Big Data Meta-Analyses of Transcriptional Responses of Human Samples to Orthohantavirus Infection and Shotgun Metagenomics From Crohn's Disease Patients.

John L. Krapohl
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Life Sciences Commons

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/9667

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Big Data Meta-Analyses of Transcriptional Responses of Human Samples to *Orthohantavirus* Infection
and Shotgun Metagenomics from Crohn’s Disease Patients

John L. Krapohl

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Brett E. Pickett, Chair
K. Scott Weber
Bradford K. Berges

Department of Microbiology and Molecular Biology
Brigham Young University

Copyright © 2022 John L. Krapohl
All Rights Reserved
ABSTRACT

Big Data Meta-Analyses of Transcriptional Responses of Human Samples to Orthohantavirus Infection and Shotgun Metagenomics from Crohn’s Disease Patients

John L. Krapohl
Department of Microbiology and Molecular Biology, BYU
Master of Science

Hantavirus is a dangerous zoonotic viral pathogen that is found across Asia, Europe, and the Americas. This virus causes a range of symptoms from flu-like malaise to heart failure and death. It is normally transmitted to humans via the aerosolized feces or urine of infected rodents. Currently, there are no known treatments for the disease, and it continues to threaten human health in endemic areas. In order to identify possible future therapeutic targets, we ran a meta-analysis of existing transcriptomic data collected from infected human tissue. Several genes and cellular pathways were identified, in addition to several potential therapeutics that warrant additional testing as potential future therapeutics for hantavirus infection. Such genes include, but are not limited to SLC27A3, NOG, AMIGO1, NUSAP1, and CDC25C which have not been previously associated with hantavirus infection. In addition, we identified that RIG-I and MDA5-associated anti-viral response genes are downregulated, while downstream elements of these pathways are upregulated, indicative of immune activation via alternate pathways. Finally, among the potential therapeutics we identified are dinaciclib, alvodicib, and ruxolitinib, which limit cellular replication, as well as ruxolitinib, baricitinib, and tofacitinib, which target other human intracellular pathways that may aid in successful viral infection.

Crohn’s disease is an autoimmune disorder that affects the digestive system of more than six million people worldwide, with most cases found in North America and Europe. Although the disease can occur throughout the entire digestive tract, the classical sign of disease progression is inflammation of the intestine. There are a number of factors that have been associated with the onset and progression of the disease including diet, antibiotics, stress, and bacterial infections, but no putative cause has been found. As diet and the gut biome play a significant role in disease progression, we aimed to find commonalities in the gut microbiomes of Crohn’s patients, even when located in different geographical areas.

Keywords: hantavirus, meta-analysis, Crohn’s, transcriptomics, metagenomics, MAGs
ACKNOWLEDGEMENTS

I would like to express thanks to all who have supported me in my education here at Brigham Young. I would like to acknowledge the extensive expertise and efforts of my committee, especially my advisor Dr. Pickett, in helping me delve into new areas of research in which I had little prior experience. Next, I’d like to thank my family for their support of these many years, which have been fun but not easy. Finally, I’d like to thank all the friends I have made in the MMBIO department, especially the Pickett and Berges lab members for always brightening my day, and also to Alyssa Applegate for patiently helping me with training.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Hantavirus and Health Outcomes</td>
<td>1</td>
</tr>
<tr>
<td>Crohn’s Disease and Metagenomics</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>6</td>
</tr>
<tr>
<td>META-ANALYSIS OF TRANSCRIPTOMIC DATASETS QUANTIFIES DIFFERENTIAL GENE</td>
<td>6</td>
</tr>
<tr>
<td>EXPRESSION, AFFECTED PATHWAYS, AND PREDICTED DRUGS FROM HANTAVIRUS-</td>
<td></td>
</tr>
<tr>
<td>INFECTED HUMAN MATERIAL</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>7</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>7</td>
</tr>
<tr>
<td>METHODS</td>
<td>9</td>
</tr>
<tr>
<td>Data Acquisition</td>
<td>9</td>
</tr>
<tr>
<td>Data Processing</td>
<td>10</td>
</tr>
<tr>
<td>RESULTS</td>
<td>10</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER 3</td>
<td>27</td>
</tr>
</tbody>
</table>
LIST OF FIGURES
Figure 1-1: Overview of Hantavirus Replication in Host Cell ........................................................... 2
Figure 2-1: A Volcano Plot of Significance and Expression of DEGs in our Primary PBMC Analysis ... 14
Figure 2-2: A Volcano Plot of Significance and Expression of DEGs in the In vitro Analysis .......... 15
Figure 2-3: A Protein-Protein Interaction Network of the Significant In vitro Genes Generated from the STRING Database ........................................................................................................ 16
Figure 3-1: Gut Bacterial Composition of CD Patients .............................................................................. 33
Figure 3-2: Gut Bacterial Composition of Controls .................................................................................. 33
Figure 3-3: Top 15 Toxin Genes Found in Bins Generated by the Bin Refinement Module ............... 34
LIST OF TABLES

Table 2-1: Top 10 Significant Signaling Pathways During Primary PBMC Human Infection .................... 15

Table 2-2: All 11 Significant Signaling Pathways During In Vitro Human Infection ................................. 17

Table 2-3: Top 20 Therapeutics Predicted for the Primary PBMC Datasets Analyzed in This Study ...... 19

Table 2-4: Top 20 Therapeutics Predicted for the In Vitro Datasets Analyzed in This Study .................... 21
INTRODUCTION

Hantavirus and Health Outcomes

*Orthohantavirus* is a genus that belongs to the family *Hantaviridae* in the order *Bunyavirales*, meaning it has a three-segment negative-sense RNA genome[1]. While the genome is linear, each segment can form a panhandle structure through complementary base pairing of the 5’ and 3’ ends, which is then bound by the viral RdRP (RNA-dependent RNA polymerase) and nucleocapsid protein[2]. The entirety of each genomic segment is bound by nucleocapsid protein forming three unique nucleocapsid macromolecules in each virion[3]. The large (L) segment codes for the RdRP. The medium (M) segment encodes a glycoprotein precursor that is cleaved into the Gn and Gc subunits prior to dimerization into the functional glycoprotein[3]. The small (S) segment encodes the nucleocapsid (N) protein that binds to the viral nucleic acid. Some hantavirus species have been shown to have a functional nonstructural protein encoded on the reverse strand of the S segment as well, but it is largely uncharacterized. The main entry receptor for hantaviruses is the host b3-integrin, which transports the virus into the cell through cellular endocytosis[2](Figure 1-1). In addition, hantaviruses belonging to the new world clade rely on protocadherin-1, and old-world hantaviruses rely on CD55 as an entry receptor [4, 5].
The first member of the viral family *Hantaviridae* was originally isolated in 1978 by Dr. Ho Wang Lee during his search for the cause of an outbreak of Korean hemorrhagic fever. The causative agent was subsequently named the Hantaan virus, due to the outbreak’s proximity to the Hantaan river[6]. Within this viral family is the genus *Orthohantavirus*, which consists of viruses that are endemic within rodent populations and are commonly referred to as hantaviruses. These viruses do not cause disease within their rodent host; rather, they are able to maintain a persistent infection without becoming pathogenic or symptomatic[7]. However, as human populations and cities have expanded, there has been increasing contact with rodent species resulting in viral spillover and increased human infection[8]. Depending on

Figure 1-1: Overview of Hantavirus Replication in Host Cell. After hantavirus binds to integrin beta, it is endocytosed. After entry, hantavirus replicates its genome and creates viral components, packages itself, and escapes the cell[1, 2]. Note: Some new world hantaviruses assemble at the cellular membrane[1].
weather patterns, rodent populations may rapidly increase, resulting in a greater risk of hantavirus spread. This phenomenon generally causes hantavirus epidemics every few years throughout Asia and Europe.

However, despite the ever-present risk of disease, with cases reaching over 10,000 in China annually, no treatment has been shown to be completely effective in treating hantavirus infection, especially once viremia occurs[9]. One of the primary reasons for this lack of therapeutic options is that the virus has a 2-6 week incubation time between infection and symptoms, at which point infection is already well established and difficult to treat[10]. While tracing and identifying the initial infection is difficult, hantavirus infection is not transmitted human-to-human, but solely by exposure to rodents and rodent droppings[11].

Hantavirus infection in humans is known to cause two manifestations of disease: Hantavirus Cardiopulmonary Syndrome (HCPS) and Hemorrhagic Fever with Renal Syndrome (HFRS). Old-world hantaviruses, being those found natively in Asia and Europe, are responsible for HFRS, while new-world hantaviruses are responsible for HCPS. HFRS causes flu-like aches and fever, often followed by shock and renal failure which can lead to death[2]. Depending on the viral species, mortality can occur in 1-12% of cases(12). However, HCPS is much more deadly, due to the loss of barrier integrity in the heart and lung endothelium, causing death in about 40% of cases[12].

Hantaviruses target human endothelial tissue by binding to integrin beta-3, which is found in high quantities on most human endothelial cells[13] (Figure 1-1). When a human cell is infected, hantaviruses replicate without causing cytopathic effects or apoptosis[14]. Instead, host immune cells are recruited and attack the tissue, causing intercellular junctions to break down. This causes edema in the tissue and the symptoms of hantaviral disease[15]. It is widely accepted that a cytokine storm results in excessive T-lymphocyte activity, which is the etiological factor that drives HCPS and HFRS; however, recent studies have shown that other immune cells, such as neutrophils and dendritic cells, may play an important role in pathogenesis[16-18]. Interestingly, both old and new world hantaviruses are capable of infecting and
replicating in a broad range of tissues[13]. However, old world hantavirus infections typically only result in disease in the kidneys, and new world infections usually only result in disease in the lungs and heart. While there are some reported instances of hantavirus pathogenesis in other organs, the defining feature of HFRS and HCPS is disease in their respective primary target organ[13]. Researchers have not yet been able to determine the underlying molecular mechanisms that influence tissue-specific pathogenesis for hantaviruses. Part of this thesis work was to perform a meta-analysis of all publicly-available RNA-seq data of hantavirus infected tissue to identify genes and signaling pathways that previously have not been identified but may play a role in of host-based mechanisms of hantavirus pathogenesis. Chapter 2 will show how this methodology was used to determine underlying genes and cellular pathways associated with hantavirus infection both within primary cells and cultured cell lines.

**Crohn’s Disease and Metagenomics**

Crohn’s disease (CD) affects millions of people of worldwide, especially in developed countries[19]. Over 3 million people in the US have been diagnosed with this disease, and over the past 30 years, incidence rates of CD in the US are increasing[19, 20]. This has been attributed to unhealthy diet and increased use of antibiotics along with better sanitation[21, 22]. While there is no widely-accepted singular cause of disease incidence or progression of the disease, bacterial dysbiosis, as well as small intestinal bacterial overgrowth (SIBO), have been associated with severe disease[23-25]. Several bacteria including *Mycobacterium paratuberculosis* and *Helicobacter pylori* have been suggested in the past as possible causes; however, while the treatment of these pathogens can alleviate symptoms, it cannot be described as a cure[26-28].

Although patients with CD can be found around the world, most patients are of European descent[29, 30]. However, this observation is changing as incidence rates in Asian countries are steadily climbing[31, 32]. Given this shift and the increasing amount of publicly available data from studies performed in human patients in these countries, we have the opportunity to analyze CD at a more global
scale and to identify any underlying commonalities between different populations of CD patients. We used multiple shotgun metagenomic datasets of the gut microbiome collected from case and control patients in the United States to better characterize the patterns that were consistently observed among each group of patients.

This project used data from shotgun metagenomic sequencing data rather than using amplicon-based 16S sequencing. In 16S amplicon sequencing, after the DNA is extracted from the community of microbes, the 16S ribosomal RNA gene is isolated and amplified through PCR. Because this gene is highly conserved between different bacterial and archaeal species within a genus, this allows researchers to identify and quantify bacterial diversity within a population with relatively little sequencing needed. This is much cheaper and bioinformatically straightforward than normal sequencing and requires fewer sequencing reads to achieve excellent sequencing depth. The drawback to this approach is the inability to determine any additional qualities or attributes of any individual species. In addition, since only known 16S sequences are matched against a database of reference sequences, any novel microbe remains unidentified.

With shotgun metagenomic sequencing, the DNA from the community of microbes is processed and sequenced together without selecting for any particular locus. This generates substantially larger datasets and consequently requires much more time- and computationally-intensive bioinformatic processing to process, analyze, and interpret the data. The increased resources associated with this approach are required to obtain sufficiently high sequencing depth and to ensure adequate coverage of the sample. A shotgun metagenomics approach can identify both genes and novel species and can give insight into microbial interactions within a biological system. Thus, shotgun metagenomics can have a greater impact in healthcare settings when trying to understand the microbiome of human systems and the underlying mechanism(s) of infection and disease.
After sequencing data from a shotgun metagenomics sample is collected and processed for quality control, it must go through an assembly workflow. If normal read assembly is compared to assembling a puzzle, metagenomic read assembly would be like assembling puzzles from several hundred puzzles where all the pieces are mixed together. Naturally, this is computationally expensive, and only somewhat recently has the required computational capacity become sufficient to make this problem tractable. Reads are assembled into contigs, and then passed to a step known as binning. Contigs are grouped together based on relative abundance and other biological features such as GC content. Contigs grouped together into a bin are considered to be a draft genome of a single species. From there the contigs can be reassembled into a viable genome. A genome is considered good enough to work with and representative of a real organism if the bin is calculated to be at least 70 percent complete with less than 5 percent contamination. Chapter 3 uses this binning technique to identify species within our samples and identify potentially harmful genes within those bins to determine if there are significant differences in the gut biome between those with Crohn’s in the US and the controls. Chapter 4 is a description of code I wrote used to simplify and automate the necessary processes to create and characterize those bins.

CHAPTER 2

META-ANALYSIS OF TRANSCRIPTOMIC DATASETS QUANTIFIES DIFFERENTIAL GENE EXPRESSION, AFFECTED PATHWAYS, AND PREDICTED DRUGS FROM HANTAVIRUS-INFECTED HUMAN MATERIAL

**ABSTRACT**

We performed a meta-analysis that consisted of re-analyzing publicly available RNA-sequencing datasets across representative species within *Orthohantavirus* and across host cell types to identify differentially expressed genes associated with hantavirus infection. In addition, we determined signaling pathways that were significantly perturbed in human tissue, as well as predicted existing drugs that could be repurposed to potentially reduce the effects of acute hantavirus infection. The National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO) and European Nucleotide Archive (ENA) databases were searched to find datasets for this analysis, then processed using the Automated Reproducible Modular Workflow for Preprocessing and Differential Analysis of RNA-seq Data (ARMOR) to identify differentially expressed genes (DEGs). The results from this analysis were then processed using signaling pathway impact analysis (SPIA) to identify enriched signaling pathways. Finally, those data were processed using the Pathways2Targets program to identify existing therapeutics that target the impacted cellular pathways. Our results include affirming the roles of the previously identified interferon pathway and the *MYBL1* and *CCR7* genes, as well as newly identified genes, including *SLC27A3*, *TSEN34* and neuron associated *NOG* and *AMIGO1*. We predicted dinaciclib, alvodicib, and ruxolitinib, among many other drugs, that could be useful as potential HFRS and HCPS treatments. These data can be used in the design of future wet lab experiments to better characterize the human transcriptional response to infection, as well as to improve host-based diagnostics and/or therapeutics to reduce the severity of hantavirus infection.

**BACKGROUND**

*Orthohantavirus* is a viral genus that includes the hantaviruses that are pathogenic in humans. Members of this viral genus are transmitted by various rodent hosts throughout Asia, Europe, the Americas, and parts of Africa [8]. These hantaviruses are generally well-adapted to their rodent host and are unlikely to cause disease in that host. However, when humans come into direct contact with infected rodent vectors or their feces, virions can enter host cells and potentially cause a productive infection [14].
Once human infection occurs, hantaviruses are far more likely to cause severe disease, permanent injury, or even death. Human-to-human spread of hantaviruses generally does not occur, with the exception of some rare cases of Andes virus [33]. However, as metropolitan areas continue to expand, the threat of hantavirus spillover into humans is increasingly becoming a threat. This is especially true in eastern Asia, where thousands of cases of hemorrhagic fever with renal syndrome (HFRS) due to hantavirus infection are reported each year [8]. New world hantavirus disease, known as hantavirus cardiopulmonary syndrome (HCPS), is far less common, but has been reported to have up to a 50% fatality rate [34]. It was the new world hantavirus Sin Nombre virus that was responsible for the 1993 hantavirus outbreak in the Four Corners region of the western United States [35].

Like all members of the Bunyavirales family, an Orthohantavirus virion is composed of a segmented negative-sense RNA genome enclosed in a spherical capsid. Each of its three segments encodes reading frames that code for the RNA-dependent RNA polymerase (RDRP), glycoprotein, and the nucleocapsid (L, M, and S segments respectively) [1]. The glycoprotein precursor is cleaved by endogenous proteases to form a heterodimeric protein receptor [3]. The integrin β3 surface protein serves as the primary receptor for all infectious hantaviruses [36]. Past work has shown that hantaviruses do not directly cause cytopathic effects in their host; rather that the host immune response to infection causes the breakdown of endothelial integrity, leading to severe symptoms or death. To better understand the pathogenesis of hantavirus, it is essential to characterize the underlying host intracellular transcriptional responses. Currently, there are no FDA-approved treatments for hantavirus infection besides supportive therapy, but effective vaccines are being investigated for several strains of hantavirus [37]. However, vaccination may be difficult to administer due to the sporadic and relatively limited nature of many hantavirus outbreaks.

Hantavirus infections are characterized by the loss of endothelial barrier integrity leading to severe symptoms and death, as well as high levels of T-cell proliferation in affected tissue [38]. Interestingly, several prior experiments have shown that hantavirus infection causes no cytopathic effects
Instead, the highly elevated levels of cytokines along with T-cell levels in infected tissue indicate that host inflammatory responses are primarily responsible for the observed clinical signs and symptoms.

While several studies have examined the proteomics and health outcomes of infected individuals, relatively little has been done to transcriptionally characterize human infection. Furthermore, no studies have compared the host transcriptional response during hantavirus infection both in primary human peripheral blood mononuclear cells (PBMC) and in human cell lines in vitro. Evidence of differentially expressed genes (DEGs) across tissues would provide evidence that further research is required to understand how these DEGs affect differences in morbidity in human tissues. This study seeks to further examine how infection impacts host gene expression and intracellular signaling pathways by re-analyzing existing human RNA-seq data collected during hantavirus infection. We anticipate that this approach will enable us to better compare the DEGs across tissue types to predict more effective host-based therapeutics.

METHODS

Data Acquisition

A search of the Gene Expression Omnibus (GEO) database, hosted at the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/geo/), was performed in mid-2021 to find RNA-sequencing datasets for hantaviruses using terms including “hantavirus”, “Seoul virus”, “andes virus”, “sin nombre virus” and “Hantaan virus”. In addition, these same terms were used to find relevant datasets within the European Nucleotide Archive (ENA) database. The corresponding RNA-sequencing data for five GEO series were retrieved from the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) at NCBI: GSE158712, GSE161354, GSE133319, GSE133634, GSE133751[39-41]. Sequencing data with the experiment accession number PRJEB41624 was also downloaded from ENA and used in this analysis[42]. These datasets were generated from either human
cell cultures (endothelial cell lines HUVEC or A549) infected with Hantaan Virus (HNTV) or Puumala
virus (PUUV) which are two similar viruses that are causative agents of HFRS, or peripheral blood
mononuclear cells (PBMC) harvested from patients who were infected with Hantaan Virus (HTNV) and
currently experiencing HFRS. The endothelial cell datasets and the primary PBMC datasets were
processed separately to generate data. The endothelial cell datasets will be referred to as the in vitro
analysis, while the PBMC datasets will be referred to as the primary PBMC analysis. The SRA files were
downloaded and converted to fastq format using GNU wget and version 2.10.1 of the NCBI sratools

Data Processing

Much of the data processing was managed by the Automated Reproducible Modular Workflow
for Preprocessing and Differential Analysis of RNA-seq Data (ARMOR) [43]. Within this workflow the
data was processed as follows: quality control of the reads was performed using fastQC, adapters and
low-quality reads were trimmed using TrimGalore, reads were mapped and quantified to the GRCh38
human transcriptome using Salmon [44], and differential gene expression was calculated using edgeR
[45]. The list of DEGs from this workflow was subsequently analyzed with the Signaling Pathway Impact
Analysis (SPIA) algorithm [46], which uses pathway enrichment to identify cellular pathways using the
Biocarta, KEGG, NCI, Panther, and Reactome databases [47-50]. The SPIA output file from this analysis
was then analyzed with the Pathways2Targets program to identify existing therapeutics that target the
impacted cellular pathways [51] from the opentargets.org database [52]. The Search Tool for the Retrieval
of Interacting Genes/Proteins (STRING; https://string-db.org) database was used to identify and visualize
relevant protein-protein interactions [53].

RESULTS

We began by collecting the relevant fastq files for human-derived samples infected with
Orthohantavirus strains. In total, we included 73 samples from six datasets in this meta-analysis,
including 23 controls, and included all pertinent data and metadata that could be extracted from the NCBI-GEO and ENA databases [39-41]. Fastq files were divided into in vitro and primary PBMC groups for the analysis according to sample type and each group was separately input to ARMOR. The primary PBMC dataset consists of sequencing data from peripheral blood mononuclear cells (PBMC) of HFRS (Hantaan virus infected) patients, with relevant control fastq files. The in vitro dataset consisted of A549 and HUVEC cells infected with either Hantaan or Puumala virus, with relevant control fastq files.

*Primary PBMC Differentially Expressed Genes (DEGs)*

We next used the automated ARMOR bioinformatics workflow to calculate genes with significantly altered expression from the raw fastq files. We observed approximately 5900 significant DEGs in the primary PBMC dataset (Figure 2-1), which consisted of human PBMC samples (corrected p-value ≤ 0.05). This relatively high number of DEGs was expected since the infected material was being affected by the host immune response. Interestingly, four of the most significant 50 genes have been previously identified in earlier hantavirus studies including: *OLFM1, CCR7, KLRG1*, and *PTGDR2*—all of which were downregulated [54-57]. We observed the most significant upregulated genes found were *NUSAP1, KIF20A, CDC25C*, and *DEPDC1*. 
In Vitro Differentially Expressed Genes (DEGs)

We then wanted to perform a similar analysis on the in vitro dataset (cultured HUVEC and A549 infected cells). In this case, our analysis identified only 26 significant DEGs during infection including: *IFIT1, OAS2, CMPK2, IFIT3, IFI6*. We found that three genes were downregulated, while the remaining 23 significant genes were upregulated (Figure 2-2).
To better understand whether these gene products directly interacted with each other, we used the STRING database to retrieve protein-protein interaction data [53]. We found that this protein network revealed most of the proteins directly interact with each other, with two proteins not connected to the rest of the graph and a third protein not represented in the database (Figure 2-3.)

Figure 2-2: A Volcano Plot of Significance and Expression of DEGs in the In Vitro Analysis of Hantavirus Infection in Humans. Genes with a log₂ fold-change of less than |2| and/or an adjusted p value < 0.05 are represented by grey dots. Blue dots represent significantly downregulated genes, while red represent significantly upregulated genes.
Figure 2-3: A Protein-Protein Interaction Network of the Significant *In Vitro* Genes Generated from the STRING Database. Known interactions (edges) connect proteins (nodes). (Not pictured: AC090527.2 due to unknown function and interactions)

**Pathway Analysis**

After identifying the significant DEGs, we used them as input to the signaling pathway impact analysis (SPIA) algorithm, which performs robust statistical analyses to identify cellular pathways that were significantly enriched in DEGs [46]. We identified 79 enriched pathways in the primary PBMC dataset, largely associated with transcription, cellular repair, homeostasis, and the cell cycle, with a few associated with immune response (Table 2-1). In contrast, the *in vitro* dataset enrichment identified 11 pathways, all of which are associated with the antiviral response (Table 2-2).

<p>| Table 2-1: Top 10 Significantly Affected Signaling Pathways During Primary PBMC Human Hantavirus Infection. |</p>
<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Status</th>
<th>Corrected p-value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytokine-cytokine receptor interaction</td>
<td>Inhibited</td>
<td>4.96E-12</td>
<td>KEGG</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Activated</td>
<td>0.000115</td>
<td>KEGG</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>Activated</td>
<td>0.000276</td>
<td>KEGG</td>
</tr>
<tr>
<td>Carbohydrate digestion and absorption</td>
<td>Activated</td>
<td>0.000529</td>
<td>KEGG</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em> infection</td>
<td>Activated</td>
<td>0.00062</td>
<td>KEGG</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>Activated</td>
<td>0.006032</td>
<td>KEGG</td>
</tr>
<tr>
<td>Generic Transcription Pathway</td>
<td>Activated</td>
<td>7.52E-12</td>
<td>Reactome</td>
</tr>
</tbody>
</table>
Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3.

RNA Polymerase I Promoter Opening
Inhibited 5.01E-11 Reactome

Meiotic recombination
Activated 5.26E-11 Reactome

Table 2-2: All 11 Significantly Affected Signaling Pathways During *In Vitro* Human Hantavirus Infection.

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Status</th>
<th>Corrected p-value</th>
<th>Source Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza A</td>
<td>Activated</td>
<td>1.13E-08</td>
<td>KEGG</td>
</tr>
<tr>
<td>Herpes simplex infection</td>
<td>Activated</td>
<td>9.39E-06</td>
<td>KEGG</td>
</tr>
<tr>
<td>Measles</td>
<td>Activated</td>
<td>0.000117</td>
<td>KEGG</td>
</tr>
<tr>
<td>Pathway</td>
<td>Status</td>
<td>p-value</td>
<td>Database</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>RIG-I-like receptor signaling pathway</td>
<td>Activated</td>
<td>0.005427</td>
<td>KEGG</td>
</tr>
<tr>
<td>Interferon Signaling</td>
<td>Inhibited</td>
<td>3.68E-31</td>
<td>Reactome</td>
</tr>
<tr>
<td>Interferon alpha/beta signaling</td>
<td>Activated</td>
<td>1.21E-27</td>
<td>Reactome</td>
</tr>
<tr>
<td>Cytokine Signaling in Immune system</td>
<td>Inhibited</td>
<td>8.60E-18</td>
<td>Reactome</td>
</tr>
<tr>
<td>Antiviral mechanism by IFN-stimulated genes</td>
<td>Inhibited</td>
<td>5.06E-10</td>
<td>Reactome</td>
</tr>
<tr>
<td>ISG15 antiviral mechanism</td>
<td>Inhibited</td>
<td>5.06E-10</td>
<td>Reactome</td>
</tr>
<tr>
<td>Negative regulators of RIG-I/MDA5 signaling</td>
<td>Activated</td>
<td>5.49E-06</td>
<td>Reactome</td>
</tr>
<tr>
<td>RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways</td>
<td>Inhibited</td>
<td>7.24E-05</td>
<td>Reactome</td>
</tr>
</tbody>
</table>

*Potential Therapeutic Prediction*
We next wanted to predict existing drugs that could be repurposed as host-based therapies that limit viral replication. We consequently used the enriched pathway data generated by SPIA as input to the Pathways2Targets algorithm [51]. This software identifies known drug targets in the significant pathways, then searches the public opentargets.org database to generate a list of potential therapeutics that could affect each pathway. This method was justified since reversing the phenotype of the significantly impacted pathways should reduce viral replication and/or pathogenesis. We sorted the results of this analysis to prioritize therapeutics that were predicted to affect multiple pathways. We intentionally selected this approach to maximize the potential effect of each drug by simultaneously interfering with multiple intracellular processes that the virus uses to replicate. We generated predictions for the top 20 drugs for the primary PBMC and *in vitro* datasets (Table 2-3, Table 2-4). The top hits for the primary PBMC analysis included dinaciclib, alvocidib, and roniciclib.

Table 2-3: Top 20 Therapeutics Predicted for the Primary PBMC Datasets Analyzed in This Study.

<table>
<thead>
<tr>
<th>Predicted Therapeutic Name</th>
<th>Number of Target Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>DINACICLIB</td>
<td>33</td>
</tr>
<tr>
<td>ALVOCIDIB</td>
<td>33</td>
</tr>
<tr>
<td>RONICICLIB</td>
<td>33</td>
</tr>
<tr>
<td>SELICICLIB</td>
<td>33</td>
</tr>
<tr>
<td>AT-7519</td>
<td>33</td>
</tr>
<tr>
<td>Drug Name</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------</td>
</tr>
<tr>
<td>MILCICLIB</td>
<td>33</td>
</tr>
<tr>
<td>ATALUREN</td>
<td>33</td>
</tr>
<tr>
<td>ELX-02</td>
<td>33</td>
</tr>
<tr>
<td>MT-3724</td>
<td>33</td>
</tr>
<tr>
<td>UCN-01</td>
<td>24</td>
</tr>
<tr>
<td>PANOBINOSTAT LACTATE</td>
<td>23</td>
</tr>
<tr>
<td>ROMIDEPSIN</td>
<td>23</td>
</tr>
<tr>
<td>PANOBINOSTAT</td>
<td>23</td>
</tr>
<tr>
<td>BELINOSTAT</td>
<td>23</td>
</tr>
<tr>
<td>ENTINOSTAT</td>
<td>23</td>
</tr>
<tr>
<td>ABEXINOSTAT</td>
<td>23</td>
</tr>
<tr>
<td>VORINOSTAT</td>
<td>21</td>
</tr>
<tr>
<td>TACEDINALINE</td>
<td>19</td>
</tr>
<tr>
<td>AMG-232</td>
<td>17</td>
</tr>
</tbody>
</table>
Table 2-4: Top 20 Therapeutics Predicted for the *In Vitro* Datasets Analyzed in This Study.

<table>
<thead>
<tr>
<th>Predicted Therapeutic Name</th>
<th>Number of Target Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUXOLITINIB</td>
<td>7</td>
</tr>
<tr>
<td>BARICITINIB</td>
<td>7</td>
</tr>
<tr>
<td>RUXOLITINIB PHOSPHATE</td>
<td>7</td>
</tr>
<tr>
<td>TOFACITINIB</td>
<td>7</td>
</tr>
<tr>
<td>TOFACITINIB CITRATE</td>
<td>7</td>
</tr>
<tr>
<td>FILGOTINIB</td>
<td>7</td>
</tr>
<tr>
<td>UPADACITINIB</td>
<td>7</td>
</tr>
<tr>
<td>RONTALIZUMAB</td>
<td>6</td>
</tr>
<tr>
<td>SIFALIMUMAB</td>
<td>6</td>
</tr>
</tbody>
</table>
DISCUSSION

The purpose of this meta-analysis was to better characterize and compare the *in vitro* and primary PBMC human intracellular transcriptional response to infection with various hantaviruses. Specifically, we identified unique sets of DEGs in both comparisons that have been identified previously, as well as
novel DEGs that could provide additional knowledge into the underlying mechanisms of viral pathogenesis. We determined statistically significant signaling pathways that were enriched with the identified DEGs for each set of samples, and subsequently used these pathways to predict potential therapeutics that could be relevant to natural infections.

We believe that this analysis was somewhat hindered by the relatively small number of samples in the public datasets. Specifically, the public data that we used came from HTNV-infected tissue, rather than a variety of several hantaviruses. This is somewhat unsurprising due to various logistical reasons including the difficulty and danger of culturing hantavirus in a laboratory setting, or a perceived lack of interest in or access to hantavirus. Such studies can be quite useful in identifying potential causes or compounding factors in HFRS and HCPS. Other hantaviruses, such as Andes virus and Sin nombre virus, may cause far fewer infections each year, but infections by these pathogens have a higher mortality rate. Much research is still needed to determine whether the differential expression patterns vary between species of hantaviruses and between infected human tissue types, but this study attempts to address hantaviruses as a whole in an attempt to identify common expression patterns between different viral species. Our findings could potentially be applied across multiple hantaviruses.

We ran our data as two separate sets of data, differentiating between infected primary PBMC taken from HFRS patients, and cells that were both cultured and infected in vitro in the lab. This was done to analyze different components of the infection, both in the target cell type (endothelial cells) isolated from the immune response and immune cells directly involved in the response. The cell lines from the studies used in this investigation, human umbilical vein endothelial cells (HUVEC) and A549 (alveolar adenocarcinoma derived cells) are both commonly used models for infection and are a representative model for identifying genes that are modulated as a direct response to infection, and not due to any damage from the immune response. The primary cells are, on the other hand, not the primary target of infection, although in more severe cases of HFRS, CD8+ T-cells, monocytes, and dendritic cells may be potential targets for infection[58, 59]. The purpose of processing this data was to find underlying
genes and pathways associated with the immune response to predict potential contributing factors to the extreme “cytokine storm” response to hantavirus infection.

In the primary PBMC comparison, we found that of the top 10 differentially expressed genes, only the NUSAP1 gene, which plays a role in microtubule organization [60], was upregulated. In contrast, MYBL1 was the top hit, which is a protooncogene that is predicted to be associated with promoting piRNA and PIWI expression, was found to be strongly downregulated[61]. PIWI-interacting RNA, or piRNA, are small RNAs that play a role, along with PIWI proteins, in silencing mobile DNA elements in the genome [62]. A recent study found that some PIWI pathways may be useful in antiviral defense against viruses in insects, but nothing has been investigated in mammals, although these pathways have been proposed to exist [63, 64].

As stated above, of the top 100 genes identified in primary PBMCs, only OLFM1, CCR7, KLRG1, and PTGDR2 have been identified in previous hantavirus publications. OLFM1, which produces olfactomedin and was found to be associated with hantavirus in a previous study, is also found to be highly expressed in neuronal tissue and may aid in nervous system development; however, relatively little is understood about many of its functions in other cell types [65]. Further, while it was identified as downregulated in a prior study, we found it to be upregulated in our meta-analysis, but its potential role in hantavirus has not been studied. Interestingly, while this gene is expressed in monocytes, its function is unknown and warrants future investigation [66]. Additionally, three of the top 10 genes found in this analysis, ZNF365, NOG and AMIGO1, regulate proper neuronal development[67-70]. Hantavirus rarely affects the nervous system and has not been observed to directly infect nervous tissue, but these genes may also play a role in the development of other cell types that can be affected during hantavirus infection of PBMCs. While not fully understood, they may have a useful function when expressed in activated PBMCs, or they may share common transcriptional activators with other genes that are activated during hantavirus infection.
**CCR7**, or C-C Motif Chemokine Receptor seven, is upregulated in lymphocytes and monocytes during a variety of viral infections. A previous study showed that **CCR7** induces monocyte egress during hantavirus infection, and therefore a decrease in the number of **CCR7**-expressing monocytes occurs in the bloodstream [55]. While this may account for the decrease in overall expression, it is unknown if lymphocyte expression of this gene is modulated by hantavirus infection, as has been indicated by our meta-analysis. Similarly, **KLRG1**, an NK cell receptor, is upregulated in T-cells generally during viral infection, but previous studies have not indicated whether or not it is modulated by hantavirus infection[56]. Our data indicates that it is downregulated during infection with hantavirus, which is novel, especially when current research shows that NK cells actively proliferate during hantavirus infection[71]. Further investigation is needed to determine if hantaviruses can actively infect NK cells, which would be a novel addendum to hantavirus research, or if the downregulation is due to signals from other cytokines.

**PTGDR2**, or the prostaglandin D2 receptor, promotes inflammation during a viral infection [72]. While no studies of this gene in association with hantavirus have been performed in humans, a 2014 study found that deer mice, a natural non-symptomatic host for the virus, upregulate this gene during active infection of Andes virus, a species of hantavirus [57]. Our analysis showed that this gene is strongly downregulated in humans during infection of HTNV, another hantavirus. Although our observation could be due to a difference in viral species or strain, the gene product may play a role in human pathogenesis of the virus that is not seen in murine species, and likely warrants additional investigation, as it may contribute to the excessive cytokine signaling seen in infected tissue.

The most significant upregulated genes primary PBMCs were **NUSAP1, KIF20A, CDC25C, and DEPDC1** and were associated with cellular division, especially in rapidly dividing cells, which is expected during active viral infection and tissue repair [73-77]. In addition, **CDC25C** has been shown to play a role in inhibiting viral replication of HSV, however there is no research indicating whether this is true in other viruses such as hantavirus [78].
In vitro, of all the DEGs, only SLC27A3, AC090527.2 (a novel, uncharacterized gene not shown on Figure 2-3), and TSEN34 were not associated with interferon stimulation and are not part of the antiviral response. SLC27A3, which is a Very Long-Chain Acyl-CoA Synthetase, is both a synthetase and a transport protein. It has a role in brain development but has not previously been associated with viral pathogenesis or the immune response [79]. The tRNA Splicing Endonuclease 34 (TSEN34) gene, and novel gene AC090527.2 similarly have no immune associated function. Finally, Immunoglobulin Heavy Constant Gamma 2 (IGHG2), is shown to be strongly downregulated. This is unexpected since the endothelial cells used in this analysis normally do not produce immunoglobulin. This may be a result of the some of the cell lines being immortalized, or as an artifact of the analysis. However, at least one prior study has shown that primary endothelial cells can produce antibodies in vitro [80].

The pathway analysis revealed a much wider variety of genes and perturbed pathways In the primary PBMC dataset as compared to the in vitro dataset. We believe that one of the primary causes of this difference is likely the higher content and complexity of cellular signals that are present in the primary PBMC dataset. As the putative cause for the disease is a cytokine storm, we expected to see cytokine signaling along with the interferon response in addition to repair pathways that would not be seen in vitro [16]. Of interest in the in vitro enriched pathways are the inhibition of the RIG-I/MDA5 signaling pathways associated with viral detection. Although they were inhibited, the interferon pathway downstream from them was still upregulated. This observations could mean another cellular pathway inducing the interferon response, such as through TLR4 or IRF3, could play a role [81-83]. This is in accordance with much of the current research, which shows that RIG-I is a part of the interferon response in hantavirus infection and that glycoproteins from hantavirus can inhibit RIG-I activity, but there is uncertainty on which pathways are essential to the activation of interferon stimulated genes (ISGs) during a hantavirus infection. Our data indicates that while the RIG-I complex may certainly play a role during infection, but it is not essential to the downstream activation of interferon.
Using data from the pathway analysis, medications identified for the primary PBMC dataset generally treat rapidly dividing and surviving cells, such as in cancer. Specifically, the drugs we predicted would affect the highest number of primary PBMC pathways in this analysis included those that target cyclin-dependent kinases and JAK kinases such as dinaciclib, alvocidib, and roniciclib. All three of these therapeutics have shown promise as a host-based therapy that limits viral replication in other pathogens [84-87], but none have been identified or tested for hantaviruses previously. Interestingly, our approach ranked these drugs higher than other potential therapeutics on the list, which have previously been used as host-based therapies that limit viral replication, such as vorinostat [88-90]. In contrast, the therapeutic prediction analysis from the in vitro dataset identified baricitinib and interferon derivatives, which are commonly used for viral infections [91-93]. As of now, there are no effective treatments for hantavirus infection, and the current suggested remedy is supportive therapy such as bed rest and IV fluids [37]. We believe that testing a subset of these drugs in a laboratory setting would be justified, and that laboratory validation of these findings could improve treatment options for patients in the clinic.

The primary purpose of this study was to analyze the existing public RNA-seq datasets to improve our understanding of the response of human cells during hantavirus infection and to predict existing drugs that could be repurposed as host-based therapeutics. Specifically, we identified seven genes that are novel to the human response to hantaviruses, as well as multiple signaling pathways and drugs that could be repurposed as therapeutics. Such treatments could benefit patients by reducing the occurrence of HFRS and HCPS. The results of this study remain speculative until they can be validated in vitro, however we hope this will guide future research that is needed to understand how hantaviruses impact gene expression in the human cell and to improve the identification, development, and testing of effective treatments.
CHAPTER 3

METAGENOMIC META-ANALYSIS FOR IDENTIFYING COMMONALITIES IN GUT BIOMES BETWEEN CROHN’S DISEASE PATIENTS IN MULTIPLE POPULATIONS

The content from this chapter is currently being prepped for submission for publication

BACKGROUND

Crohn’s disease is becoming an increasingly prevalent health problem in developed nations. In the United States alone, it is estimated that about 1 in 400 adults has CD, and that number is expected to rise if current trends persist[20, 94]. This results in a roughly 10 billion dollar burden on the US economy, and those with severe disease generating as much as 9 times more in medical costs as compared to those in remission [95]. A recent study found that more than one in four Crohn’s disease patients were dependent on steroids to manage their condition, and about half depend on thiopurine medications[96]. Both of these classes of medications have a well-documented history of use in the medical field, and the potential dangers of their use are well known. They can be especially risky for developing children, but in severe disease they may be necessary [97, 98]. Despite these advancements in diagnosis and treatment, about one fourth of CD patients may be hospitalized each year [99].

The introduction of biologics, i.e. monoclonal antibody medications, has significantly reduced the burden of disease [100, 101]. However, even with these advanced medications, identifying root causes of initiation and progression of Crohn’s disease is a surer route to finding a cure. Environmental factors play a large role in the development and diversity of the gut biome, and by extension the severity of CD symptoms. Diet, water sourcing, radiation, and medicine—especially antibiotics—are a few among a plethora of environmental sources that constantly impact the microbiome health of an individual. While these factors may differ between communities, or even between people, we hope to look at a national scale and compare CD patients in the United States. The western diet is oversaturated with fatty foods and peoples of the United States rely heavily on medications and antibiotics. While these are generalizations,
by analyzing large datasets we use these cultural, dietary, and other differences among these populations to identify differences in gut diversity generally. Beyond this, we identified underlying similarities that exist between CD patient populations in the US that we hope represent general patterns of CD patient microbiomes.

To get a wide degree of sampling, shotgun metagenomic data from five projects across the United States were reprocessed using the metaWRAP metagenomic pipeline. Each dataset was processed separately to produce high quality bins. A bin represents a hypothetical genome assembly of a single organism from within the sequenced microbial community, and individually contains the previously generated contigs that have been sorted, or “binned” together. This binning process is informed by the relative abundance of reads and contigs across all the submitted samples. A high-quality bin would be one calculated to be over 70 percent complete with less than 5 percent contamination, which are the standardized criteria that the ENA considers to be “real”. We then found the relative abundance of these bins across all samples in the study to identify patterns of bacterial presence both within populations and bins associated with CD patents. In addition to finding and quantifying relative abundance, we classified and annotated each bin. Some bins represent previously unidentified bacterial species and may require isolation in a wet lab setting for more comprehensive fully identification and characterization.

RESULTS

Between all datasets, over 1000 high-quality bins were produced by the metaWRAP bin refinement module. Although these bins aren’t specifically designed for colony composition studies, we can get an idea of the composition by using the bin classification. The specificity of the classification, potentially down to the strain, depends on the bin’s completion. With a standard of at least 70 percent, most of the bins could be at least identified within a phylum, though many could be identified within a family or genus.
Based solely on the bacterial phylogeny, the controls of the United States population are remarkably similar to those with CD (Figure 3-1, Figure 3-2). Indeed, a phylogenetic-based approach to understanding metagenomics is limited, but deeper inspection allows for more specific understanding of potential underlying patterns, especially when compared to 16S amplicon-based sequencing. Five of the bins that were exclusive to the Crohn’s set have been identified as *Fusobacterium*, of which two are *Fusobacteria ulcerans*. The presence of these bacteria has been noted in cases of ulcerative colitis, which is another irritable bowel syndrome (IBD) and is similar to CD, but it has only been weakly tied to potentially playing a role in CD[102, 103]. We calculated percentages, rather than using raw numbers as a normalization method since more bins were generated from the Crohn’s population. Overall, there were many more samples used to create bins, so this observation was expected.
Figure 3-1: Gut Bacterial Composition of CD Patients. Estimations are made based classification output from taxator-tk of refined bins generated from samples collected across studies.

Figure 3-2: Gut Bacterial Composition of Controls. Estimations are made based classification output from taxator-tk of refined bins generated from samples collected across studies.
Next, we looked at the annotations generated by prokka (Figure 3-3). Using a custom script, we retrieved all toxin related genes from the dataset and created an overall count of each toxin gene. Again, due to having a larger dataset, more genes are present in the Crohn’s set, however when compared as a percentage of total toxin genes identified, it appears that the control has more toxins present than Crohn’s patients. This observation was rectified when we excluded the large number of less-common genes that were only identified in the CD dataset.

![Top 15 Toxin Genes Found in Bins Generated](image)

Figure 3-3: The Top 15 Bacterial Toxin Genes Found in Bins Generated by the Bin Refinement Module. Percentage represents percentage of the number of a toxin gene found in comparison to all toxin genes found within the bins of that condition (Crohn’s or Control). The differences were found not to be statistically significant (p < 0.05).

DISCUSSION

Our analysis revealed several interesting patterns from this data. First, is that the bin diversity is not indicative of disease status. Bins that reached over 70 percent completion were included, so more uncommon bacterial strains in the gut were, the nature of the study, not included. Another reason for this might be that diet is masking some of the underlying markers of the disease state. In concurrence with current research, if the gut bacterial dysbiosis caused by western diet or environment is similar to one that
is indicative, or even partially causative, of disease, this could explain prevalence in western countries[27].

A recent study found that Actinobacteria, Fusobacteria, Firmicutes and Bacteroidetes are all higher abundance in CD patients, but our data only showed an increase in Fusobacteria and Bacteroidetes[104]. It was interesting to note the presence of Fusobacteria ulcerans, as while several studies have analyzed the role of other Fusobacteria species in CD, F. ulcerans remains largely uncharacterized in CD. There are several studies in current literature that indicate that Mycobacterium avium subspecies paratuberculosis may be a causative agent of CD; however, it is a common pathogen and most likely cannot explain all cases of CD [28, 105]. In addition, elimination of the pathogen does not effectively reverse the effects of the disease completely in those that are affected [106]. However, this does not exclude the possibility that multiple bacteria contribute to the induction and persistence of severe disease [107, 108].

Additionally of note is that while CD progression is associated with bacterial toxin in serum, these data indicate that the presence of any toxin gene alone is not indicative of disease state [109]. Instead, our data does not show any significant difference in the quantity or diversity of toxin genes between the CD patients and the control. Either a meta-transcriptomic or a meta-proteomic analysis would more definitively indicate if these genes have a change in expression. TabA, the most highly present toxin, is associated with biofilm production and those with Crohn’s have a high amount of E. coli generated biofilm in the gut and become particularly adherent and worsen disease symptoms.

One of the weaknesses of our study is that it could not differentiate between active flaring CD and CD in remission. The gut biome changes based on the activity of the disease, although it is uncertain if the reverse is true, if the gut biome can independently induce remission or flare states[110]. However, we wished to identify pathogens that were present in all CD patients regardless of disease state. Further, we could not definitively quantify any particular gene from bins alone, but instead bins are more useful
for understanding bin diversity and gene detection. However, using heat mapping of bin quantity per sample could elucidate if particular bins are enriched in the samples.

In the process of this study, terabytes of data were produced. While we were interested in toxins, any gene or series of genes could be investigated. An example of this would be biofilm production or protein products that are well recognized PAMPs that perhaps increase inflammation. Further, several of the generated bins have high predicted completion and low contamination percentages, but were unable to be sorted into any phylum. This requires searching extensively through the genes and making classification predictions based on predicted genes. This search is likely to identify previously uncharacterized bacteria, whose role has not been studied in the context of gut health.
ABSTRACT

Generating high-quality genome assemblies of complex microbial populations from shotgun metagenomics data is often a manually intensive task involving many computational steps. SnakeWRAP is a novel tool, implemented in the Snakemake workflow language, to automate multiple metaWRAP modules. Specifically, it wraps the shell scripts provided within the original metaWRAP software, within Snakemake. This approach enables high-throughput simultaneous assembly and analysis of multiple shotgun metagenomic datasets using the robust modular metaWRAP software. We expect this advancement to be of import in institutions where high-performance computing infrastructure is available, especially in the context of big data. This software tool is publicly available at https://github.com/jkrapohl/SnakeWRAP

INTRODUCTION

As sequencing technology has become cheaper and more readily accessible, the need for the increased computational capacity to process these data has become apparent. In particular, high-throughput sequencing has been particularly useful when applied to the field of metagenomics. Substantial effort has been devoted to developing software and computational pipelines, such as MetaWRAP, which cater to this growing area of research. MetaWRAP is a modular wrapper that combines many of the necessary tools to process reads, create bins, and visualize data within a robust modular design (Uritskiy, DiRuggiero, & Taylor, 2018). The primary limitation arising from this design
is the inability to automatically scale its usage to massive datasets. However, its unique code that produces exceptionally high quality bins from bins generated from other binning programs is invaluable. Snakemake is a widely-used Python based workflow management system that automates repetitive tasks, allowing software processes to be both scalable and reproducible (Mölder et al., 2021). By integrating the original MetaWRAP processes into Snakemake, the customizable modular nature of MetaWRAP can be preserved even when automatically processing large datasets through a single workflow. Our SnakeWRAP software automates the tasks performed within MetaWRAP, allowing for individual modules to be toggled on and off using Snakemake-defined rules.

METHODS

Implementation

As this tool makes use of Snakemake, individual steps of the workflow are broken into rules, many of which can be toggled [111]. To do this, Snakemake requires a YAML configuration file as an input to determine which steps are specified for the run, as well detailed parameters such as which assembly tool to use. This configuration file also specifies the location of a metadata file, which contains a list of files associated with the run. An example of both the YAML and metadata file are included in the Github repository. SnakeWRAP can be submitted for scheduling from the command line by using the submission script also found within the repository. Running the script requires the complete installation of both Snakemake and METAWrap, as well as all of their dependencies. It is recommended that Snakemake and METAWrap be installed using the latest version of Miniconda, following instructions found on their corresponding Github repositories [43, 111].

Operation

The advantage of using SnakeWRAP over the base metaWRAP program is the ability to appropriately support largescale job sizes while requiring less user input by automating the entire process. This enables jobs to be better organized and allows improved consistency in outputs. We recommended
that this tool be run on cluster computing or within a high-performance infrastructure for larger jobs. In such cases, we have found that a minimum of 100 GB of RAM across 24 CPUs is sufficient, with larger jobs requiring additional resources. The rules found within the Snakemake file correspond to different steps within the original metaWRAP software. Some initial required steps include read quality control, assembly, binning, and bin refinement and reassembly. However, we have provided toggle settings for memory-heavy modules that generate figures, which are Blobology, Kraken, bin quantification, bin annotation, and bin classification. This design reduces resource usage by only generating the desired figures and allows the modification of this setting to create these figures at a later time by toggling the module on in the YAML file, deleting any outputs created for that module, and resubmitting the job. Snakemake will skip any rules already completed and only run the required rule(s).

Use Cases

As stated above, this tool works best with a high-performance computing infrastructure. We anticipate that the most effective use of this tool is applying it to big metagenomics data, such as meta-analyses of existing datasets or datasets with many samples. To assess the capabilities of this tool, the publicly available samples SRR13296364 and SRR13296365, accessed from the NCBI Sequence Read Archive from the study SRP299130, were run together successfully. We then ran a larger set of files from the SRP257563 study through the metaWRAP pipeline using this software. We found that this tool will run as expected as long as sufficient memory is provided and all dependencies are installed correctly. Inputs include sequencing files in fastq format, a metadata file listing all fastq filenames, and a YAML file describing paths for all file sources and destinations. Outputs include high-quality genomic bins, as wells as a number of generated figures such as a blobplot, heatmap, and kronogram.

DISCUSSION

While the processing of metagenomics datasets is untractable for most personal computers, researchers with access to high-performance computing infrastructure can take full advantage of this
software. The core functions found within the original MetaWRAP include read quality control, assembly, and binning are required for the analysis and cannot be toggled off. These functions include several refinement steps that are unique to MetaWRAP, which allow it to create higher-quality bins than existing stand-alone programs [112]. Our design enables the user to decide whether computationally-expensive modules, such as Kraken and Blobology, that assist with figure generation can be skipped.

Snakemake automatically generates a directed acyclic graph (DAG) to order tasks, track the progress of each task for each sample, and eliminate duplicate tasks for the same sample [111]. This is vital to efficiently processing and analyzing large datasets, as jobs can fail due to insufficient memory or timing out. An advantage of implementing this workflow in Snakemake is the ability to resume the workflow from the same step that was underway if a process fails. The input paths, output paths, and parameters for each job are assigned by the end-user within a configuration file. This file is read by Snakemake to prevent data from being incorrectly assigned or lost and to facilitate reuse and customization of the workflow. We anticipate that this advance in automating data processing will facilitate the generation, analysis, and re-analysis of larger datasets in the future.
CONCLUSION

In the microbiological and health fields of research, computation is becoming increasingly essential. In this thesis two computational approaches are described, transcriptomics and metagenomics. While they are but two of countless sequence-based tools that have been developed, they have practical uses in the study of human health. Transcriptomics has allowed us to understand the human response to stimuli of interest, in this case being hantavirus infection, and using that to identify potential routes of research. In this study, we identified genes, pathways, and drugs that have the potential to treat infection, and would be worth investigating. With these data, future studies can narrow their focus on specific interactions when developing novel treatments and investigating underlying mechanisms of Hantavirus infection.

Outbreaks of hantavirus infection are sporadic, but can be incredibly dangerous, with death tolls of over 100,000 people per year [113]. Several vaccines are under investigation, but none of them has seen great efficacy or range of viral species in part due to the difficulty to have widespread vaccinations in impoverished areas. [10, 37]. Despite this, hantavirus remains a lower priority due to its inability to be transmitted between human hosts. However, with the recent Covid-19 pandemic, the threat of a novel zoonotic virus is now regarded as much more of a threat. Certainly, Orthohantavirus is well-established in rodent populations around the world and will likely continue to be a threat to human health. By using and building upon identified potential therapeutics, we can lower current mortality rates and be prepared in the event of an outbreak or epidemic.

The gut microbiome has been described as the “hidden organ” and remains neglected as a component of human health due to the difficulty in understanding its function and role in disease. Metagenomics allows us to approach gut microbial biomes and identify patterns that exist within the community. In particular, shotgun metagenomics, or whole genome metagenomics, has only become possible due to recent advances in computational power. Because of this, a rich new layer of understanding can be extracted from metagenomics sequencing without any need for bacterial isolation or
16S amplification. While we identified bacterial abundance and the presence of harmful bacterial genes, any community could be studied and lead to the discovery of new genes and species. Through our work, we have generated terabytes of data to analyze. We could build upon this further by including more control datasets, or by including other gut disorders in the analysis, such as ulcerative colitis or colon cancer. Such an approach could provide interesting comparisons and contrasts between gut disorders. Future work can better characterize the bacterial role in disease, which can then be translated into the development of new treatments, and medication that can be better tailored to the individual.

Overall, we believe that our findings improve our mechanistic understanding of these diseases and may contribute to the future development, testing, and availability of novel treatments to improve human health.
FUTURE DIRECTIONS

In my time as a Master’s student, I found time to be the most limiting factor to my research, as only so much can be done over the course of two years. There are several components of my research that could be expanded upon and further developed if given the time. Within the hantavirus project, I would test to see if any of the predicted drugs had the potential of limiting infection and replication of hantavirus within cell culture, and if successful, collaborate with another university to test that drug within a hamster model. Next, I would test to see how hantavirus transcriptional modulation occurs within different cultured primary cells. We could compare RNA isolated from primary lung tissue, kidney tissue, cardiac tissue, and even blood cells during an *in vitro* infection. By performing RNA-seq we could potentially predict which genes and pathways contribute to a symptomatic or asymptomatic phenotype during hantavirus infection and the potential for more targeted treatments.

If I were to continue with the Crohn’s disease project, there are several avenues I would like to explore. First, I would expand my datasets to include as many countries as possible as part of the analysis to identify commonalities between Crohn’s disease patients on a more global scale. In addition, I would want to do a more comprehensive review of *F. ulcerans* as well as do a comparison of our assembled bins against the NCBI reference genome to see if there are genes of interest that are present only in the assembled bins. Finally, there are several bins that were well-assembled with little contamination but were not classified. They most likely represent novel strains of bacteria and would be of interest in future gut microbiome projects moving forward. More importantly, they could be undiscovered contributing factors in Crohn’s disease that have yet to be isolated and cultured. To identify them, we would have to through the genome and manually identify genes using BLAST (Basic Local Alignment Search Tool) [114]. Additionally, we could use a common or universal bacterial gene, such as a sigma factor or 16S sequence to place the bins within a tree alongside reference genomes to see how similar they are to known bacteria. In this same way, other genes, besides toxins, could be analyzed in relation to Crohn’s disease. There are still several potential ways to expand these projects, whether *in silico* or *in vitro*. 
REFERENCES


Appendix A: Secondary Author Publication

The following is a copy of “Preprocessing of Public RNA-Sequencing Datasets to Facilitate Downstream Analyses of Human Diseases”, of which I was a contributing, but not primary author. The citation for this article is:

Data Descriptor

Preprocessing of Public RNA-Sequencing Datasets to Facilitate Downstream Analyses of Human Diseases

Naomi Rapier-Sharman, John Krapohl, Ethan J. Beausoleil, Kennedy T. L. Gifford, Benjamin R. Hinatsu, Curtis S. Hoffmann, Makayla Komer, Tiana M. Scott and Brett E. Pickett

Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA;
naomirapier.sharman@gmail.com (N.R.-S.); jkrapohl@gmail.com (J.K.); ebeausoleil@gmail.com (E.J.B.);
hinatsu.ben@gmail.com (B.R.H.); hoffmann@byu.edu (C.S.H.);
komer135@gmail.com (M.K.); tianamnna29@gmail.com (T.M.S.)
* Correspondence: brett@byu.edu

Abstract: Publicly available RNA-sequencing (RNA-seq) data are a rich resource for elucidating the mechanisms of human disease; however, preprocessing these data requires considerable bioinformatic expertise and computational infrastructure. Analyzing multiple datasets with a consistent computational workflow increases the accuracy of downstream meta-analyses. This collection of datasets represents the human intracellular transcriptional response to disorders and diseases such as acute lymphoblastic leukemia (ALL), B-cell lymphomas, chronic obstructive pulmonary disease (COPD), colorectal cancer, lupus erythematosus; as well as infection with pathogens including Francisella tularensis, bartonella henselae, influenza A virus, Middle East respiratory syndrome coronavirus (MERS-CoV), Streptococcus pneumoniae, respiratory syncytial virus (RSV), severe acute respiratory syndrome coronavirus (SARS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We calculated the statistically significant differentially expressed genes and Gene Ontology terms for all datasets. In addition, a subset of the datasets also includes results from splice variant analyses, intracellular signaling pathway enrichments as well as read mapping and quantification. All analyses were performed using well-established algorithms and are provided to facilitate future data mining activities, wet lab studies and to accelerate collaboration and discovery.


Dataset Licence: CC-BY.

Keywords: transcriptomics; RNA-sequencing; autoimmune diseases; cancer; pathogens; bacteria; viruses; data preprocessing

1. Summary

The number of publicly available RNA-sequencing (RNA-seq) datasets is increasing, and we expect this momentum to continue. However, comprehensive results from statistical analyses such as differential gene expression are not consistently available in public transcriptomics repositories such as the Gene Expression Omnibus (GEO). Additionally, in the subset of cases where multiple differentially expressed gene (DEG) lists from different experiments are available, directly comparing them is difficult due to the differing parameters, assumptions, and biases present within each of the preprocessing algorithms (e.g., trimming, mapping, quantification; see sampling of pipelines and methods) [1–5]. A survey of the literature confirms that transcriptomic preprocessing pipelines utilize a variety of underlying statistical models, further complicating comparison between two datasets processed by different pipelines.

Our motivation for publishing these preprocessed public datasets was to make the results from these computational methods accessible to facilitate hypothesis generation, as
well as for subsequent analysis and interpretation by researchers. In particular, to aid those who may not have the necessary computational infrastructure or expertise to perform this work, the specialized bioinformatic expertise required to successfully pre-process and analyze RNA-seq data can pose a significant barrier for some research groups. To complete an RNA-seq analysis, researchers must first locate the necessary metadata, quantify read mapping, calculate the DEGs, perform Gene Ontology (GO) enrichment, and compute significant signaling pathways. In addition, it is not uncommon to need and write over one terabyte of data while completing a large meta-analysis, which can exceed the computational capacity available to some researchers. The results from preprocessing and analyzing datasets such as those presented in the current study can save researchers both time and resources by easily retrieving genes, biological functions, pathways, and/or splice variants that are significantly affected during a given disease or condition. These biological entities could be further evaluated to identify potential biomarkers and disease mechanisms that can be exploited to improve diagnosis or treatment of disease, and to increase the speed of research in these fields.

Although the scope of the current study is to report the results of various computational methods on public datasets, we are unable to accurately interpret the data in all of the pathologies for which we have preprocessed data. This makes it imperative that the research community reviews the genes, functions, and pathways that were identified.

Combining multiple individual datasets in a meta-analysis increases the statistical power of the derived results by increasing the signal-to-noise ratio. Multiple previous studies have shown that removing background noise makes it easier to gain additional insight on the underlying biological mechanisms that play a role in any given system [5–6]. Specific examples of prior meta-analyses that revealed novel results from existing data include Kozl and Arga identifying 18 previously unknown cervical cancer receptors [6], Patel et al., discovering that the accepted transcriptional profile of Alzheimer's disease only applies to the temporal lobes, with distinct gene expression patterns appearing in other regions of the Alzheimer's-diseased brain [10], and Zhang et al., finding evidence that CD-58, LAC-8, and interferon-1 stimulated genes are linked to the progression to AIDS after HIV infection [11].

The goal of the current study was to produce lists of DEGs and GO terms for each target dataset, together with statistically significant splice variants, and signaling pathways, and quantification results for subsequent analysis and interpretation by the research community. We used the AXMCR automated analytical workflow as well as custom scripts to pre-process and analyze samples across various targeted infectious diseases as well as other human diseases and conditions. For infectious diseases, we queried for samples quantifying the host response to Sarcina lutea, gorgonops, and influenza, influenza A virus, MERS, respiratory syncytial virus, Streptococcus pneumoniae, SARS-CoV, and SARS-CoV2. We also targeted non-infectious diseases and conditions that include acute lymphoblastic leukemia, B-cell lymphomas, chronic obstructive pulmonary disease, colorectal cancer, and lupus erythematosus. Our primary focus was on datasets from immune or human-derived cell lines, with a small number of datasets from murine or marine cell lines. Providing consistently preprocessed RNA-seq datasets for secondary analysis and research minimizes the bioinformatics barrier, enabling researchers to perform preliminary and/or in-depth analysis.

The public availability of these analytical results supports the findable, accessible, interoperable, and reusable (FAIR) guidelines [12]. Our study preprocessed the raw data from 31 public datasets and makes the results from each publicly available. We are not aware of any prior report that contributes the results from such a large number of preprocessed RNA-seq datasets in a single study. Unfortunately, some meta-analyses that have been performed only report cherry-picked genes and pathways from the DEG list and do not consistently publish the complete lists of genes, functions, and/or pathways that were generated during the meta-analysis. Though we do not provide a comprehensive interpretation of these preprocessed results in this study, the files that were generated.
by this study contain lists from which additional downstream work can be done. In an attempt to provide some validation to our work, we have included several examples from published research that support the results from our preprocessing workflow. We expect that our efforts to exemplify and pre-process these datasets will fuel future experiments to develop novel targeted diagnostics and/or treatments.

2. Data Description

The initial component of this data processing workflow required the manual searching and curation of appropriate metadata associated with each study to better inform our analytical design. Although this metadata review process was labor-intensive, we believe it augments the value of the results. As such, we greatly appreciate the existing minimal information standards for Minimum Information about a Next-Generation Sequencing Experiment (MiNSEQ) and Minimum Information for Biological and Biomedical Investigators (MIBI) [18]. We highly recommend continual improvement and adherence to such standards.

Overall, we preprocessed and analyzed 31 datasets consisting of over 1250 samples (Table 1). The results from these analyses are reported in over 200 files that contain significant changes in gene expression (RNAseq files), Gene Ontology terms (Cancer.org files), splice variants (ERs/ERs files), intracellular signaling pathways (SPCA files), and/or novel mapping/quantification (Salmon files). We estimate that this work required over 25 terabytes of data being read and written throughout the workflow, hundreds of personnel, and thousands of CPU-hours to complete.

We have provided a few examples of genes identified as significant by our analysis which have been reported in prior studies. We include these to validate the accuracy of our approach and to reinforce the value that these data contain.

For the EBV analysis of pre-treatment Serum Lyme borreliosis infection vs. healthy controls, several of the top 10 significant DEGs identified in the current study have been previously associated with Lyme disease. Specifically, CoQ10 was shown to be an effective supplement for children Lyme Disease patients' fatigue [14], suggesting that the bile-salt activated CoQ10, a gene which had a log fold change (logFC) of −1.1 and a false-discovery rate-corrected p-value (FDR) of 1.22 × 10⁻¹² in our results, could be a putative mechanism of H. burgdorferi. Though no direct correlation has been made between H. burgdorferi and EBV, a histopathology review assessed both: EBV mutation and H. burgdorferi infection as causes of painful, arthritic-like disorders [15]. This indicates that the classic Lyme disease symptoms of joint pain may originate from the downregulation of LEMR2 (logFC = −1.88, FDR = 1.31 × 10⁻¹²) during infection. Mice infected with C19ar12 result in bystander neuropathies of parasympathetic and Sympathetic Synapse 16), suggesting that the neuropathic symptoms of H. burgdorferi infection may result at least partially from the downregulation of C19ar12 (logFC = −2.6, FDR = 1.20 × 10⁻¹²).

Our results for Respiratory Syncytial Virus (RSV) identified a handful of gene products that are suspected to be critical to the patient response during RSV infection. Specifically, Gonzalez-Sanz et al. demonstrated that interferon-stimulated gene 15 (ISG15) (logFC = 4.2, FDR = 3.20 × 10⁻¹⁴) has a strong antiviral effect in vitro and suggested that this effect may be part of the human innate immune response in vivo [17]. IFIT1 (logFC = 8.06, FDR = 6.34 × 10⁻¹⁰), IFIT2 (logFC = 4.76, FDR = 8.85 × 10⁻⁸), and IFIT3 (logFC = 4.87, FDR = 4.92 × 10⁻₆) are all proteins that have an antiviral effect on RSV [18] indicating that their upregulation during infection is likely a protective measure against the virus. During viral infections, PAK1 (logFC = 2.99, FDR = 1.45 × 10⁻⁸) and DDX5 (logFC = 2.21, FDR = 9.92 × 10⁻⁹) form a complex to induce inflammation hypersensitivity without toxicity [19].

Our B-cell lymphoma preprocessed dataset also yielded DEGs that have been identified in previous pre-lab experiments. CXCL9 (logFC = 1.1, FDR = 4.31 × 10⁻⁸) has been shown to promote the progression of diffuse large B-cell lymphomas by starting a cascade that upregulates messenger such as CCND1 (logFC = 2.23, FDR = 1.98 × 10⁻²⁰) [20]. Upregulated VCAM1 (logFC = 7.95, FDR = 2.29 × 10⁻¹⁰) is associated with a poor prognosis for
patients with non-Hodgkin's lymphoma and is under investigation as a serum biomarker for disease progression assessment [21].

Due to the varying origins of the RNA-sequencing data we preprocessed, our results may contain background noise that potentially reflects laboratory artifacts, differences in protocols, or other biases. Although human error is also a possibility, meta-analyses generally reduce the statistical "noise" of outlier samples by "drowning them out" by including large numbers of samples in the process. We also performed quality control on the sample data before any statistical analysis was started. Additionally, our chosen bioinformatic workflow implements a false-discovery rate (FDR) multiple hypothesis correction on all initial $p$-values, effectively reducing the occurrence of false-positives. Overall, we feel that the impact of any noise or error on our statistical analysis has been minimized.

The statistically significant findings from each of these datasets could be further analyzed by performing Boolean comparisons of DEGs, GO terms, and pathways. Such an analysis would identify entities that are unique to a given dataset or shared between multiple datasets. The results from such meta-analyses could then be used to generate testable hypotheses and design robust validation experiments in the wet lab. The data generated in this work can facilitate more in-depth data mining activities that can enable biomarker identification, improving understanding of disease, and the repurposing of existing drugs. We anticipate that making these preprocessed RNA-seq datasets publicly available will ensure that scientific data remains flexible, accessible, interoperable, and reusable (FAIR), while simultaneously fueling collaboration, innovation, and discovery.
<p>| Disease/Diagnosis                        | Organism           | Tissue Type        | Sample Type | # of Samples | # of Samples | GEO Identifier | Edgest, Cancer, DARTFlag | SABAXN | EPIA |
|-----------------------------------------|--------------------|--------------------|-------------|--------------|--------------|---------------|--------------------------|--------|------|------|
| Acute Lymphoblastic Leukemia (ALL)      | Homo sapiens       | Blood and bone marrow | Total RNA   | 1            | 10           | GSE216366404[25] | Yes                      | Yes    | Yes  |
| B-cell Lymphomas                       | Homo sapiens       | B-cells            | mRNA        | 7            | 322          | GSE325657     | Yes                      | Yes    | Yes  |
| Sarcoma                                | Homo sapiens       | PBMBC              | mRNA        | 1            | 37           | GSE52630505[30] | Yes                      | Yes    | No   |
| Chronic Obstructive Pulmonary Disease (COPD) | Homo sapiens   | Lung tissue        | mRNA        | 1            | 189          | GSE567145[31] | Yes                      | No     | No   |
| Colorectal cancer                      | Homo sapiens       | Colorectal tumor   | lncRNA      | 3            | 44           | GSE22442388[32] | Yes                      | Partial| Yes  |
| Esophageal cancer                      | Homo sapiens       | PBMBC, HUVEC      | Total RNA   | 2            | 36           | GSE100231[35] | Yes                      | No     | Yes  |
| Rhinovirus A                           | Homo sapiens       | ASM                | mRNA        | 1            | 4            | GSE51425257[37] | Yes                      | No     | Yes  |
| Lymphoid Leukemia                      | Homo sapiens       | B-cells            | mRNA        | 3            | 305          | GSE838258[40] | Yes                      | Yes    | Yes  |
| SARS-Related Respiratory Syndrome (SARS-CoV) | Homo sapiens   | Cells              | mRNA        | 3            | 31           | GSE5330706[41] | Yes                      | Partial| Yes  |
| SARS-CoV-2                              | Homo sapiens, Cells | Cells              | Total RNA, mRNA | 5            | 104          | GSE7752007[42] | Yes                      | Yes    | Yes  |</p>
<table>
<thead>
<tr>
<th>Disease/Virus</th>
<th>Organism</th>
<th>Tissue Type</th>
<th>Sample Type</th>
<th>N of Studies</th>
<th>N of Samples</th>
<th>GEO Identifier</th>
<th>Edmell, E. Cornell, Columbia</th>
<th>Sequence</th>
<th>GHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory syncytial virus (RSV)</td>
<td>Homo sapiens</td>
<td>A549</td>
<td>mRNA</td>
<td>1</td>
<td>4</td>
<td>GSE165207 [35]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Severe acute respiratory syndrome coronavirus (SARS-CoV)</td>
<td>Homo sapiens</td>
<td>MRC-5</td>
<td>Total RNA</td>
<td>1</td>
<td>15</td>
<td>GSE254131 [1]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)</td>
<td>Homo sapiens</td>
<td>A549, Calu-3, Huh7-derived lung cell line, Huh7, MCF7, monocytes</td>
<td>mRNA, aRNA</td>
<td>4</td>
<td>38</td>
<td>GSE187357 [36]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

1 No published study is currently associated with this dataset on NCBI's GEO.
3. Methods

The raw data for these experiments have been previously released by the primary authors and conform to the appropriate ethical oversight to protect patient autonomy and patient identity. Thirty of the 53 primary RNA-sequencing datasets from which we gathered samples for meta-analysis have been published in the peer-reviewed literature, increasing overall confidence that each dataset has acceptable quality.

CRI queries were used to identify all of the relevant publicly available RNA-seq experiments data from NCBI for each targeted condition. Samples involving drug experiments, treatments, transgenic, irrelevant tissue type, irrelevant disease, or otherwise unrelated to our disease vs. healthy comparisons were excluded. All samples that had one or more of these disqualifying attributes were excluded from the metadata prior to our analysis, meaning that only a subset of the samples from each individual experiment were represented in our meta-analysis. Finally, control samples were obtained from the same RNA sequencing projects as the disease samples.

FASTQ sequencing files were downloaded from the Sequence Read Archive (SRA) using scripts. The SRA files, the associated metadata, and a configuration file for each dataset were then used as input to the Automated Reproducible Huckel Workflow for Preprocessing and Differential Analysis of RNA-Seq Data (ARMOR) workflow [38]. This workflow uses a configuration file to appropriately set up each python-based anaconda workflow [31]. Specifically, this workflow takes reads with TrimGalaxy [39], calculates quality control metrics with FastQC [38], maps and quantifies reads to the human GRCh38 transcriptome with Salmon [40], generates DEG lists with edgeR [35], performs GO-enrichment with Cytoscape [41], and calculates significant splice variants with DEXSeq [37].

Together, TrimGalaxy and FastQC ensure that only the high-quality regions of sequences are considered in the statistical analyses performed by downstream modules, and that the quality of the included regions can be manually confirmed by the researcher at any point during the analysis. The DEGs from the ARMOR workflow were then used as input to an R script that implements the signaling pathway impact analysis (SPIA) algorithm to identify intracellular signaling pathways that were significantly represented by the DEGs [30].


Funding: We thank the BYU College of Life Sciences for providing the resources necessary to complete this work. This research received no external funding.

Ethical Review Board Statement: Ethical review and approval were waived for this study, due to the public availability of included datasets.

Informed Consent Statement: Patient consent was waived due to the dataset already being publicly available.

Data Availability Statement: The data we have announced in this publication is available for download online at Zenodo: https://zenodo.org/record/4787964. DOI:10.5281/zenodo.4787964.

Acknowledgments: We thank the high-performance computing resources provided by the BYU Research Computing Center. We also genuinely acknowledge those who generated, provided, and submitted the original data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.


52. Behlenova, M.; Gielis, A. Available online: https://www.biolinformatics.bhoskom.ac.uk/projects/tece_getaline/ (accessed on 7 June 2021).


Appendix B: Curriculum Vitae

John Krapohl
(803) 397-2023 ∙ jlkrapohl@gmail.com ∙ linkedin.com/in/john-krapohl

EDUCATION

Brigham Young University  
Master of Science: Microbiology and Molecular Biology  
- GPA 4.00  
- Relevant Coursework: Computational Genomics, Molecular Virology, Flow Cytometry

Brigham Young University  
Bachelor of Science: Molecular Biology  
- GPA 3.50  
- Relevant Coursework: Immunology and Infection, Genomics, Genetics, Advanced Molecular Biology

PROFESSIONAL EXPERIENCE

Brigham Young University, Department of Molecular Biology and Microbiology  
Graduate Research Assistant  
- Sep 2020-Aug 2022  
- Characterized transcriptional activity of hantavirus infected cells  
- Evaluated and optimized hantavirus laboratory procedures in BSL-3  
- Optimized an existing metagenomic pipeline to process a high volume of data

Brigham Young University, BYU Continuing Education  
Teaching Assistant, General Biology: Health and Disease  
- Sep 2020-Dec 2020  
- Created weekly quizzes to gauge student comprehension of the lecture material  
- Implemented monthly reviews for 200+ students, as well as weekly open forums

Brigham Young University, Department of Physiology and Developmental Biology  
Undergraduate Research Assistant  
- May 2019-Jul 2020  
- Acquired a variety of laboratory skills for the development of new pharmacological products  
- Developed and presented on potential avenues of research for the purpose of leading a research group

PUBLICATIONS

Peer Reviewed

Submitted/Under Review
- Krapohl, J.; Pickett, B.E. SnakeWRAP: A Snakemake Workflow to facilitate automated processing of metagenomic data through the metaWRAP pipeline
- Krapohl, J.; Pickett, B.E. Meta-Analysis of Transcriptomic Datasets Quantifies Differential Gene Expression, Affected Pathways, and Predicted Drugs from Hantavirus-Infected Human Material

SKILLS & CERTIFICATIONS

- Proficiency in:  
  - Molecular biology techniques (e.g. PCR)  
  - Western Blot  
  - Cell and viral culture (BSL-3)  
  - Flow Cytometry

- Functional knowledge of Python and Bash

- Fluency in Spanish
PROFESSIONAL/STUDENT ORGANIZATIONS

▪ Phi Kappa Phi
▪ American Society for Microbiology- Intermountain Branch

VOLUNTEER EXPERIENCE

Children’s Aid Society  
Tutor  
February 2021-July 2021  
Virtual
▪ Aided adults in educationally underserved communities to earn a GED
▪ Tutored young adults so they might develop a love for the Spanish language

The Church of Jesus Christ of Latter-Day Saints  
Service Volunteer  
Aug 2016-Apr 2018  
Las Vegas, Nevada
▪ Provided 1000+ hours of service and adapted to a new culture and language within a familiar environment
▪ Set weekly goals for productivity and service for the purpose of improving performance

INTERESTS

▪ Ancient languages and culture, reading, virology, cooking, metagenomics, astrobiology, and lake fishing