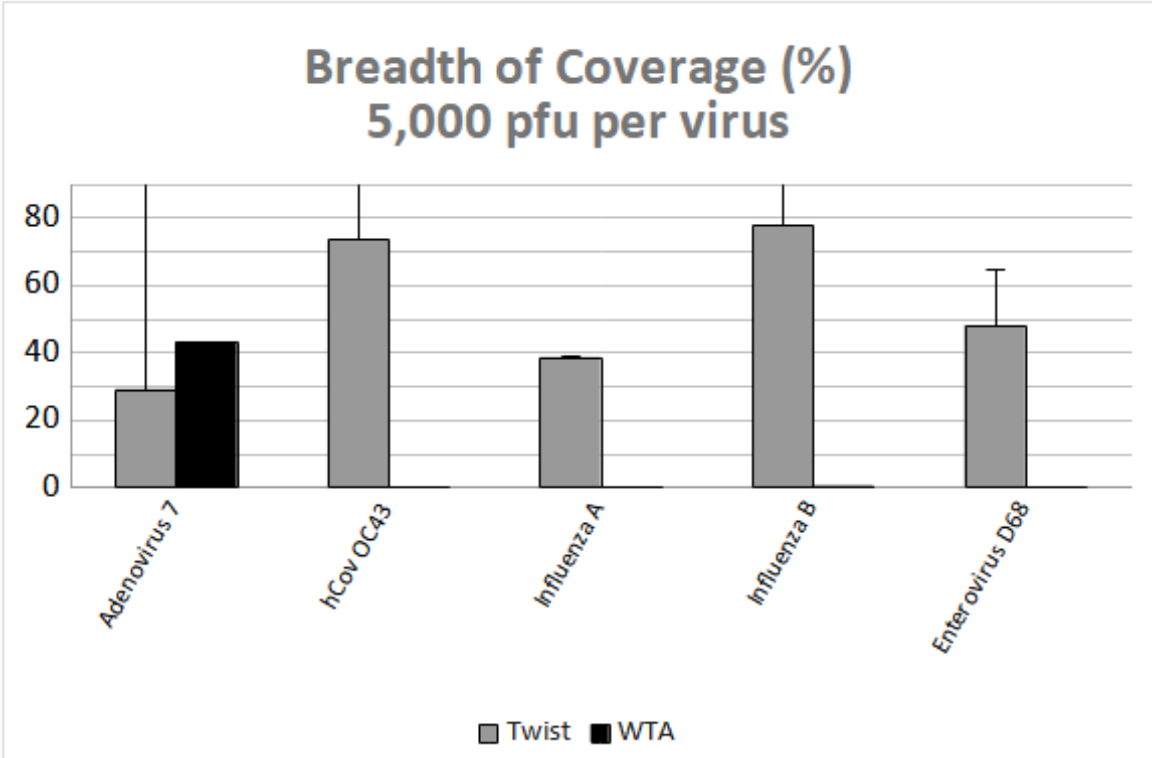

Figure 4 Comparison of breadth of coverage between Twist and WTA at 5,000 pfu.

Twist (gray bars) produced inconsistent coverage across all samples. WTA (black bars) failed to produce any significant coverage, with the exception of hAdv7. Values are displayed as percentage of genome covered at 1x

At 200 pfu, WTA significantly outperformed Twist for hAdv 7, however, it again failed to generate results for any other target (Figure 5). Coverage percentages dropped for Twist, but were still above 20% for hCoV OC43, Influenza A, and Influenza B. Strikingly, Influenza B maintained a coverage of 65.8%.

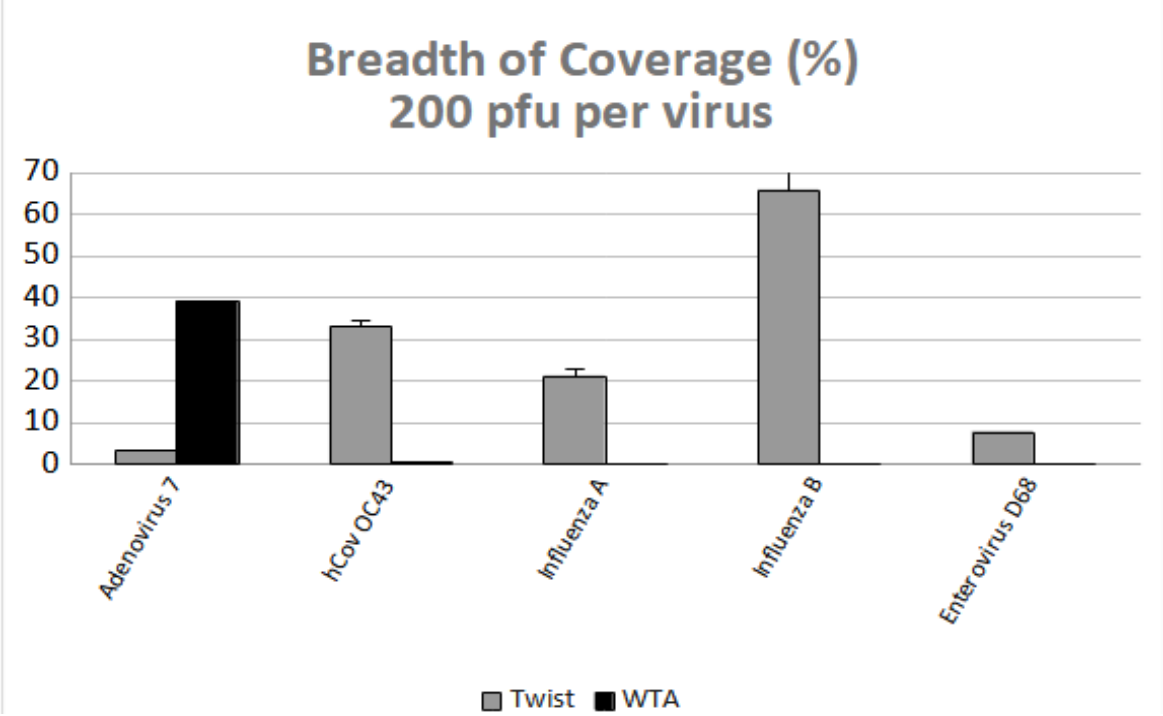

Figure 5 Comparison of breadth of coverage between Twist and WTA at 200 pfu.

WTA (black bars) again only produced coverage for hAdv7. Twist (gray bars) had appreciable decreases in coverage from the 5,000 pfu values. Values are displayed as percentage of genome covered at 1x

The average depth per nucleotide at 5,000 pfu was low for Twist, and non-existent for WTA on all targets except hAdv7 (Figure 6). When comparing the breadth and depth of coverage obtained by WTA for hAdv7, the high depth and less than 50% coverage at 1x likely indicates that the average is due to extremely heterogeneous depths across the genome. Comparatively, Twist has much lower depth of coverage, but higher breadth, thus indicating a

more even distribution across the target genome.

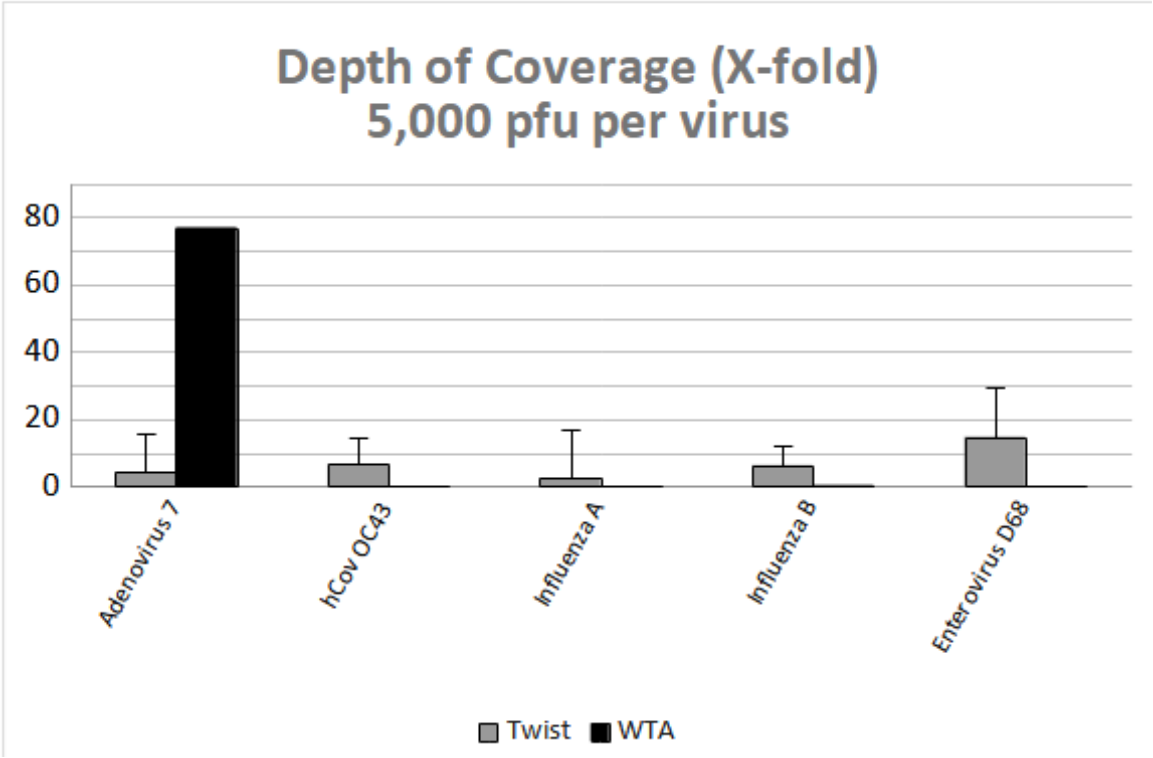

Figure 6 Comparison of depths of coverage between Twist and WTA at 5,000 pfu.

WTA (solid black bars) performed significantly better than the Twist (gray bars) for hAdv with an almost 20x increase in depth of coverage. However, WTA failed to generate any depth > 0.01 for the rest of the targets. Values are shown as average depth of coverage at each nucleotide across the breadth of the genome.

As was the case with depth of coverage at 200 pfu, WTA outperformed Twist for detecting hAdv7, but failed for every other target (Figure 7). As is expected, the overall depth for Twist dropped, however when combined with breadth of coverage, still provides meaningful alignment that could be utilized for viral identification, even if only at the family level.

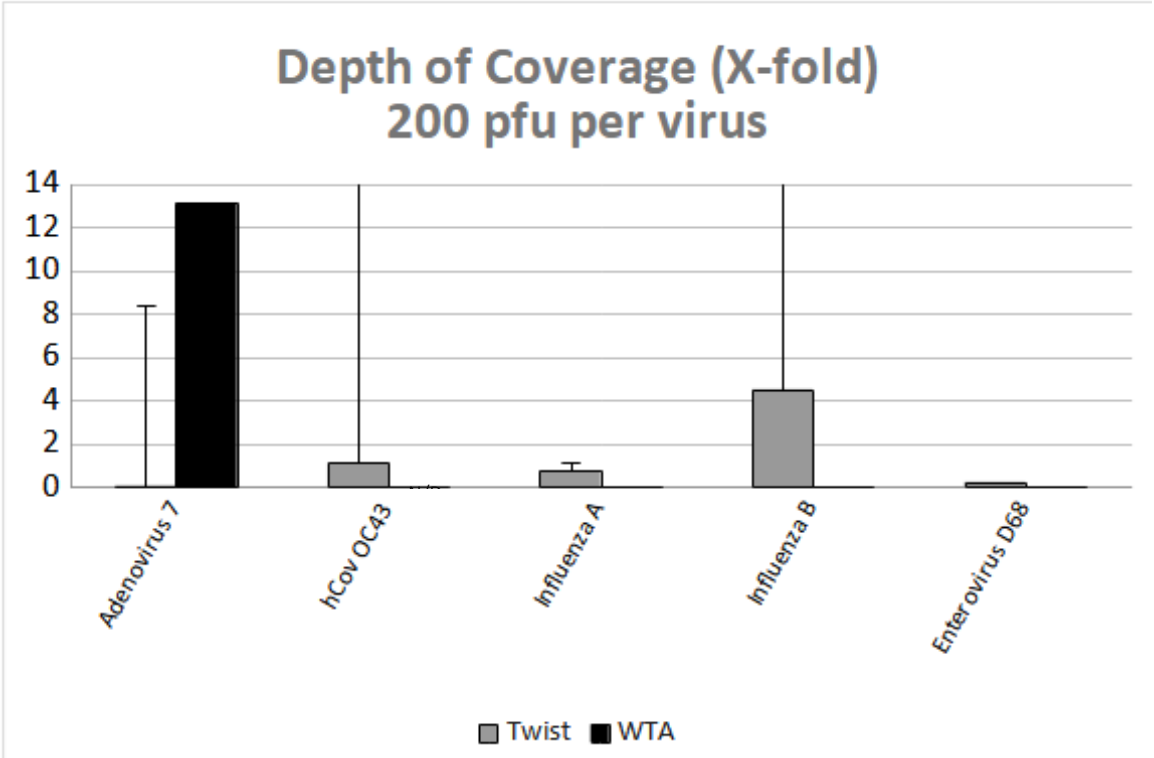

Figure 7 Comparison of depths of coverage between Twist and WTA at 200 pfu.

WTA (black bars) performed better than Twist (gray bars) by over 3 logs. Again, WTA failed to produce significant depth for every other target. Values are shown as average depth of coverage at each nucleotide across the breadth of the genome. Only 1 replicate for Enterovirus D68 was successful at this concentration, thus, there are no error bars

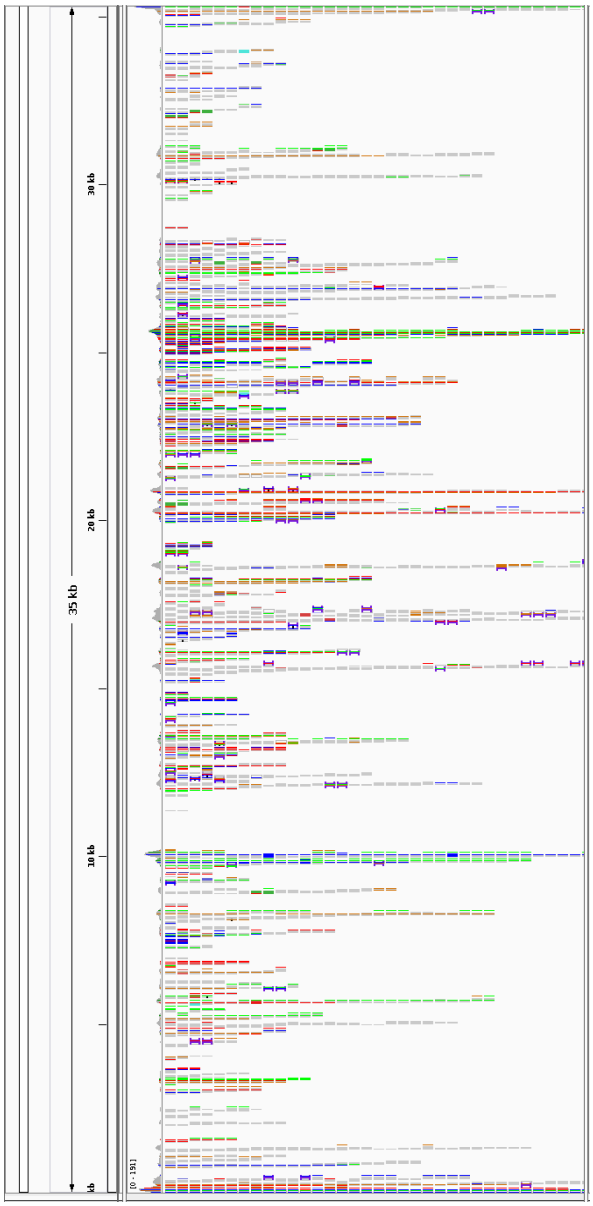

Figure 8: AC_000018.1 (Adeno 7) Graphic representation of the breadth and depth of coverage for one of the Twist Adenovirus 7 samples.

Each colored mark represents one read at that base. Deep coverage is in highly conserved regions of the Adenovirus genome and are likely representative of probe locations. This graphic shows that while the breath of coverage and depth may appear to be adequate, there are still numerous gaps in coverage.

4.3.4 **qPCR for RNAseP**

Using the RefSeq Viral database, most hits in the WTA blank and spiked samples can be categorized as belonging to phages endemic to human microbial colonies and human endogenous retroviruses. In a study examining possible sources of phage/viral contamination, Asplund et al.

(88) showed that certain library preparation components are indicated as origins of contamination by viruses such as Shamonda and Simu viruses. Both of these viruses were identified in almost every sample sequenced, regardless of library prep kit used.

BeAn virus 58058 (BAV) was identified as a major contributor to aberrant reads in all WTA samples. BAV was first isolated from a rodent in the Brazilian rainforest (86) and is implicated as a contaminant in multiple studies utilizing shotgun sequencing (87,88). Using the nt/nr database with the same parameters previously used, hits identified in the WTA samples mapped to the human genome. Further investigation revealed that, using word length of 32 and E Value of 0.0001, the refseq genome for BAV (Accession number KY094066.1) aligns to multiple locations on the human genome (txid9605). This prompted additional examination of potential sources of this seemingly human contamination.

RNAseP is a common positive control target used in clinical testing to assess nucleic acid extraction success and specimen quality (89). In this instance, it was used as a potential indicator of human contamination in the samples tested. As shown in Figure 9, 3 out of 5 nucleic acid extractions tested positive for RNAseP prior to any library preparation. All WTA samples tested had a Ct< 45 for RNAseP, including the blank, whereas only the Twist preparation of hCoV-OC43 was positive. The pre-sequencing pool for Twist was notably negative for RNAseP, indicating that the interfering sequences were likely excluded in the pull-down. These results show that the human contamination was present in at least some of the nucleic acid extractions, and given the incredible sensitive enzymes used for WTA, was likely present in all extractions and/or samples contaminated by human DNA and amplified to a sequenceable level over the course of library preparation. The blank having a Ct of 29.42 is the most curious result pointing

to unintentional technician contamination, or potential known contamination issues using the WTA method, as further explored in the Discussion.

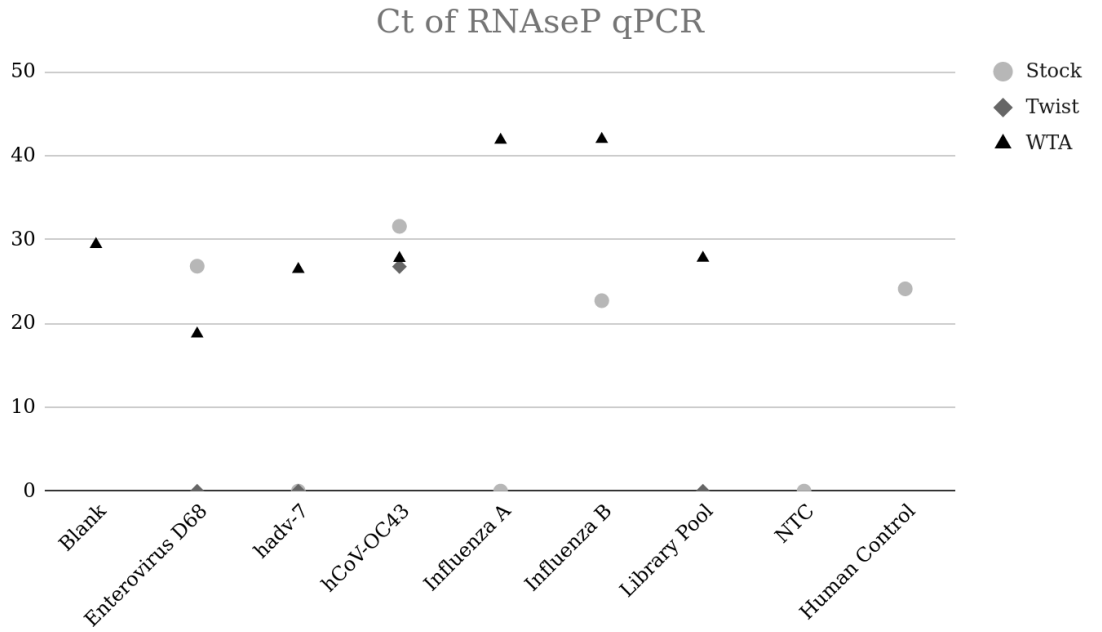

Figure 9 Results of qPCR using RNAseP target primers/probe.

All WTA samples had Cts < 45, while only Twist hCoV-OC43 had a Ct <45. This positive Ct for Twist hCoV-OC43 was absent in the pooled specimen, showing that the capture probes likely are designed to omit human DNA. Human control was extracted from nasopharyngeal swab residuals from previous studies. Und: Undetermined, showing no amplification in qPCR. NTC: No template control, used nuclease-free water.

4.4 **Discussion**

User investigation of manufacturer claims is a critical step in transitioning new technologies to practical use, especially if the technology is to be used for clinical surveillance and/or diagnosis. Ease of use and cost are also critical factors to consider when upscaling production of surveillance data. The ideal assay for global surveillance of new, emerging, and unknown respiratory pathogens is low cost, easy to use, and provides efficient data that requires

little post-acquisition analysis to get an accurate identification. These factors become even more important when one discusses the use of sentinel laboratories in austere or resource limited environments. Data analysis can be outsourced from these areas, however, there remains the importance of ease of use for the wet lab and cost to implement the process.

WTA was the least expensive method of sequencing used in this study, with a 26% reduction in price per sample. Also, time from extraction to sequencing was two-thirds faster than time needed for Twist (1 day versus 3 days). While both kits were easy to use, Twist did require extra steps for the pull-down portion of the assay which could potentially contribute to user error.

However, it does not matter how inexpensive, quick, and easy to use an assay is if it does not provide meaningful data. WTA was prone to failures in sequencing and identification, to the point that one concentration (10 pfu) of targets was unsequenceable. Even when sequencing was successful, WTA only correctly identified the target 20% of the time. These false negatives are problematic, as the phenomena of false negatives in symptomatic patients is the motivating factor for this method comparison study.

An additional point of concern is the high background present in the blank WTA samples. Per Qiagen, high molecular weight DNA is produced even in blank samples due to formation and extension of random primer-dimers (68). As shown in the results, RNAseP was present in 60% of nucleic acid extractions even before any other library preparation steps are taken. This is likely due to trace contamination of cellular host material in the viral stock. All target viruses used in this study are propagated in mammalian cell lines, with the exceptions of Influenzas A and B which are grown in chicken eggs. The primers and probes used for qPCR

(90) map to the ribonuclease P/MRP subunit p30 which is conserved among multiple species, including but not limited to humans and chickens (91). Twist is designed to eliminate non-viral DNA prior to sequencing, so it makes sense that the probes used in the pull-down step would exclude human targets. The biggest question that still remains is why the WTA blank was positive for RNAseP. The easiest explanation is accidental contamination by the author, however, again referencing Asplund et al., introduction of spurious reads by the library prep and sequencing reagents themselves is not uncommon, and thus warrants further investigation.

It could be argued that deeper sequencing would potentially pull out the viral targets as the ratio of target:non-target nucleic acid is generally low. Using random primers for reverse transcription is theorized to be an unbiased way of increasing the total amount of cDNA in the sample for sequencing. However, basic principles of biochemical reactions tell us that even though the primers are truly random, it is statistically more probable that they will disproportionately amplify nucleic acids that are already present in higher amounts in the original sample. The Qiagen WTA Handbook protocol for using already extracted nucleic acids suggests starting with >50pg of RNA. Quantitation of RNA at those low levels is unreliable and differentiation of RNA from DNA is difficult when working with such low amounts. It is likely that the starting amount of target RNA was too low for WTA to accurately capture. Using Influenza B as an example, based on a viral genome copy number to TCID50 ratio of 60:1 (92), there was at maximum of 6.02pg of target RNA in the 5000 pfu samples. This is far lower than the suggested starting amount. Additionally, the second round of sequencing for WTA was done with fewer samples on the same size Illumina kit. More than 3 million reads were available per sample, and except for hAdv7, WTA still failed to capture any of the target viruses.

Twist is more expensive, but much more successful in sequencing and identification than WTA in this study. With an LOB of 2, there is a minimal amount of background in the Twist samples, producing clean alignments and correct identification of the pathogens tested. Of the successful sequences, 92.1% identified the correct virus, with one of the failed identifications being one of the low concentration Influenza A replicates. The other two failures were from Influenza B, with one at 200 pfu and one at 10 pfu. The Influenza A replicates correctly identified the virus but failed to deliver correct strain specificity. Further evaluation of this issue showed that Influenza A H2N2 and Influenza A H3N2 have high segment homology, with the exception of the hemagglutinin gene (segment 4). This demonstrates a limit of the Twist kit, and a need for more specific bioinformatics to accurately strain delineate closely related and segmented genomes.

While the strain delineation for Influenza A requires further investigation, the Twist method performed at or better than the performance metrics of some clinical testing. The BioFire® Film Array® Respiratory Panel (RP) is a popular clinical point-of-care, PCR-based system for rapid identification of respiratory pathogens. Per the manufacturer's documentation (83) the LODs in pfu of the BioFire® Film Array® RP are above the lowest successful concentrations sequenced using Twist (Figure 8).

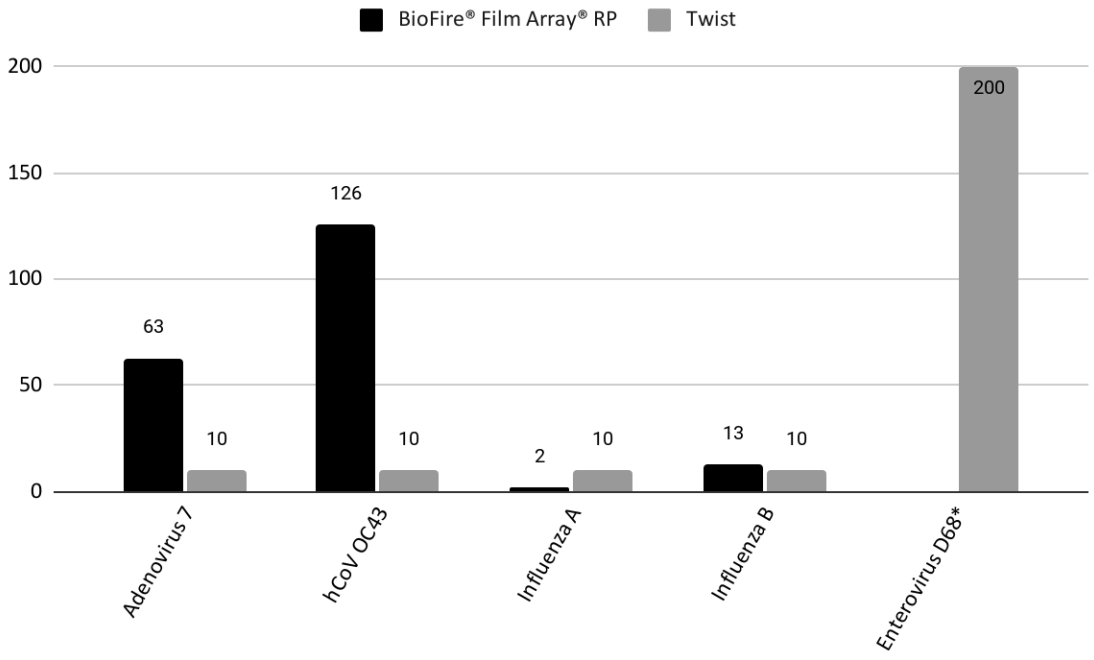

Figure 10 Comparing LODs in pfu between Twist and BioFire® Film Array® RP.

Twist (gray bars) detected target viruses at 10 pfu, with the exception of Enterovirus D68 which was detected at 200 pfu. This is lower than the manufacturer supplied LODs of most targets BioFire® Film Array® RP, with the exception of Influenza A. Values are displayed as pfu per virus per method. *Biofire cannot delineate Enterovirus D68 from other enteroviruses and rhinoviruses, thus depending on the version of the kit used, samples that are actually positive for Enterovirus D68 output as positive for Enterovirus/Rhinovirus and require follow-up testing.

In addition to having lower detection thresholds for most viruses surveyed, sequencing also has the advantage of being able to identify pathogens even if there are mutations in the genome. Clinical PCR testing relies heavily on highly conserved regions of the viral genome in order to produce a test that is both sensitive and specific. Should these conserved regions change or, as was the case with SARS, MERS, and SARS-CoV-2, an entirely new pathogen emerge, these tests may fail to identify a causative agent of illness. Additionally, with deeper sequencing, it is likely that there would be a significant increase in breadth and depth of coverage. The next step in this study would be to test the ability of Twist to accurately identify mutated pathogens at multiple concentrations. Additionally, inclusion of bacterial and fungal targets is necessary in

order to reliably conclude that Twist can capture pathogen targets that are not explicitly included in the pull down probes.

5 CONCLUSIONS

Largely, WTA failed to sequence and to identify the target pathogens. Even though it was able to properly identify hAdv, the resulting data was difficult to parse into meaningful results. Statistical comparisons between Twist and WTA were nearly impossible due to the numerous failures of WTA. Twist, while more expensive per sample, provided results on low concentration samples and needed little post-sequencing processing. WTA as a method has many advantages when looking at true transcriptomes, however, it would require more optimization in order to be successful in identifying viral nucleic acids from clinical specimens.

The Twist results from this study correlate with results published by the manufacturer (31). Twist Biosciences included limited data on depth and breadth of coverage on RSV, the HA segments of different Influenza A H1N1 strains, and a SARS-CoV-2 analog. This study further challenged the abilities of this assay to detect respiratory pathogens with varying genetic and biochemical arrangements with accuracy, ease, and minimal specialized training.

While there have been vast technical and programmatic improvements in surveillance, recent influenza and coronavirus pandemics have exposed gaps in these systems. Establishing a reliable method for sequencing unknown respiratory pathogens is critical for improving the process in which we handle outbreaks and potential pandemics. Evaluation and comparison of both WTA and Twist is simply a first step in the effort to improve surveillance. If SARS-CoV-2 has been any indication, it requires a global multi partisan effort to identify and contain the pathogens that threaten our health.

6 RECOMMENDATIONS

This project was a pilot study to examine the performance of a targeted pull-down sequencing assay and compare its ability to produce meaningful data when compared to a more traditional shotgun approach. It is my recommendation that evaluation of the assay continues, with the goal of incorporating it into a larger scale surveillance program targeting unknown pathogens. Listed below are my recommended expansions of the study:

1. Evaluate hybrid capture on bacterial targets. Virulence and antibiotic resistance markers are of particular interest, especially when looking at colonization of wounds, changes in the native human microbiome, and tracking nosocomial infections. Changes in both species composition and genetic makeup of the human microbiome can provide meaningful insight in to the illness-wellness continuum and how is correlates to optimized warfighter performance.

2. Expand the list of targets to include gastrointestinal pathogens. Viral and bacterial GI illness is a major cause of lost duty days among warfighters (93). While some cases can be attributed to a singular pathogen, many causative agents remain unknown. Sequencing fecal samples of ill patients that test negative using clinical testing could elucidate answers about the origins of many GI illness cases.

3. Evaluate the performance of hybrid capture targeted sequencing on complex environmental samples such as wastewater and soil. Identifying pathogens that are inherently in soil and/or wastewater, combined with metadata on disease incidence and population metrics, could serve as an early warning system for a

potential large scale outbreak as well as fill in the missing gaps of the disease transmission cycle.

7 REFERENCES

1 Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Yi Hu, Zhi-Gang Song, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes, Yong-Zhen Zhang. Complete genome characterization of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. bioRxiv 2020.01.24.919183; doi: https://doi.org/10.1101/2020.01.24.919183

2 Coronavirus Disease (COVID-19). World Health Organization. https://www.who.int/health-topics/coronavirus. Published 2020. Accessed January 19, 2022.

3 Corman, Victor M et al. “Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.” *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* vol. 25,3 (2020)

4 Dong J, Olano JP, McBride JW, Walker DH. Emerging pathogens: challenges and successes of molecular diagnostics. *J Mol Diagn*. 2008;10(3):185-197. doi:10.2353/jmoldx.2008.070063

5 Shrestha SS, Swerdlow DL, Borse RH, Prabhu VS, Finelli L, Atkins CY, Owusu-Edusei K, Bell B, Mead PS, Biggerstaff M, Brammer L, Davidson H, Jernigan D, Jhung MA, Kamimoto LA, Merlin TL, Nowell M, Redd SC, Reed C, Schuchat A, Meltzer MI. Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009-April 2010). Clin Infect Dis. 2011 Jan 1;52 Suppl 1:S75-82

6 Perlman S, Zumla A. MERS-CoV in Africa-an enigma with relevance to COVID-19. Lancet Infect Dis. 2020 Oct 6:S1473-3099(20)30578-8. doi: 10.1016/S1473

7 Detection and characterization of respiratory viruses, including SARS-CoV-2, using Illumina RNA Prep with Enrichment. Illumina. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-rna-enrichment-coronavirus-app-note-470-2020-005/illumina-rna-enrichment-coronavirus-app-note-470-2020-005.pdf Published 2020. Accessed October 2021

8 Middle East respiratory syndrome coronavirus (MERS-CoV). World Health Organization. https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers. Published 2012. Accessed October 2021

9 Chapleau RR, Christian M, Connors B, Premo C, Chao TC, Rodriguez J, Huntsberger S, Meyer J, Javorina A, Reynolds K, Riddle D, Lisanby M, Starr C. Early Identification of SARS-CoV-2 Emergence in the Department of Defense via Retrospective Analysis of 2019-2020 Upper Respiratory Illness Samples. MSMR. 2021 Jun 1;28(6):2-5. PMID: 34379377.

10 Institute of Medicine 2008. Review of the DoD-GEIS Influenza Programs: Strengthening Global Surveillance and Response. Washington, DC: The National Academies Press. https://doi.org/10.17226/11974.

11 Influenza (Flu). Centers for Disease Control and Prevention https://www.cdc.gov/flu/index.htm. Accessed October 2021

12 History. Global Initiative on Sharing Avian Influenza Data. https://gisaid.org/about-us/history/. Published May 2008. Accessed October 2021

13 Twist Biosciences. *NGS Universal Blockers*. 2020. ProductSheet_NGS_UniversalBlockers_8JUL20_Rev1.1.pdf

14 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170

15 Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data (Online). Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

16 SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation Shen W, Le S, Li Y, Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11(10): e0163962. https://doi.org/10.1371/journal.pone.0163962

17 Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and WebbMiller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

18 Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

19 James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).

20 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9 (19505943)

21 Illumina. *Index Adapters Pooling Guide (document #1000000041074)*. 2018

22 Bouvier NM, Palese P. The biology of influenza viruses. Vaccine. 2008 Sep 12;26 Suppl 4(Suppl 4):D49-53. doi: 10.1016/j.vaccine.2008.07.039. PMID: 19230160; PMCID: PMC3074182.

23 Palese P, Shaw ML. Orthomyxoviridae: The Viruses and their Replication. In: Knipe DM, Howley PM, editors. *Fields Virology.* Philadelphia: Lippincott Williams & Wilkins; 2007

24 Influenza: Are We Ready?. World Health Organization.https://www.who.int/news-room/spotlight/influenza-are-we-ready. Published October 12, 2018. Accessed September 2022

25 National Notifiable Diseases Surveillance System (NNDSS). Centers for Disease Control and Prevention. Published 2015 August 19. Accessed November 2022

26 Select Agents and Toxins List. Centers for Disease Control and Prevention. https://www.selectagents.gov/sat/list.htm?CDC_AA_refVal=https%3A%2F%2Fwww.selectagents.gov%2FSelectAgentsandToxinsList.html. Last Reviewed January 17 2023. Accessed January 2023.

27 Noor R, Maniha SM. A brief outline of respiratory viral disease outbreaks: 1889-till date on the public health perspectives. Virusdisease. 2020 Dec;31(4):441-449. doi: 10.1007/s13337-020-00628-5. Epub 2020 Sep 2. PMID: 32905186; PMCID: PMC7466926.

28 The Deadly Virus: The Influenza Epidemic 1918. National Archives and Records Administration. https://www.archives.gov/exhibits/influenza-epidemic/#:~:text=World%20War%20I%20claimed%20an,other%20illness%20in%20recorded%20history. Accessed October 2022

29 Zinsser, Hans (1996) [1935]. *Rats, Lice and History: A Chronicle of Pestilence and Plagues*. New York: Black Dog & Leventhal. ISBN 978-1-884822-47-6.

30 Ochman, Sophie; Roser, Max. Polio (graph "Reported paralytic polio cases and deaths in the United States since 1910") . Our World in Data. OurWorldInData.org. Retrieved 15 May 2021. Front. Microbiol., 15 January 2021 Sec. Infectious Agents and Diseases https://doi.org/10.3389/fmicb.2020.631736

31 Twist Bioscience. Twist Total Nucleic Acids Library Preparation Kit for Viral Pathogen Detection and Characterization. https://www.twistbioscience.com/resources/protocol/twist-total-nucleic-acids-library-preparation-kit-viral-pathogen-detection-and. Accessed September 2021

32 1968 Pandemic (H3N2). Centers for Disease Control. https://www.cdc.gov/flu/pandemic-resources/1968-pandemic.html. Published January 2, 2019. Accessed February 2021

33 Leland DS, Ginocchio CC. Role of cell culture for virus detection in the age of technology. Clin Microbiol Rev. 2007 Jan;20(1):49-78. doi: 10.1128/CMR.00002-06. PMID: 17223623; PMCID: PMC1797634.

34 Middle East respiratory syndrome coronavirus (MERS-CoV). World Health Organization https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers. Published 2012. Accessed October 2021

35 Honigsbaum M. Revisiting the 1957 and 1968 influenza pandemics. *The Lancet*. 2020;395(10240):1824-1826. doi:10.1016/s0140-6736(20)31201-0

36 Jackson C. History lessons: the Asian flu pandemic. Br J Gen Pract. 2009 Aug;59(565):622-3. doi: 10.3399/bjgp09X453882. PMID: 22751248; PMCID: PMC2714797.

37 Upper Respiratory Infection. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/4022-upper-respiratory-infection. Published May 5, 2021. Accessed January 2023

38 Koutsakos, M., Wheatley, A.K., Laurie, K. *et al.* Influenza lineage extinction during the COVID-19 pandemic?. *Nat Rev Microbiol* 19, 741–742 (2021). https://doi.org/10.1038/s41579-021-00642-4

39 Valesano, Andrew & Fitzsimmons, William & McCrone, John & Petrie, Joshua & Monto, Arnold & Martin, Emily. (2019). Influenza B viruses exhibit lower within-host diversity than influenza A viruses in human hosts. 10.1101/791038.

40 Koutsakos, M., Wheatley, A.K., Laurie, K. *et al.* Influenza lineage extinction during the COVID-19 pandemic?. *Nat Rev Microbiol* 19, 741–742 (2021). https://doi.org/10.1038/s41579-021-00642-4

41 García-Sastre A. Influenza virus receptor specificity: disease and transmission. Am J Pathol. 2010 Apr;176(4):1584-5. doi: 10.2353/ajpath.2010.100066. Epub 2010 Mar 4. PMID: 20203283; PMCID: PMC2843447.

42 Jonathan A. McCullers, Frederick G. Hayden, Fatal Influenza B Infections: Time to Reexamine Influenza Research Priorities, *The Journal of Infectious Diseases*, Volume 205, Issue 6, 15 March 2012, Pages 870–872, https://doi.org/10.1093/infdis/jir865

43 Stewart P.L.et al. Image reconstruction reveals the complex molecular organization of adenovirus. *Cell.* 1991; 67: 145-154

44 Virus Taxonomy: 2020 Release. International Committee on Taxonomy of Viruses (ICTV). Published March 2021. Retrieved 28 Feb 2023

45 Dhingra A, Hage E, Ganzenmueller T, Böttcher S, Hofmann J, Hamprecht K, et al. (January 2019). "Molecular Evolution of Human Adenovirus (HAdV) Species C". *Scientific Reports*. 9 (1): 1039. Bibcode:2019NatSR...9.1039D. doi:10.1038/s41598-018-37249-4. PMC 6355881. PMID 30705303.

46 Adenovirus. Centers for Disease Control and Prevention. https://www.cdc.gov/adenovirus/index.html. Published January 31 2020. Retrieved January 6 2021.

47 Fu, Y., Tang, Z., Ye, Z. *et al.* Human adenovirus type 7 infection causes a more severe disease than type 3. *BMC Infect Dis* 19, 36 (2019). https://doi.org/10.1186/s12879-018-3651-2

48 Marie E Killerby, Faye Rozwadowski, Xiaoyan Lu, Mardea Caulcrick-Grimes, Lisa McHugh, Ann Marie Haldeman, Tara Fulton, Eileen Schneider, Senthilkumar K Sakthivel, Julu Bhatnagar, Demi B Rabeneck, Sherif Zaki, Susan I Gerber, John T Watson, Respiratory Illness Associated With Emergent Human Adenovirus Genome Type 7d, New Jersey, 2016–2017, *Open Forum Infectious Diseases*, Volume 6, Issue 2, February 2019, ofz017, https://doi.org/10.1093/ofid/ofz017

49 Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev. 2008 Aug;29 Suppl 1(Suppl 1):S49-52. PMID: 18852857; PMCID: PMC2556583.

50 Wang H, Jean S, Wilson SA, Lucyshyn JM, McGrath S, Wilson RK, Magrini V, Leber AL. A deletion in the N gene of SARS-CoV-2 may reduce test sensitivity for detection of SARS-CoV-2. Diagn Microbiol Infect Dis. 2022 Apr;102(4):115631. doi: 10.1016/j.diagmicrobio.2021.115631. Epub 2021 Dec 29. PMID: 35045382; PMCID: PMC8715644

51 CDC Museum COVID-19 Timeline. David J. Sencer CDC Museum: In Association with the Smithsonian Institution. https://www.cdc.gov/museum/timeline/covid19.html#Early-2020. Last Reviewed March 17, 2023. Accessed December 2022

52 S Pavia C, M Plummer M. The evolution of rapid antigen detection systems and their application for COVID-19 and other serious respiratory infectious diseases. J Microbiol

Immunol Infect. 2021 Oct;54(5):776-786. doi: 10.1016/j.jmii.2021.06.003. Epub 2021 Jun 26. PMID: 34272205; PMCID: PMC8234251.

53 Clinical Laboratory Improvement Amendment: CLIA 88. https://www.govinfo.gov/content/pkg/STATUTE-102/pdf/STATUTE-102-Pg2903.pdf

54 Accuracy of the Tests. *Rabies*. Centers for Disease Control and Prevention. https://www.cdc.gov/rabies/diagnosis/accuracy.html. Published April 22, 2011. Accessed January 2023

55 Vrbpac March 7, 2023 meeting announcement. U.S. Food and Drug Administration. https://www.fda.gov/advisory-committees/advisory-committee-calendar/vaccines-and-related-biological-products-advisory-committee-march-7-2023-meeting-announcement. Published March 7, 2023. Accessed March 7, 2023.

56 Renata Soares, Tiana Tasca. Giardiasis: an update review on sensitivity and specificity of methods for laboratorial diagnosis, Journal of Microbiological Methods, Volume 129, 2016, Pages 98-102, ISSN 0167-7012, https://doi.org/10.1016/j.mimet.2016.08.017.

57 Herrmann B, Larsson C, Zweygberg BW. Simultaneous detection and typing of influenza viruses A and B by a nested reverse transcription-PCR: comparison to virus isolation and antigen detection by immunofluorescence and optical immunoassay (FLU OIA). J Clin Microbiol. 2001 Jan;39(1):134-8. doi: 10.1128/JCM.39.1.134-138.2001. PMID: 11136761; PMCID: PMC87692.

58 Shelton H, Ayora-Talavera G, Ren J, Loureiro S, Pickles RJ, Barclay WS, Jones IM. Receptor binding profiles of avian influenza virus hemagglutinin subtypes on human

cells as a predictor of pandemic potential. J Virol. 2011 Feb;85(4):1875-80. doi: 10.1128/JVI.01822-10. Epub 2010 Nov 24. PMID: 21106732; PMCID: PMC3028872.

59 Wang, H., Li, ZY., Liu, Y. *et al.* Desmoglein 2 is a receptor for adenovirus serotypes 3, 7, 11 and 14. *Nat Med* 17, 96–104 (2011). https://doi.org/10.1038/nm.2270

60 Crenshaw BJ, Jones LB, Bell CR, Kumar S, Matthews QL. Perspective on Adenoviruses: Epidemiology, Pathogenicity, and Gene Therapy. Biomedicines. 2019 Aug 19;7(3):61. doi: 10.3390/biomedicines7030061. PMID: 31430920; PMCID: PMC6784011.

61 Notes from the Field: Fatalities Associated with Human Adenovirus Type 7 at a Substance Abuse Rehabilitation Facility — New Jersey, 2017. *MMWR.* 2018;67:371–372.

62 Clinical Overview. Adenoviruses. Centers for Disease Control and Prevention. https://www.cdc.gov/adenovirus/hcp/clinical-overview.html#:~:text=Some%20people%20infected%20with%20adenoviruses,virus%20for%20weeks%20or%20longer. Last Reviewed November 23, 2022. Accessed February 2023

63 Sinclair W, Omar M. Enterovirus. StatPearls Publishing;Treasure Island, Fl. January 2022. https://www.ncbi.nlm.nih.gov/books/NBK562330/

64 Hixon AM, Frost J, Rudy MJ, Messacar K, Clarke P, Tyler KL. Understanding Enterovirus D68-Induced Neurologic Disease: A Basic Science Review. Viruses. 2019 Sep 4;11(9):821. doi: 10.3390/v11090821. PMID: 31487952; PMCID: PMC6783995.

65 Oberste M.S., Maher K., Schnurr D., Flemister M.R., Lovchik J.C., Peters H., Sessions W., Kirk C., Chatterjee N., Fuller S., et al. Enterovirus 68 is associated with respiratory illness and shares biological features with both the enteroviruses and the rhinoviruses. *J. Gen. Virol.* 2004;85:2577–2584. doi: 10.1099/vir.0.79925-0.

66 Enterovirus D68. Centers for Disease Control https://www.cdc.gov/non-polio-enterovirus/about/ev-d68.html. Last Reviewed September 9, 2022. Accessed February 2023

67 Dyda Amalie, Stelzer-Braid Sacha, Adam Dillon, Chughtai Abrar A, MacIntyre C Raina. The association between acute flaccid myelitis (AFM) and Enterovirus D68 (EV-D68) – what is the evidence for causation?. Euro Surveill. 2018;23(3):pii=17-00310. https://doi.org/10.2807/1560-7917.ES.2018.23.3.17-00310

68 Repli-g® Single-Cell Handbook. Qiagen. https://www.qiagen.com/us/resources/resourcedetail?id=f2430d44-7f49-4544-b2de-5a51b904ea39&lang=en Published October 2012

69 Nextera XT DNA Library Prep. Illumina. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-06.pdf. March 30, 2022

70 Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev. 2008 Aug;29 Suppl 1(Suppl 1):S49-52. PMID: 18852857; PMCID: PMC2556583.

71 Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, Vandamme AM, Van Ranst M. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol. 2005 Feb;79(3):1595-604. doi: 10.1128/JVI.79.3.1595-1604.2005. PMID: 15650185; PMCID: PMC544107.

72 Kim MI, Lee C. Human Coronavirus OC43 as a Low-Risk Model to Study COVID-19. Viruses. 2023 Feb 20;15(2):578. doi: 10.3390/v15020578. PMID: 36851792; PMCID: PMC9965565.

73 Keshavarz Valian N, Pourakbari B, Asna Ashari K, Hosseinpour Sadeghi R, Mahmoudi S. Evaluation of human coronavirus OC43 and SARS-COV-2 in children with respiratory tract infection during the COVID-19 pandemic. J Med Virol. 2022 Apr;94(4):1450-1456. doi: 10.1002/jmv.27460. Epub 2021 Nov 24. PMID: 34786736; PMCID: PMC8661600.

74 CDC SARS Response Timeline. Centers for Disease Control and Prevention. https://www.cdc.gov/about/history/sars/timeline.htm. Last Reviewed April 26, 2013. Accessed February 2023

75 SARS-Associated Coronavirus (SARS-CoV) Sequencing. Centers for Disease Control and Prevention. https://www.cdc.gov/sars/lab/sequence.html. Last reviewed December 6, 2017. Accessed February 2023.

76 Liu J, Xie W, Wang Y, Xiong Y, Chen S, Han J, Wu Q. A comparative overview of COVID-19, MERS and SARS: Review article. Int J Surg. 2020 Sep;81:1-8. doi: 10.1016/j.ijsu.2020.07.032. Epub 2020 Jul 26. PMID: 32730205; PMCID: PMC7382925.

77 MERS-CoV, SARS-CoV, SARS-CoV-2 Agent Information Sheet. Boston University. https://www.bu.edu/researchsupport/safety/rohp/agent-information-sheets/coronaviruses-agent-information-sheet/#:~:text=Approximately%2020%25%20of%20COVID%2D19,case%20fatality%20rate%20was%2034%25. Accessed February 2023.

78 Saif LJ. Bovine respiratory coronavirus. Vet Clin North Am Food Anim Pract. 2010 Jul;26(2):349-64. doi: 10.1016/j.cvfa.2010.04.005. PMID: 20619189; PMCID: PMC4094360.

79 Liu Q, Wang HY. Porcine enteric coronaviruses: an updated overview of the pathogenesis, prevalence, and diagnosis. Vet Res Commun. 2021 Sep;45(2-3):75-86. doi: 10.1007/s11259-021-09808-0. Epub 2021 Jul 12. PMID: 34251560; PMCID: PMC8273569.

80 Pathogen Genomics: Sequencing and Surveillance. American Society of Microbiology. https://asm.org/Resource-Pages/Advanced-Molecular-Detection-AMD. Last reviewed 2023. Accessed March 2023

81 What is the Optimal Length of a Primer. Frequently Asked Questions. Integrated DNA Technologies. https://www.idtdna.com/pages/support/faqs/what-is-the-optimal-length-of-a-primer-#:~:text=IDT%20recommends%20you%20aim%20for,secondary%20structures%20and%20self%2Dcomplementarity. Accessed March 2023

82 James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

83 Amiott E, Kanack K, Robbins T, Poritz M, Lingenfelter B, Ririe K. Analytical evaluation of the FilmArray Respiratory Panel. https://www.biofiredefense.com/wp-content/uploads/2018/08/Analytical-Evaluation-of-the-FilmArray-Respiratory-Panel-2011-Pstr.pdf. Published 2011. Accessed March 2023. Presented at the ASM General Meeting, 2011

84 Dadonaite, B., Gilbertson, B., Knight, M.L. *et al.* The structure of the influenza A virus genome. *Nat Microbiol* **4**, 1781–1789 (2019). https://doi.org/10.1038/s41564-019-0513-7

85 A CONFIDENCE INTERVAL FOR A POPULATION STANDARD DEVIATION, KNOWN OR LARGE SAMPLE SIZE. INTRODUCTORY BUSINESS STATISTICS, RICE UNIVERSITY. HTTPS://OPENSTAX.ORG/BOOKS/INTRODUCTORY-BUSINESS-STATISTICS/PAGES/8-1-A-CONFIDENCE-INTERVAL-FOR-A-POPULATION-STANDARD-DEVIATION-KNOWN-OR-LARGE-SAMPLE-SIZE

86 WANZELLER AL, SOUZA AL, AZEVEDO RS, JÚNIOR EC, FILHO LC, OLIVEIRA RS, LEMOS PS, JÚNIOR JV, VASCONCELOS PF. COMPLETE GENOME SEQUENCE OF THE BEAN 58058 VIRUS ISOLATED FROM *ORYZOMYS* SP. RODENTS IN THE AMAZON REGION OF BRAZIL. GENOME ANNOUNC. 2017 MAR 2;5(9):E01575-16. DOI: 10.1128/GENOMEA.01575-16. PMID: 28254970; PMCID: PMC5334577.

87 CONTAMINATING VIRAL SEQUENCES IN HIGH-THROUGHPUT SEQUENCING VIROMICS: A LINKAGE STUDY OF 700 SEQUENCING LIBRARIES, CLINICAL MICROBIOLOGY AND INFECTION, VOLUME 25, ISSUE 10, 2019, PAGES 1277-1285, ISSN 1198-743X, HTTPS://DOI.ORG/10.1016/J.CMI.2019.04.028.\

88 Slavov SN. Viral Metagenomics for Identification of Emerging Viruses in Transfusion Medicine. *Viruses*. 2022; 14(11):2448. https://doi.org/10.3390/v14112448

89 Fernandes-Monteiro AG, Trindade GF, Yamamura AM, Moreira OC, de Paula VS, Duarte AC, Britto C, Lima SM. New approaches for the standardization and validation of a real-time qPCR assay using TaqMan probes for quantification of yellow fever virus on clinical samples with high quality parameters. Hum Vaccin Immunother. 2015;11(7):1865-71. doi: 10.4161/21645515.2014.990854. PMID: 26011746; PMCID: PMC4514303.

90 Lu X, Wang L, Sakthivel SK, Whitaker B, Murray J, Kamili S, Lynch B, Malapati L, Burke SA, Harcourt J, Tamin A, Thornburg NJ, Villanueva JM, Lindstrom S. US CDC Real-Time Reverse Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis. 2020 Aug;26(8):1654–65. doi: 10.3201/eid2608.201246. Epub 2020 May 15. PMID: 32396505; PMCID: PMC7392423.

91 U.S. National Library of Medicine. (2023, April 9). *RPP30 ribonuclease P/MRP subunit P30 [homo sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. Retrieved May 4, 2023, from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=10556

92 Parker, Jayme; Fowler, Nisha; Louise Walmsley, Mary; Schmidt, Terri; Scharrer, Jason; Kowaleski, James; et al. (2015): Relationship between TCID50/mL concentrations and copy number.. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0143164.t003

93 Riddle MS, Tribble DR, Cachafiero SP, Putnam SD, Hooper TI Development of a travelers' diarrhea vaccine for the military: how much is an ounce of prevention really worth? Vaccine 2008; 26(20): 2490–502.

APPENDIX

APPENDIX A:

#Massblast- local instance blast against refseq_viral NCBI

```
#!/bin/bash

#for big ole nucleotide blasty time

for file in $PWD/*fastq.fasta

do

blastn -query $file -db refseq_viral.fasta -num_threads "4" -outfmt "6 stitle pident evalue
bitscore"

-max_target_seqs "1" -evalue "0.001" -out ""$file"_refseq_viral.out3"

done
```

APPENDIX B

#hit count. Output csv

```
#!/bin/bash

#get number of hits per ID from blast file

for file in $PWD/*out3

do

cut -f 1 "$file" | sort | uniq -c > "$file"_hits.txt

done
```

APPENDIX C

#ptrimfa- fq to fna using seqtk

```
#!/bin/bash

for fq in *.fastq
```

```
do
seqtk seq -a $fq >$fq\.fasta
done
```

APPENDIX D

```
#autotrim
#!/bin/bash
#for trimming multiple files in TrimmomaticPE
#adjust adapter set after ILLUMINACLIP: use TruSeq for Twist
for infile in *_R1_001.fastq
do
base=$(basename ${infile} _R1_001.fastq)
TrimmomaticPE ${infile} ${base}_R2_001.fastq \
${base}_R1.001.trim.fastq ${base}_R1.001.notrim.fastq \
${base}_R2.001.trim.fastq ${base}_R2.001.notrim.fastq \
LEADING:3 TRAILING:3 MINLEN:36 ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:2
done
```

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ACRONYM	DESCRIPTION
AFM	Acute flaccid myelitis
ATCC	American type tissue culture
BAV	BeAn 58058
BSL	Biosafety level
CDC	The United States Center for Disease Control
DNA	deoxyribonucleic acid
dsDNA	double stranded DNA
dsRNA	double stranded RNA
EV-D68	Enterovirus D68
gDNA	Genomic DNA
GISAID	Global Initiative on Sharing Avian Influenza Data
GISRS	Global Influenza Surveillance and Response System
HA	Hemagglutinin
hAdv3	human adenovirus 3
hAdv7	Human adenovirus 7
hCoV-OC43	Human coronavirus OC43
IVD	In vitro diagnostic
LOB	Limit of blank
LOD	Limit of detection
MDA	multiple strand displacement amplification
MERS	Middle eastern respiratory syndrome
mRNA	Messenger RNA
NA	Neuraminidase
NNDSS	National Notifiable Diseases Surveillance System
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
Pfu	Plaque forming unit
QC	Quality control
qPCR	Quantitative PCR
RNA	ribonucleic acid
RNAseP	Ribonuclease P
SARS	Sudden acute respiratory syndrome
SARS-CoV-2	Sudden acute respiratory syndrome coronavirus 2
ssRNA	single stranded RNA
WHO	World Health Organization
WTA	Whole transcriptome amplification