

TAXONOMER: A FAST AND ACCURATE
METAGENOMICS TOOL AND ITS
USES ON CLINICAL SPECIMENS

by

Keith Eugene Simmon

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2016

Copyright © Keith Eugene Simmon 2016

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Keith Eugene Simmon
has been approved by the following supervisory committee members:

<u>Karen Eilbeck</u>	, Chair	<u>02-26-2016</u> Date Approved
<u>Mark Yandell</u>	, Member	<u>02-26-2016</u> Date Approved
<u>Robert Schlberg</u>	, Member	<u>02-26-2016</u> Date Approved
<u>Catherine Staes</u>	, Member	<u>02-26-2016</u> Date Approved
<u>Brain Chapman</u>	, Member	<u>02-26-2016</u> Date Approved

and by Wendy Chapman, Chair/Dean of
the Department/College/School
of Biomedical Informatics

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Advances in sequencing technologies have made it possible to generate large amounts of microbiological sequence data without culture methods. The data generated pose a significant data analysis challenge. This is especially true in clinical diagnostics where accurate and timely diagnoses are key. To enable infectious disease diagnostics, we created Taxonomer, a kmer-based metagenomics software tool, which can rapidly process large amounts of sequence data with accuracy and precision similar to slower alignment-based approaches. A kmer is a nucleotide subsequence of k length. Kmer exact matching is performed in RAM, utilizing data structures with rapid query times, making kmer approaches magnitudes faster than alignment methods. Prior to Taxonomer, other kmer-based methods were subject to high false positive rates. Taxonomer differs by 1) providing a workflow that reduces false-positives, 2) including host-transcript profiling, and 3) providing a novel protein kmer tool to identify viruses, which are typically too divergent to reliably identify using nucleotide sequence.

A web-based front-end was created with the D3 enabled iobio framework. Reference sets utilized in Taxonomer were obtained from NCBI, GreenGenes, unite, and uniprot databases. A wide-range of simulated datasets and real clinical specimens were created or obtained to evaluate Taxonomer. Taxonomer was compared to previously published pipelines (SURPI), classifiers (Kraken, RDP classifier), and sequence alignment methods (BLAST, SNAP, RapSearch2, DIAMOND). Taxonomer was also

compared to a commercially available respiratory virus panel and utilized on a large cohort of pneumonia positive patients that had previously undergone extensive microbiological diagnostics.

Taxonomer had agreement at 98.7% with SURPI to assign reads at the phylum level. Taxonomer, RDP classifier, and Kraken classified simulated 16S rRNA reads correctly at the species level at 59.5, 61.7, and 46.0%, respectively. Protein classification using reads derived from viruses showed similar sensitivity to alignment-based methods with RapSearch2, and DIAMOND but with slightly decreased analysis times.

Taxonomer provides an accurate workflow for processing samples in a diagnostic setting. It identifies bacteria, fungi, virus, and human transcripts from clinical specimens with accuracy comparable to alignment methods. Its web-based front-end makes it accessible to laboratories without significant compute resources.

TABLE OF CONTENTS

ABSTRACT.....	iii
LIST OF TABLES.....	vii
LIST OF FIGURES	ix
ACKNOWLEDGEMENTS.....	xii
Chapters	
1 INTRODUCTION	1
1.1 Kmer Based-Methods	3
1.2 Bacterial Identification.....	4
1.3 Fungal Identification.....	5
1.4 Bacterial and Fungal Identification Resources	6
1.5 Molecular Viral Identification	7
1.6 Metagenomics.....	8
1.7 References.....	11
2 TAXONOMER: INTERACTIVE WEB-BASED METAGENOMICS ANALYSIS PORTAL FOR UNIVERSAL PATHOGEN DETECTION AND HOST RESPONSE- BASED DIAGNOSIS AND DISCOVERY	16
2.1 Introduction.....	16
2.2 Materials and Methods.....	18
2.3 Results.....	32
2.4 Discussion.....	40
2.5 References.....	42
3 UNBIASED DETECTION OF RESPIRATORY VIRUSES BY NEXT- GENERATION SEQUENCING AND TAXONOMER, A RAPID, INTERACTIVE, WEB-BASED DATA ANALYSIS TOOL.....	91
3.1 Introduction.....	91
3.2 Materials and Methods.....	93
3.3 Results.....	96

3.4 Discussion.....	102
3.5 References.....	108
4 VIRAL PATHOGEN DETECTION BY METAGENOMICS AND PANVIRAL PCR IN CHILDREN WITH PNEUMONIA WITH NO IDENTIFIABLE ETIOLOGY, RESULTS FROM THE CDC ETIOLOGY OF PNEUMONIA IN THE COMMUNITY (EPIC) STUDY	117
4.1 Introduction.....	117
4.2 Materials and Methods.....	119
4.3 Results.....	124
4.4 Discussion.....	128
4.5 References.....	132
5 CONCLUSIONS.....	143
5.1 References.....	148

LIST OF TABLES

Tables

2.1 Reference databases for the ‘Binner’ module, source, version, number of reference sequences, and kmers.	76
2.2 Bin assignment for reads with equal numbers of kmer matches to multiple Binner databases and kmer matches below threshold.	78
2.3 Contents of visualized pie charts in the web portal. Sub-bin assignments are summarized for interactive visualization at Taxonomer.iobio.io as indicated.....	79
2.4 Optimal kmer cutoffs for bin assignments based on the Youden’s Index and F1 Score.....	80
2.5 Viruses, percent nucleotide-level identity to reference sequences in the NCBI nt database, as well as numbers of total and viral reads for pediatric upper respiratory tract specimens used to compare ‘Protonomer’, RAPSearch2, and DIAMOND for protein-level classification of viral sequences.....	81
2.6 Flux Simulator parameters used to generate simulated RNAseq reads for benchmarking transcript assignment.....	82
2.7 Processing time of Taxonomer compared to rapid classification pipelines SURPI and Kraken.....	83
2.8 Broad taxonomic classification of read 1 versus read 2 by SURPI differs for 2-9% of mate pairs.....	84
2.9 Accessions for published 16S amplicon data used in for bacterial abundance estimates, numbers of reads, and analysis times for the RDP Classifier and Taxonomer. Number of reads for reference is based on mate pairs.....	85
2.10 Taxonomer bacterial abundance estimates compared to those of the RDP Classifier using recently published 16S amplicon sequencing metagenomics studies.....	86
2.11 Accession numbers for human brain RNAseq data used to compare with MAQC qPCR data.....	87

2.12	Genes that are differentially regulated in nasopharyngeal and oropharyngeal swabs from children with pneumonia who tested positive for influenza virus compared to asymptomatic controls.....	88
2.13	Gene ontology assignments for enrichment of biological processes and molecular functions are shown.....	89
2.14	Taxonomer is compatible with different sequencing protocols.....	90
3.1	Primer and probe sequences for the unpublished monoplex, real-time PCR assays for respiratory viruses	116
4.1	Demographic and clinical information of children with pneumonia with no identifiable etiology and controls	141
4.2	Performance of RNA-seq and panviral PCR compared to pathogen-specific real-time PCR performed per EPIC protocol.....	142

LIST OF FIGURES

Figures

1.1	The number of bacterial genera and species described in 1975 around the introduction of 16S rRNA phylogenetic studies and 2013.....	13
1.2	Venn diagram showing the shared 31 bp kmer composition of <i>A. baumannii</i> and <i>S. aureus</i>	14
1.3	The taxonomic lineage annotations of sequences deposited into International Nucleotide Sequence Database Collaboration (INSDC) databases.....	15
2.1	Taxonomer’s architecture.....	48
2.2	Intersections of kmers in Binner module databases	49
2.3	Receiver operator characteristics curves for classification of human and microbial sequences by the Binner module.....	50
2.4	Agreement between read binning by the Binner module versus SURPI’s assignments	51
2.5	Taxonomer web-service	52
2.6	False-positive and false-negative classifications of query sequences not represented in the reference database	53
2.7	Sensitivity for binning of bacterial and viral reads can be low for phylogenetically distant species	54
2.8	Read-level and taxon-level bacterial classification accuracy of BLAST, the RDP Classifier, Kraken, and Taxonomer.....	55
2.9	Taxonomer sensitivity and specificity for read-level bacterial classification compared to two other rapid classification tools SURPI and Kraken.....	56
2.10	Comparison of three commonly-used reference databases in Taxonomer	57

2.11 Bacterial 16S rRNA classification accuracy	58
2.12 Fungal ITS classification accuracy	59
2.13 Effect of Sequence length on classification 16S rRNA reads.....	60
2.14 Impact of sequencing error rates.....	61
2.15 Classification of metagenomics data compared to the RDP classifier	62
2.16 Analysis times for the RDP Classifier, Taxonomer, and Kraken	63
2.17 Performance of the Protonomer module for virus detection.....	64
2.18 Sensitivity of tools correlated with phylogenetic distance of viral strains to reference sequences	65
2.19 Detection of highly pathogenic viruses.....	66
2.20 Published RNA-seq data from a commercially available RNA analyzed by Taxonomer.....	67
2.21 Application of Taxonomer to metagenomic RNA-seq data from routine respiratory samples from patients with influenza infection	68
2.22 Classification of viral sequencing reads by Protonomer and typing of this strain as influenza A (H1N1)	69
2.23 Differential gene-level mRNA expression profiles from 4 patients with influenza A virus compared to asymptomatic controls infection.....	70
2.24 Sub-analysis of differential gene-level mRNA expression profiles from 4 patients with influenza A virus compared to asymptomatic controls infection	71
2.25 Sample applications of Taxonomer	72
2.26 Taxonomer in clinical samples	74
2.27 Phylogenetic tree of consensus sequence of novel Anellovirus with reference sequences for Torque teno mini viruses	75
3.1 Respiratory Virus detection by RNA-seq plus Taxonomer and performance comparison with a commercial multiplex PCR panel	112
3.2 Overall taxonomic composition of RNA-seq reads and numbers of viral reads by respiratory virus.....	113

3.3	Correlation of normalized read counts with viral burden and precision of viral read abundance within and between sequencing runs.....	114
3.4	High-resolution, sequence-based typing of 14 human rhinovirus strains based on RNA-seq directly from NP swabs.....	115
4.1	Proportion of children with CAP with no identifiable etiology during the 2½-year EPIC study.....	136
4.2	Detection of additional human viruses by RNA-seq and panviral PCR in EPIC participants with positive pathogen-specific tests.....	137
4.3	Viruses detected by RNA-seq and/or panviral PCR in children with pneumonia with no identifiable etiology and asymptomatic controls.....	138
4.4	An abundant bacterial flora dominated by a single potential pathogen was detected by RNA-seq in NP/OP samples of two children.....	139
4.5	Pathogen detection in children with CAP with no identified etiology and asymptomatic controls by method.....	140

ACKNOWLEDGEMENTS

As a student, I was given the opportunity to experiment. The result: I learned a lot. I am thankful for that. The drive, intelligence, and work ethic of Mark Yandell and Robert Schlberg are to be admired. I am in awe of it. Being able to work in Mark's lab and with his students and staff was a great experience. Karen Eilbeck is family and I cannot help but relate to her. Under her snarky exterior is a kind and generous person. I am grateful for her mentorship, friendship, and hand-me-downs because we (Anna and I) like her style. While I did not get to work with Brian Chapman and Catherine Staes on a daily basis, I appreciate the time they have taken to be on my committee. Working with Steven Flygare was great. I can only wish you the best of luck on your future endeavors.

While finishing my PhD, my kids and I were hit by a car and seriously injured. Many students and staff came to visit me in the hospital as well as offered assistance. I cannot express how deeply touching this was for an introverted curmudgeon like myself. Thank you. Thanks to my ARUP bosses (Clint Wilcox and David Nix). You allowed me to work from home and take the time I needed to recover. This eased the burden on our family and was wonderful. Thank you to my parents for supporting our family for two months while we recovered. My god, that was incredible. Thanks to my wife, Anna, and our goober-bears. We do have pretty cute kids.

CHAPTER 1

INTRODUCTION

Advances in diagnostic technology have provided greater resolution to examine the natural world. In the field of microbiology, new technologies have enabled researchers to better distinguish between microorganism species and strains. This has led to the discovery of many novel bacteria and fungi and led to a doubling of new bacterial species named nearly every 10 years (1). In 1977, Carl Woese et al. analyzed the 16S rRNA gene from bacteria and established a gene-based approach to understand the phylogenetic relationship of bacteria (2). In the nearly 40 years that have followed, the 16S rRNA gene is still the standard marker for bacterial evolutionary genetics studies and molecular identification. The growth of named bacterial species has increased from ~3,000 to ~12,000 (see Figure 1.1), since the introduction of gene-based phylogenetics. Similar gene-based approaches utilizing the rRNA genes for fungal organisms have been employed to broaden our understanding of eukaryotic microorganisms. These gene-based approaches remain popular because of the rich reference sets that have been created, standardized protocols, and assessed interpretation criteria.

DNA sequencing is central to these gene-based methods for microorganism detection, identification, and phylogeny. Recent developments in sequence technologies, in particular massively parallel sequencers often called Next-generation Sequencing

(NGS), have allowed the expansion of these techniques. NGS has opened up a new field of research called metagenomics, the study of complex microbial communities. Metagenomics and NGS represent a significant advancement in microbiological research because it allows the investigation of environments without culture.

Clinical metagenomics is emerging as a field of important research. Measuring the composition of bacterial species present at sites on the human body has shown associations with allergies (3), eczema (4), and irritable bowel syndrome (5). While cost and turn-around-time remain a barrier to routine implementation of clinical metagenomics in reference laboratories, it is clear that metagenomics techniques will be important for future diagnostics as more knowledge is gained about the human-microbiome interaction.

Metagenomics sequencing techniques can generate massive amounts of data that may take weeks to analyze. This has created an opportunity for computational biologists to develop new techniques, data resources, and strategies for analyzing metagenomics data. While techniques and tools exist for rapid analysis, the diagnostic implementation is challenging because of poor precision, lack of clarity on whether the information is actionable, and complexity in interpretation.

This dissertation describes the creation of Taxonomer, a rapid metagenomics tool targeting clinical and diagnostic microbiology. Taxonomer is distinct from other metagenomics tools in that it focuses on speed and maximizing sensitivity and specificity for the identification of bacterial and fungal species, identification of sequences from known divergent and novel viruses, and performs human transcriptional analysis to detect fluctuations in gene expression involved in immunological response. Taxonomer works

in a high Random Access Memory (RAM) environment and utilizes short DNA sequences (*kmers*) to perform sample analysis in minutes. Taxonomer exists as a web-based platform placed into the iobio framework (6), which makes it accessible to laboratories without significant computational resources.

The introduction chapter seeks to provide additional context explaining Taxonomer's design choices. Additional background is provided to illustrate the need for Taxonomer. Chapter 2 fully describes the components of Taxonomer, benchmarks the tool against current tools, and includes several use cases. Chapter 3 compares metagenomics methods to a standard respiratory viral diagnostic panel, and Chapter 4 demonstrates how Taxonomer can be employed to understand the etiology of pneumonia using a large cohort of well-characterized samples.

1.1 Kmer-Based Methods

Alignment methods for DNA compute a scoring matrix to determine the optimal alignment between two sequences. Matching bases receive high scores; 'mismatches' and 'gaps' receive low scores. From the scoring matrix, the best possible alignment can be determined. Many metagenomic tools use alignment methods to determine the taxonomic origin of a sequence (7,8). However, producing an alignment is not essential to answer the underlying question: Does this sequence belong to a human, bacterial species, or fungal species? A quicker alternative to alignment is exact kmer matching.

Kmer matching does not compute a score but simply queries a database or hash table to determine the kmer present. The utility of kmer matching is provided in the following illustration (see Figure 1.2): imagine we want to determine if a sequence is

derived from *Acinetobacter baumannii* or *Staphylococcus aureus*. First, we analyze the reference sequences; in this illustration we use GenBank references NC_021733 (*A. baumannii*) and NC_002745 (*S. aureus*) and count the 31 base pair (bp) kmers. The *A. baumannii* strain's genome is 4,001,621 bp and the number of distinct 31 bp kmers in that genome is 3,925,464. This shows that only 1.9% of the *A. baumannii* genome contains redundant 31 bp sequences. For *S. aureus*, the genome size is 2,814,816 bp and the 31 bp kmer count is 2,743,338; 2.5% of the kmers are redundant. When comparing the two genomes' 31bp kmers ($n=6,668,802$), only 102 (0.002%) are seen in both species. In this simple illustration, it is clear to see that a single kmer derived from one organism is unlikely to be confused with the other. NGS sequence length can vary from 50 – 300 bp, meaning that for a single sequence, as many as 20 to 270 31bp kmers will be queried for each sequence. Even with multiple queries per sequence, interrogated kmer-based methods can be 900 times faster than alignment methods like BLAST (9). In Chapter 2, we provide data and comprehensive benchmarks that show the efficacy of kmers to both discriminate sequences from numerous taxa and identify the origin of a sequence down to the species level in some cases.

1.2 Bacterial Identification

Microbial culture independent studies have led to the discovery of over 200,000 hypothetical bacterial species based on 16S rDNA sequence comparison. When you contrast this with the number of microorganisms that have been named ($n = \sim 14,000$), it is apparent that the majority of the known bacterial world has not been named or even grown in a laboratory. Currently, there are over 5,000,000 bacterial sequences in public

databases of which ~3,000,000 are 16S rRNA gene sequences (10), and only around ~5,000 species have complete bacterial genomes (11). NGS sequencing makes it possible to sequence genomic regions outside of the 16S rRNA gene; however, the decades of work have created a robust and complete database to identify and compare microorganisms. As NGS has become more available, the use of 16S rDNA in studies has helped us understand bacterial communities of both environmental and health significance. It is estimated that through current research, the microbial world as understood by the 16S rRNA gene will be saturated by the end of the current decade, meaning that nearly all bacteria species will have had its 16S rRNA sequenced (12).

The 16S rRNA gene is a well-established marker, but it provides a limited view of a microorganism's phenotype. Additionally, the evolutionary rate at which it mutates can be too slow to provide adequate information to differentiate some species (13). Other DNA targets can provide more resolution to speciate; for example in mycobacteria, the *rpoB* gene is often used to differentiate members of the *Mycobacterium chelonae* and *M. abscessus* group. However, the use of alternative targets is hindered by both database coverage and the inability to create universal primers that work outside of a family or phylum.

1.3 Fungal Identification

As with bacteria, molecular fungal identification is carried out by using ribosomal sequence. While bacteria identification is largely performed with the small subunit of the ribosome (i.e. 16S rDNA), the markers for fungal identification are either the 28S gene (LSU) or Internal Transcribed Spacer (ITS). Within the last few years, the ITS has

become preferred over the LSU because it has higher variability, thus providing greater species resolution (14). The amount of reference information for fungi is significantly less than the information available for the bacterial 16S rRNA gene. While over 3,000,000 16S rRNA sequences are available, only ~10,000 LSU and ~500,000 ITS sequences are present in public reference databases (15,16).

1.4 Bacterial and Fungal Identification Resources

Several publically available resources have been created to facilitate identification of microorganisms and phylogenetic studies. The main role of these resources is to house and curate 16S, ITS, LSU sequences and provide tools which aide users to query or interact with the reference set. The need for these resources is clear when looking at the annotations present in International Nucleotide Sequence Database Collaboration (INSDC) databases, where nearly less that <20% of the sequences are annotated to the species level (see Figure 1.3). While there is redundancy between the different resources, each provides some distinct value.

Bacterial resources include GreenGenes, Silva, and the Ribosomal Database project (10,15,17). Each of these resources curate publically available 16S rRNA sequences INSDC. Sequences are filtered for quality and the taxonomic lineage annotation is reassessed and amended if necessary. In addition to the curation, the resources provide tools to classify sequences against their curated data. Both GreenGenes and Silva provide datasets that are clustered and aligned. The clustering of sequences provides a method to reduce redundant data as well as estimate the breadth of species present in the database without having to evaluate taxonomic annotations associated with

the sequences. Observation of the 16S rRNA gene have established that strains of the same species generally have <1.0% difference in percent identity (12). Therefore, clustering sequences at 99% identity yields estimates of hypothetical abundance. Sequences belonging to a cluster (i.e., within 1% sequence identity) are considered to be part of the same *Operational Taxonomic Unit* (OTU) or *Species Hypothesis* (SH). A canonical OTU sequence, usually the centroid, is selected to represent the OTU in subset databases. Publically available reference sequences for fungi are available in SILVA (LSU) and Unite (ITS) (15,16). Both SILVA and Unite provide clustered sequences collapsed to OTUs.

Taxonomer leverages the curated 16S rRNA and ITS clustered datasets provided by GreenGenes and Unite to identify bacteria and fungal sequences, respectively. These marker genes provide 100X more species coverage than is provided by whole genome sequence. In Chapter 2, we demonstrate an increase in sensitivity and specificity by using these highly curated references over other tools utilizing uncurated and/or genomic references sets.

1.5 Molecular Viral Identification

Viral identification is significantly different from the identification of bacteria and fungi for several reasons. They lack universally conserved genes, strains of the same viral species can be highly divergent, and size varies drastically between different viruses. Like bacteria and fungi, a significant portion of viruses remain unknown; however, unlike bacteria and fungi, which contain conserved genes (e.g., 16S rDNA, LSU, *rpoB*) that

allow comparison to already known species, viruses may have little shared sequence identity to other known viruses.

To detect novel viral sequences, nucleotides are translated into amino acids and compared against a protein database. The reduction in complexity when converting a DNA sequence to an amino acid sequence more readily allows for the detection of distantly related sequences. However, searching in protein space is more computationally expensive since the query needs to be translated into all 6 frames and more alignments may be seeded. In Chapters 2, 3, and 4 we demonstrate the benefit of protein searching in the detection of viral sequences from clinical specimens.

1.6 Metagenomics

In the above sections, we discussed the current practices and resources for microorganism identification. In this section, we address metagenomics. For our purpose, metagenomics is defined as the identification of bacteria, fungi, and virus, and the detections of markers associated with an infection in a complex sample. This definition succinctly describes what Taxonomer tries to achieve by identifying bacteria, fungi, viruses, and profiling human transcripts. Metagenomic methods provide a diagnostic advantage by eliminating the need for culturing. It also has the ability to detect both novel and under-appreciated bacterial, fungal, and viral species that analyte-based testing cannot.

1.6.1 Amplicon Sequencing

Amplicon sequencing was not the specific target of Taxonomer, but it is important to note because it is used extensively in human microbiome/metagenomics studies. Amplicon sequencing uses Polymerase Chain Reaction (PCR) to magnify a portion of the 16S rRNA gene, or other marker, from a sample. The resulting amplicon is sequenced. Because a single molecule is sequenced by this technique, the analysis is fairly straightforward and computationally inexpensive. Analysis steps can be reduced by counting and removing duplicate sequences, leaving only a portion of the reads to be compared against reference sets. The drawback of this approach is that it is not suitable for viruses and if the intent is to measure multiple taxa, then multiple amplicons need to be generated.

1.6.2 Shotgun Methods

Taxonomer was developed utilizing shotgun sequencing as the primary sample type. Two broad shotgun techniques exist, RNA-seq or DNA-seq, which use libraries created from isolated RNA and DNA, respectively. These sample types are more complex to analyze because they contain sequence from random genomic regions of the taxa present in the sample. This means very few sequences will be duplicates and thus each sequence read needs to be analyzed to determine its taxonomic origin.

1.6.3 Analysis of Shotgun Metagenomics

A clinical metagenomics sample can contain sequence from the patient, the microbiome community, environmental source, and/or pathogen. The diversity of the

taxa that may be present has led many to create bioinformatics pipelines that sequentially identify and bin sequences belonging to certain taxa (7,18). Central to these pipelines are the alignment tools that compare reads to different reference sets; ideally the initial screening steps will be performed by a rapid alignment tool with a small reference database to remove a large portion of reads from the samples set. This first step is most frequently performed using the human genome. Subsequent steps use larger databases, which take a longer time to process to identify reads that are from bacterial, fungal, and other organisms. A logical data flow is essential because the time to analyze samples can take hours to days to process depending on the composition of taxa present (see Chapter 2).

Metagenomics data are complex and understanding how to interpret the data is important to avoid false conclusions. In shotgun metagenomics, data can be derived from different regions of the microorganism genome. Using these random DNA fragments for identification with the same algorithm may lead to erroneous conclusions (19-21). Several factors may exacerbate this issue, including a) a lack of database references for the region being screened, b) a lack of knowledge of the evolutionary pressure and the subsequent mutation rate in order to accurately compare results to reference sequences, and contaminated reference sequences.

We created Taxonomer to minimize the issues described above. Central to Taxonomer's approach is the use of vetted marker genes for identification, and a classification system that provides confidence scores based on kmer weighting. In the subsequent chapters, we illustrate the benefits of our approach.

1.7 References

1. Tindall BJ, Kampfer P, Euzéby JP, Oren A. Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. *Int J Syst Bacteriol* [Internet]. **2006**; 56(11):2715–2720.
2. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *PNAS. National Acad Sciences*; **1977**; 74(10):4537–4541.
3. Arrieta M-C, Stiemsma LT, Dimitriu PA, Thorson L, Russell S, Yurist-Doutsch S, et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med. American Association for the Advancement of Science*; **2015**; 7(307):307ra152–307ra152.
4. Kobayashi T, Glatz M, Horiuchi K, Kawasaki H, Akiyama H, Kaplan DH, et al. Dysbiosis and *Staphylococcus aureus* colonization drives inflammation in atopic dermatitis. *Immunity*. **2015**; 42(4):756–766.
5. Kennedy PJ, Cryan JF, Dinan TG, Clarke G. Irritable bowel syndrome: A microbiome-gut-brain axis disorder? *World Journal of Gastroenterology : WJG. Baishideng Publishing Group Inc*; **2014**; 20(39):14105–14125.
6. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods. Nature Publishing Group*; **2014**; 11(12):1189–1189.
7. Kostic AD, Ojesina AI, Peadarallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnology. Nature Publishing Group*; **2011**; 29(5):393–396.
8. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research. Cold Spring Harbor Lab*; **2014**; 24(7):1180–1192.
9. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol. BioMed Central Ltd*; **2014**; 15(3):R46.
10. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research. Oxford University Press*; **2014**; 42(Database issue):D633–42.
11. NCBI Genomes [Internet]. 2015. Available from: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html

12. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Publishing Group. Nature Publishing Group; **2014**; 12(9):635–645.
13. Simmon KE, Brown-Elliott BA, Ridge PG, Durtschi JD, Mann LB, Slechta ES, et al. Mycobacterium chelonae-abscessus complex associated with sinopulmonary disease, Northeastern USA. Emerging Infect Dis. **2011**; 17(9):1692–1700.
14. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci USA. National Acad Sciences; **2012**; 109(16):6241–6246.
15. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. Nucleic Acids Research. Oxford University Press; **2014**; 42(Database issue):D643–8.
16. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et al. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol. **2013**; 22(21):5271–5277.
17. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. American Society for Microbiology; **2006**; 72(7):5069–5072.
18. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Research. **2014**; 24(7):1180–1192.
19. Ackelsberg J, Rakeman J, Hughes S, Petersen J, Mead P, Schriefer M, et al. Lack of evidence for plague or anthrax on the New York City subway. Cell Systems. **2015**; 1(1):4–5.
20. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Systems. **2015**.
21. Gonzalez A, Pettengill J, Vazquez-Baeza Y, Ottensen A, Knight R. Accurate Detection of pathogens in microbial samples (and avoiding the conclusion that the platypus rules the earth). International Symposium on Microbial Ecology [Internet]. Seoul; **2014**. Available from: <http://i.imgur.com/Up4mGEE.png>

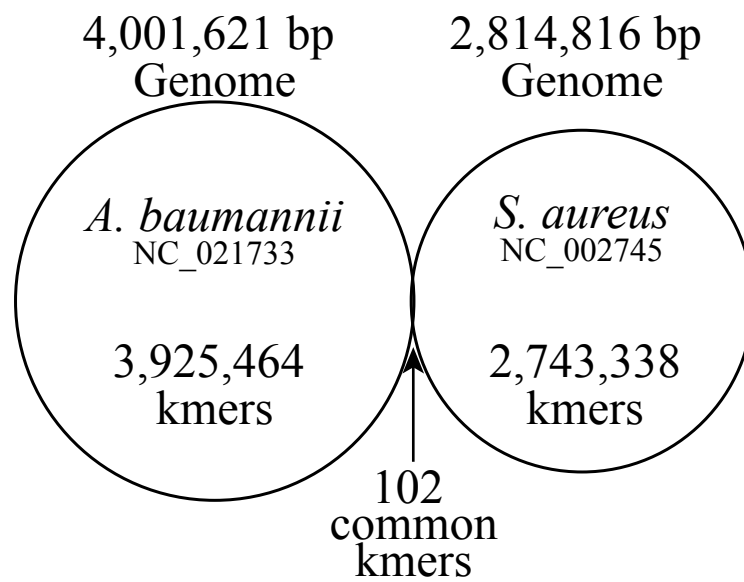


Figure 1.2. Venn diagram showing the shared 31 bp kmer composition of *A. baumannii* and *S. aureus*.

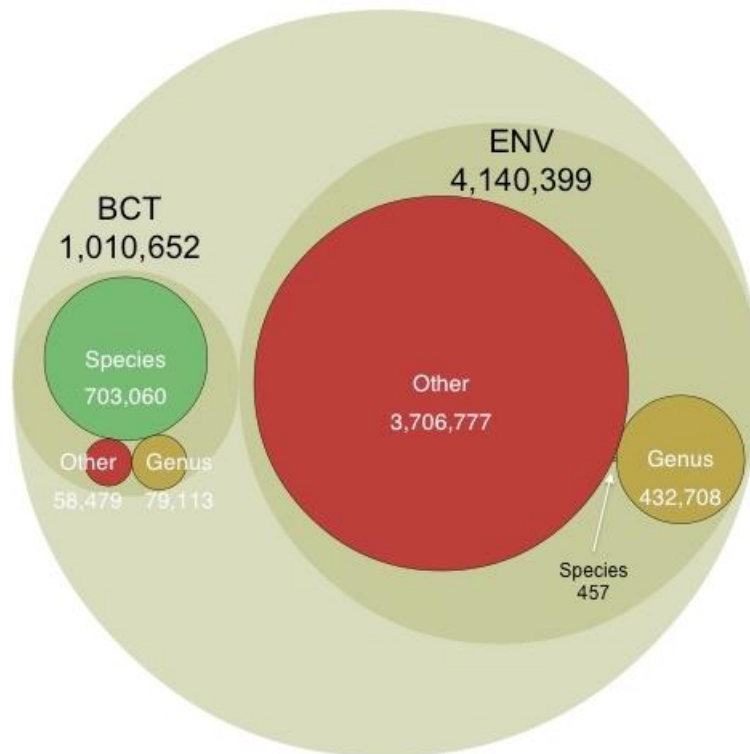


Figure 1.3. The taxonomic lineage annotations of bacterial sequences deposited into International Nucleotide Sequence Database Collaboration (INSDC) databases. The references are grouped into deposit categories bacteria (BCT) and environmental (ENV). The majority of sequences deposited do not have annotation indicating their genus or species. Some of these represent unnamed or unplaced bacterial species.

CHAPTER 2

TAXONOMER: INTERACTIVE WEB-BASED METAGENOMICS ANALYSIS PORTAL FOR UNIVERSAL PATHOGEN DETECTION AND HOST RESPONSE-BASED DIAGNOSIS AND DISCOVERY¹

2.1 Introduction

With replacement of microbial culture by molecular tests, the laboratory diagnosis of infectious diseases increasingly relies on pathogen-specific tests. While more sensitive, they require *a priori* knowledge of likely etiologic agents (i.e. answering the question ‘is pathogen X present’). For several common syndromes (e.g. pneumonia, sepsis, encephalitis), many different pathogens can cause clinically indistinguishable symptoms. Thus, increasingly large, yet inherently limited diagnostic panels are necessary for detection of common pathogens, and exhaustive follow-up testing may be required if first-line tests are negative. A unified approach for detection of all potential pathogens (panmicrobial detection) will increase diagnostic yield, decrease time to result for unexpected pathogens, and improve targeted treatment.

¹ This chapter has been submitted to Genome Biology. Co-authors include Dr. Mark Yandell, Dr. Karen Eilbeck, and Dr. Robert Schlaberg. Keith Simmon is co-first author.

Panmicrobial detection will also enable more comprehensive microbial profiling studies. For example, dysbiosis of the mucosal and cutaneous microbiota has been linked to metabolic, immunologic, cardiovascular, and neoplastic diseases (1-6). However, today most microbiome studies still rely upon PCR amplification of marker genes (e.g. bacterial 16S rRNA). This approach introduces bias (7), ignores effects of the relevant viral and phage flora for which no marker gene exists (8-10), and is unable to assess host response differences, all of which are known to influence the outcome of infectious diseases and modulate human microbial communities.

More recently, wide availability of next-generation sequencing instruments and lower reagent costs have enabled enrichment-independent metagenomics through shotgun sequencing of mixed microbial and host DNA or RNA directly from patient samples. This enrichment-independent approach to metagenomics enables hypothesis-free detection and genomic characterization of a theoretically unlimited number of pathogens (i.e. answering the question ‘what pathogen is present’). Indeed, metagenomic approaches have already led to the diagnosis of previously unrecognized infections and discovery of novel pathogens (11-13). Moreover, RNA-seq-based metagenomics potentially enables novel diagnostic approaches. For example, host transcriptional responses to pathogens can be used to inform treatment decisions, e.g. by differentiating viral from bacterial infections or colonization from true infections, thus helping to limit antibiotic treatment (14-16).

Unfortunately, analysis of the large datasets generated by high-throughput metagenomics requires a combination of bioinformatics skills, computational resources, and microbiological expertise that is absent from most laboratories, especially diagnostic

laboratories. Thus, more computationally efficient, accurate, and easy-to-use analysis tools are needed.

Here we describe Taxonomer, an ultrafast, user-friendly, web-based tool for metagenomic sequence analysis. Taxonomer supports both nucleotide and protein-based classification and enables new analysis modalities for clinical metagenomics datasets. We illustrate the power of Taxonomer using samples from a national pneumonia study (17), and published RNA-seq data from patients with highly pathogenic virus infections. Using Taxonomer, we detect previously unrecognized human infections in these data, and identify an antiviral transcript response signature directly from nasopharyngeal swabs of children with influenza, which has important implications for diagnosis and discovery. Taxonomer is publically available via an iobio (18) web-service, allowing rapid, highly interactive analyses using personal computers and mobile devices.

2.2 Materials and Methods

2.2.1 Binner Module

Identifying small numbers of pathogen sequences hidden among vast numbers of host and/or microbiota-derived sequencing reads is a major algorithmic challenge for metagenomics-based pathogen detection tools. The standard approach is to use digital subtraction (19), whereby all sequencing reads are first aligned to the host's genome sequence. This is the approach used by SURPI (20), for example. During subtraction, reads of host origin are removed. Additional subtraction steps may be used for removal of non-relevant microbial sequences, including those known to represent reagent contamination (21) or sequencing adaptors. A greatly reduced number of presumably

relevant microbial sequences are then classified by alignment to larger reference databases. Since only the remaining reads are matched with selected reference sequences, pathogens can be missed entirely if they are homologous to sequences in the subtraction database. Taxonomer overcomes this inherent limitation of digital subtraction by means of its 'Binner' module, which compares each read to every reference database in parallel, assigning them to broad, non-exclusive taxonomic categories (see Figure 2.1).

Taxonomer's Binner database is created by counting unique 21bp kmers in different taxonomic/gene datasets using Kanalyze (22) (version 0.9.7). Each taxonomic/gene dataset represents a 'bin' in which query sequences can be placed based on their kmer content. Each database is assigned a unique bit flag that allows kmers that belong to one or more bins to be recognized and counted. The database bins and flags are shown in Table 2.1. The kmer counts are merged into a binary file that contains the kmers and the database flag. This binary file shares a similar organization to our classification databases, and is organized to optimize query speed. Reads are then assigned to the taxonomic group(s) with which most kmers are shared. Ties are resolved as shown in Table 2.2 and results summarized for visualization (see Table 2.3). High binning accuracy is possible because of the minimal intersections (0.47%) of kmer content from comprehensive human and microbial reference databases (see Figure 2.2). Optimal kmer count cutoffs were determined by Youden's indexes and F1 scores (23) and ranged from 3 to 13 (default, n=11) (see Figure 2.3 and Table 2.4). To eliminate binning of reads containing adapter sequence, by default, the binner ignores kmers present in Illumina Tru-Seq adapters. A database of External RNA Controls Consortium (ERCC) control sequences allows quantification of ERCC spike-in controls.

To demonstrate the advantage of Taxonomer's non-greedy binning algorithm, we compared high-level taxonomic assignments made by SURPI, which employs greedy digital subtraction using sequence alignments by SNAP (24), to those of Taxonomer's alignment-free Binner (see Figure 2.4). While high-level taxonomic assignments agree for 73.8% of RNA-seq reads, Taxonomer assigned 16% of reads an ambiguous origin (i.e. they match equally to multiple databases), and 96% of these were classified as human by SURPI. This was mostly due to highly conserved ribosomal and mitochondrial sequences (data not shown), but similar effects were also apparent for fungal sequences (18% classified as human by SURPI). Taxonomer's Binner was also able to capture more phage/viral sequences (7,426) than the alignment-based method (5,798), and resulted in fewer unclassified sequencing reads (3.2% vs. 4.5%). Consistent with lower abundance of rRNA and mtRNA sequences in DNA sequencing data, Taxonomer had many fewer ambiguous assignments (0.04%, of which 40% were classified as human and 59% as viral by SURPI; overall agreement 98.7%).

2.2.2 Classifier Module

Classification in Taxonomer is based on exact kmer matching. Taxonomer uses databases that are optimized for rapid kmer queries that store every reference in which a kmer is found as well as an associated kmer weight for every reference. The fundamental question for classification is how likely it is that a particular kmer (K_i) originates from any reference sequence, ref_i . To answer this question, Taxonomer calculates a kmer weight:

$$KWref_i(K_i) = \frac{C_{ref}(K_i)/C_{db}(K_i)}{C_{db}(K_i)/Total\ kmer\ count} \quad (2.1)$$

where C represents a function that returns the count of K_i . $C_{ref}(K_i)$ indicates the count of the K_i in a particular reference. $C_{db}(K_i)$ indicates the count of K_i in the database. This weight provides a relative, database-specific measure of how likely it is that a kmer originated from a particular reference. In order to classify a query sequence, we calculate the sum of the kmer weights for every reference that has a matching kmer in the query sequence. Suppose that there are N possible kmers from query sequence Q. Then, for every reference, ref_i , that shares a kmer with Q, the total kmer weight for ref_i is:

$$TKW(ref_i) = \sum_{j=1}^N KWref_i(K_j) \quad (2.2)$$

Each read is assigned to the reference that has the maximum total kmer weight. In the case of a tie, the query sequence is assigned to the taxonomic lowest common ancestor (LCA).

2.2.3 Protonomer Module

We developed a mapping scheme between amino acids and their corresponding codons to facilitate mapping in protein space while using the same strategies and speed we developed for classification in nucleotide space. When the amino acid database is built for classification, Taxonomer assigns every amino acid to just one codon. This unique mapping, which we term a *non-degenerate translation*, is used to generate an

artificial DNA sequence that corresponds to the protein sequence in the database. This DNA sequence is entered into Taxonomer's nucleotide classification databases. Query reads are translated into all 6 reading frames using the same non-degenerate translation scheme used to build the database and each translated frame is then classified. Kmer weighting and read classification assignment are performed as described above. The default Protonomer database is subsets of UniRef90 and UniRef50. Empirically, we found a kmer size of 30 (10 amino acids) to perform best. We chose to classify viruses in protein space because of their high mutation rates, genetic variability, and incomplete reference databases (25). Protonomer was benchmarked against two other rapid protein search tools, RAPSearch2 (26) (employed by SURPI) and DIAMOND (27) (an ultrafast, BLAST-like protein search tool), using RNA-seq data from respiratory samples of 24 children with documented viral infections as determined by an FDA-cleared molecular test (eSensor Respiratory Virus Panel, GenMark) or targeted PCR (17), (see Table 2.5) for which complete viral genomes could be manually constructed (Geneious, version 6.1). Viral reads were defined by mapping all reads binned as 'Viral' or 'Unknown' to the manually constructed viral genomes. Sensitivity and specificity were determined based on detection of known viral reads (true positives) and non-viral reads (true negatives). Protonomer provides a single taxonomic identifier per read as the classification assignment, which makes interpretation of results extremely simple. Neither RAPSearch2 nor DIAMOND classify a read; instead they only provide BLAST-like alignment information. For benchmarking against RAPSearch2 and DIAMOND, the LCA of the alignment with the lowest E-value was assigned as the classification. All tools were benchmarked using the viral subset of UniRef90 as their database. Both

Protonomer and RAPSearch2 process paired reads by concatenating them together with a '-' between mate pairs. DIAMOND does not support paired end reads, so each pair was searched separately, and the hit with the lowest e-value from each read was used to make the classification assignments.

2.2.4 Host Gene Expression Estimations

Taxonomer also uses its nucleotide classifier to assign reads to host reference transcripts. By default, these are transcripts and corresponding gene models (GTF file) from the ENSEMBL human reference sequence, GRCh37.75. Empirically, we found that a kmer size of 25 worked best for mapping reads to human transcripts. We benchmarked Taxonomer's gene expression estimates against Sailfish's and Cufflinks' (28) using both biological and synthetic data. We had Taxonomer output all ties between transcripts during the classification step; we then randomly assigned a read to a single transcript. We used these transcript level assignments to calculate gene level expression. We next employed a linear regression to correct for transcript assignment bias in a similar fashion to Sailfish. The reported correlations were then calculated using these corrected values. This level of gene expression analysis is not currently available through the web interface because of the way data are streamed; however, the results given from the web interface are a very good approximation (Spearman correlation > 0.93 on a set of genes that both methods have positive counts and Spearman correlation > 0.75 when the gene set is unrestricted). In the first experiment, we employed qPCR results taken from the microarray quality control study (MAQC) (29); specifically, human brain tissue samples. We also compared performance using synthetic RNA-seq reads (2x76bp, n=15,000,000)

generated with the Flux Simulator tool (30) (see Table 2.6 for parameters). TopHat (31) was used to produce alignments for Cufflinks. Like Taxonomer, Sailfish does not need external alignment information.

2.2.5 Databases

The Classifier and Protonomer databases are modular and easily constructed, consisting only of multi-fasta files with a ‘parent tag’ on their definition lines. These tags describe each reference sequence’s immediate phylogenetic parent-taxon.

2.2.5.1 Bacterial Classification

Bacterial classification is based on a marker gene approach (16S rRNA gene) and the Greengenes database (reference set with operational taxonomic units, OTU, clustered at 99%, version 13_8 (32), (see Table 1.1). This reference set contains 203,452 OTU clusters from 1,262,986 reference sequences. The taxonomic lineage for each OTU was used to create a hierarchical taxonomy map to represent OTU relationships. To support the OTU ‘species’ concept, the taxonomy was completed for ranks in the taxonomic lineage that had no value. Unique dummy species names from the highest taxonomic rank available were used to fill empty values. Versions of the Greengenes database were formatted for use within BLAST, the RDP Classifier, and Kraken.

2.2.5.2 Fungal Classification

Fungal classification is also based on a marker gene approach (internal transcribed spacer, ITS, rRNA sequences) and the UNITE database (33) (version

sh_taxonomy_qiime_ver6_dynamic_s_09.02.2014). This reference set contains 45,674 taxa (species hypothesis, SH) generated from 376,803 reference sequences with a default-clustering threshold of 98.5% and expert taxonomic curation. Dummy names were created for ranks that had no value. Versions of the unite database were formatted for use with BLAST, the RDP Classifier, and Kraken.

2.2.5.3 Viral Classification and Discovery

The virus classification database consists of the viral subset of UniRef90 (release 2014_06) combined with the bacterial subset of UniRef50 (release 2015_03) (34). The viral protein database was reduced to 289,486 viral sequences based on NCBI taxonomy. Phage sequences were separated, leaving a total of 200,880 references for other viruses. NCBI taxonomy was used to determine the sequence relationship. For viral classification and discovery benchmarks and for contig-level classification, only the viral subset of UniRef90 was used.

2.2.5.4 Additional Classification Databases

For testing purposes, additional bacterial classification databases were constructed from RefSeq (identical to Kraken's full database; n=210,627 total references; n=5,242 bacterial references, using NCBI taxonomy), and the complete ribosomal database project databases download on September 24, 2014 (n=2,929,433 references, using RDP taxonomy).

2.2.5.5 Database Construction

Databases are constructed to maximize query speed. Kmers are stored in lexicographical order and kmer minimizers are used to point to blocks of kmers in the database. Once a block of kmers is isolated, a binary search is used to complete the query. This scheme provides extraordinary query speeds, as demonstrated by Wood and Salzberg (35). We employ the same basic database layout as Kraken, with the important difference that instead of storing just the LCA of a kmer, we also store the kmer count and every reference (up to an adjustable cutoff) with associated kmer weight.

2.2.6 Gene Classification Protocols

We extracted reference sequences from widely used, curated public databases for benchmark experiments (36). These reference sequences were used to generate synthetic read datasets having a variety of read-lengths and error rates using wgsim (<https://github.com/lh3/wgsim>). PCR-amplified 16S rRNA gene sequences from two metagenomics studies on stool (37) and the home environment (38) were also used. The analysis was limited to taxa with relative abundance >0.1% per sample (10 random samples were selected from each study).

2.2.6.1 Bacterial 16S rRNA

From the SILVA 119 non-redundant small-subunit ribosomal sequence reference database (36), we extracted bacterial reference sequences between 1200-1650bp of length and excluded references annotated as cyanobacteria, mitochondria, and chloroplasts. Only high-quality references without ambiguous bases, alignment quality values >50%,

and sequence quality >70% were included. All the above values are reported by SILVA. Percent identity to the closest Greengenes OTU was determined by MegaBLAST⁵⁹ using hits with a query coverage >80%. Synthetic reads (100bp single-end, 100bp paired-end, 250 paired-end) were generated from these reference sequences at 5X coverage.

2.2.6.2 Fungal ITS

To test the accuracy of identifying fungal ITS sequences that are not represented in the UNITE database, we utilized the UNITE_public_dataset (version_15.01.14). Percent identity to the closest UNITE species hypothesis (SH, OTU's clustered at 98.5%) was determined by MegaBLAST using hits with a query coverage >80%. Synthetic reads (250bp single-end) were generated from these reference sequences at 5X coverage. Due to the variable length of ITS sequences (mean 585bp, range 51-2,995bp, n=376,803), paired-end sequences were not generated.

2.2.7 Classification Criteria for Reference Methods

2.2.7.1 BLAST

Default MegaBLAST parameters were used. Top scoring references were identified and used to assign OTUs/SHs. Multiple OTUs/SHs were assigned to synthetic reads when more than one OTU/SH reference shared 100% identity. If no OTU/SH had 100% identity to a read, then all OTUs within 0.5% of the top hit were assigned to the read. The taxonomy of the assigned OTUs/SHs was compared and the highest rank in common was used to assign a taxonomic value to the read. The percent identity was used to determine the assignment of the highest taxonomic rank. Sequence reads with >97%

identity to a reference were assigned to species, >90% identity to genus, and <90% to family when lineage information was available at this rank.

2.2.7.2 RDP Classifier

RDP Classifier analyses were performed on a local server (see Section 2.29). Classifications were resolved to the rank with a minimum confidence level of ≥ 0.5 .

2.2.7.3 Kraken

Kraken analyses were performed on a local server (see Section 2.29). Kraken reports the taxon identifier for each read's final taxonomic assignment. An accessory script (Kraken-filter) can be used to apply confidence scores, although we found this value had little impact on results of our benchmarks.

2.2.7.4 SURPI

SURPI analyses were performed using an Amazon EC2 instance through the published Amazon Machine Image. SURPI reports the best hit for its mapping tools (SNAP, RAPSearch2), which were used for comparison.

2.2.8 Taxonomer Implementation

Taxonomer was written in C with Python bindings through Cython. An implementation of Taxonomer that contains the entire pipeline functionality was written in C and drives the iobio web interface.

2.2.9 Server Specifications

Benchmarking was performed on a machine with Red Hat Linux, 1TB of RAM and 80 CPUs. The number of CPUs was restricted to 16 unless otherwise noted.

2.2.10 Web-Service and Visualization

Taxonomer is publically available as a web-service built upon the iobio framework (18). It is available at taxonomer.iobio.io. Complex metagenomic data can be processed quickly and effectively interpreted through web-based visualizations. Figure 2.5 illustrates the interface. As reads are being streamed to the analysis server, a pie chart is presented summarizing the results of the binning procedure. When one of the bacterial, fungal, viral, or phage bins of the pie chart is selected, the results of the Classifier/Protonomer modules are displayed in a sunburst visualization. Additional information is provided at the top of the web page about how many reads were sampled, the number of reads classified, and the detection threshold. The detection threshold informs a user about how abundant a particular organism must be in order to be detected with the number of reads sampled. This provides an indicator of the sensitivity of detection in the sample. In addition, a slider allows the user to select an absolute cutoff for the minimum number of reads required in order to be displayed in the sunburst.

2.2.11 DNA and RNA-seq of Patient Samples

2.2.11.1 Nucleic Acid Extraction

Samples (75-200 μ L) were extracted using the QIAamp Viral RNA extraction kit (Qiagen). Extraction was carried out as described by the manufacturer with the exception

of the AW1 washing step. For this step, 250 μ L of AW1 wash buffer was added to the QIAamp Mini column before centrifugation at 8000 rpm. Then, 80 μ L of DNase I mix (Qiagen) containing 10 μ L of RNase-free DNase I and 70 μ L of Buffer RDD was added to the column for on column DNase digestion. After incubation at room temperature for 15 minutes, an additional 250 μ L of AW1 was added to the column before centrifugation at 8000 rpm. The manufacturer suggested protocol was continued at this point with column washing using Buffer AW2. After all washing steps, RNA was eluted in 60 μ L of water. Extraction for total DNA was performed using 75-200 μ L of sample with the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions. DNA was eluted in 200 μ L of nuclease-free water. This study was approved by the University of Utah (IRB_00035409) and CDC (5827) IRBs.

2.2.11.2 Depletion of Human DNA

Microbial DNA was enriched with NEBNext Microbiome DNA Enrichment Kit (NEB). Briefly, MBD2-Fc-bound magnetic beads were prepared by combining 3 μ L of MBD2-Fc protein with 30 μ L of Protein A Magnetic Beads per sample and placing the mixture in a rotating mixer for 10 min at room temperature before washing with 1X Binding Buffer. Extracted DNA (200ng in 200 μ L) was added to 50 μ L 5X Binding Buffer. The resulting 250 μ L were added to MBD2-Fc-bound magnetic beads for 15 min at room temperature with rotation. The enriched microbial DNA was cleaned-up with Agencourt AMPure XP Beads (Beckman Coulter).

2.2.11.3 Library Generation

For HiSeq and MiSeq sequencing, indexed cDNA libraries were produced from extracted RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina) omitting poly-A selection. RNA was dried and resuspended in 19.5 μ L of Elute, Prime, Fragment Mix. The remainder of the library preparation was conducted per manufacturer's instructions. Before library generation from DNA, enriched microbial DNA was fragmented with the Covaris S2 Ultrasonicator using intensity 5, duty cycle 10%, and 200 cycles/burst for 80 seconds all at 7 °C. Libraries generated from fragmented enriched microbial DNA were prepared using the KAPA Hyper Prep Kit (KAPA Biosystems) according to the manufacturer's instructions. PCR cycles used for library amplification were dependent upon the amount of input DNA and 13 cycles were used for these experiments. Libraries were quantitated by qPCR using the KAPA SYBR FAST ABI Prism qPCR Kit (KAPA BioSciences) and the Applied Biosystems 7900HT Fast Real-Time PCR System (Applied Biosciences). Library size was determined with the Agilent High Sensitivity DNA Kit and Agilent 2100 Bioanalyzer. After pooling of the indexed sequencing libraries, a second qPCR and bioanalyzer run was performed to estimate the final concentration before sequencing. For Ion Proton sequencing, indexed cDNA libraries were produced from extracted RNA using the SMARTer Universal Low Input RNA Kit (Clontech) with numbers of PCR cycles ranging from 10-15 based on RNA yield.

2.2.11.4 Sequencing

Pooled sequencing libraries were analyzed on a HiSeq 2500 (2x100bp), MiSeq (2x250bp, both Illumina), or Ion Proton (median read length 139bp, Life Technologies) instruments according to manufacturers' protocols.

2.2.12 Statistical Analyses

For gene expression analyses, we report both the Pearson and Spearman correlations as was done before (39). Correlation coefficients were calculated using the `scipy` library for python. The Pearson correlation of the log transformed gene expression estimates necessitates the removal of any genes whose estimated expression is 0. The log transform prevents outliers from dominating the correlation. We also report the Spearman correlation, for which the log transform is not as necessary since it is a correlation based on ranks. Thus, the inclusion of genes with estimates of 0 can be avoided.

2.3 Results

To demonstrate the power and utility of Taxonomer, we carried out benchmark analyses using biological and synthetic datasets. These include a large number of pediatric nasopharyngeal (NP)/oropharyngeal (OP) swabs from the Centers for Disease Control and Prevention (CDC) Etiology of Pneumonia In the Community (EPIC) study (17) as well as published data (21,40,41).

2.3.1 Speed and Completeness of Classification

We used RNA-seq data from three virus-positive NP/OP samples with a range of host vs. microbial composition profiles to compare speed and completeness of classification by Taxonomer to two other ultra-fast metagenomics tools: Kraken, and SURPI (see Table 2.5). Kraken was the fastest tool (mean 1.5 min/sample), but classified the fewest reads because it relies on nucleic acid-level classification alone, and uses a single reference database. Although SURPI enables amino acid-level searches for virus detection and discovery, this greatly extended analysis times to between 1.5 and >12 hours/sample. Taxonomer achieved run times similar to Kraken (~5 minutes/sample, 5-8x10⁶ reads/sample), while performing nucleotide and protein-based microbial classification as well as host gene expression profiling. Taxonomer also classified the largest number of reads. Collectively, these results demonstrate how Taxonomer combines the ultrafast speed of Kraken with an extended suite of analysis and search capabilities that exceed those of SURPI.

2.3.2 Bacterial and Fungal Classification

Reads derived from taxa that are absent from classification databases can result in false negative and false positive classifications, especially at the genus and species level (see Figure 2.6). Thus, comprehensive classification databases are essential and several options exist. RefSeq contains whole genome sequences of only ~5,000 bacterial taxa (www.ncbi.nlm.nih.gov/refseq/), whereas more comprehensive 16S rRNA sequence databases (36,42,43) suggest existence of 100,000-200,000 species. As a result, 16S reads from unrepresented bacteria are more readily identified than reads derived from other

genomic targets (see Figure 2.7). To maximize classification accuracy, Taxonomer employs a 16S marker gene approach and a custom Greengenes-derived database.

2.3.3 Default Benchmarks

Performance of classification tools is frequently only tested with synthetic reads derived from the reference database; i.e. perfect matches exist for all synthetic reads. This is a highly artificial challenge, as novel microbial species or strains are routinely encountered in clinical or environmental samples. To provide a more realistic challenge, we generated synthetic reads from phylogenetically diverse 16S sequences almost half (n=468, 46%) of which lacked perfect matches in Taxonomer's reference database (see Figure 2.8). The utility of Taxonomer's kmer weighting approach is illustrated in Figure 2.9, demonstrating superior accuracy compared to SURPI and Kraken when using each tool's default databases and command lines. At the species level, Taxonomer correctly classified 59.5%, incorrectly classified 15.7%, and failed to classify 24.8% of the reads. By comparison, Kraken classified 29% of the reads to the correct species but classified every remaining read (71%) incorrectly. As SURPI aligns each read from a mate pair independently and in many cases best matches are discordant (see Table 2.8), results are shown for correct classification of either (left half) or both read mates (right half). In both analyses, SURPI underperformed Taxonomer and Kraken.

2.3.4 Database Benchmarks

Next, we assessed the effect of three different databases (RefSeq, RDP, and Taxonomer's custom Greengenes-derived database) on Taxonomer's accuracy using the

same synthetic reads (see Figure 2.10). With the Greengenes-derived database, Taxonomer correctly classified 59.5% of the reads at the species level, and recovered 94.9% of species. Using RefSeq (Kraken's default database), Taxonomer's values drop to 27% and 71.6%, respectively, similar to Kraken's results when using the same database: 29% and 71%, respectively. Although Taxonomer misclassified very few reads using the RDP database, overall performance was inferior. Thus, Taxonomer's Greengenes-derived database is its default for bacterial classification.

2.3.5 Algorithmic Benchmarks

To compare accuracy of classification algorithms, we used the same database (Taxonomer's Greengenes-derived db), and classified the same synthetic reads with Taxonomer, MegaBLAST (www.ncbi.nlm.nih.gov/blast/html/megablast.html), RDP Classifier, and Kraken (see Figure 2.11). SURPI was not included, as it provides no means to replace its reference databases. Overall, Taxonomer's performance closely approximated that of the RDP Classifier, an established reference tool (59.5% and 61.4% correct species-level classifications, respectively). Kraken's performance improved using the Taxonomer's Greengenes-derived database, but Taxonomer still correctly classified 13.5% more reads, had a lower false positive rate (15.7% vs. 20.1%), recovered more taxa correctly (94.9% vs. 83%), and had a lower false recovery rate (23.3% vs. 37.9%). Similar performance advantages are also seen for fungal classification and recovery rates using Taxonomer's ITS database (see Figure 2.12). Lastly, we examined the impact of read length and sequencing error rates on classification accuracy (see Figure 2.13 and 2.14). As would be expected, performance improved for all tools as a function of read

lengths. Taxonomer and Kraken were more sensitive to sequencing errors than BLAST and the RDP Classifier, which is not surprising given their reliance on exact kmer matching. Nevertheless, these same analyses demonstrate that Taxonomer's nucleotide classification algorithm is tolerant to ~5% random error, with Taxonomer achieving greater classification accuracies than Kraken on these noisy data.

2.3.6 Bacterial Community Composition

Since quantifying microbial community composition is a frequent goal of metagenomics studies, we compared Taxonomer's bacterial abundance estimates to those of the RDP Classifier using recently published 16S amplicon sequencing (37,38) and RNA-seq-based metagenomics data (see Table 2.9 and Table 2.10). Taxonomer's abundance estimates were highly correlated with RDP's across taxonomic levels for all three datasets. Spearman correlation coefficients (ρ) were 0.96 and 0.997 (order) and 0.858 and 0.826 (genus) for 16S amplicon data as well as 0.992 (order) and 0.955 (genus) for RNA-seq (see Figure 2.15). However, Taxonomer's average analysis times were 260 to 440-fold less than RDP's (see Figure 2.16). Collectively, these benchmarks illustrate the importance of Taxonomer's classification databases and the power and speed of its classification algorithm.

2.3.7 Viral Classification and Discovery

Taxonomer uses reads from the 'viral' and 'unknown' bins for detection of viral and phage sequences via its Protonomer module (see Figure 2.1). We compared Protonomer to two rapid protein search tools, RAPSearch2 (employed by SURPI) and

DIAMOND (an ultrafast, BLAST-like protein search tool), using RNA-seq data from virus-positive, pediatric NP/OP samples (n=24). Protonomer demonstrated the best overall performance, being more sensitive (median 94.6%) than DIAMOND (90.5%) and more specific (90.7%) than RAPSearch2 (88.0%, see Figure 2.17). As expected, sensitivity of all tools correlated with phylogenetic distance of viral strains to reference sequences (see Figure 2.18). DIAMOND was most vulnerable to novel sequence polymorphisms. As DIAMOND does not support joint analysis of paired sequencing reads, results of the mate-pair with the lowest E-value were used, likely resulting in optimistic performance estimates. Protonomer was also the fastest of the three tools in classifying 10^4 to 10^6 reads/sample (median time per sample: Protonomer 14 seconds; DIAMOND 37 to 46 seconds; RAPSearch2 343 to 169 seconds (see Figure 2.17) and Figure 2.18).

To demonstrate Taxonomer's ability to detect viral pathogens in public health emergencies, we analyzed published RNA-seq data from serum of a patient with hemorrhagic fever caused by a novel rhabdovirus (Bas Congo Virus) (40); a throat swab from a patient with avian influenza (H7N9 subtype) (41), and plasma from a patient with Ebola virus (21) (see Figure 2.19). Even after removal of target sequences from the classification database, to simulate detection of unknown pathogens, all three viruses or close relatives were detected, thus demonstrating Taxonomer's utility for rapid virus detection and discovery in public health emergencies.

2.3.8 Human mRNA Transcript Profiling

Quantification of synthetic reads and a commercial RNA standard (29) by Taxonomer was accurate over a broad range of transcript abundance when compared to standard tools (Sailfish, Cufflinks, see Figure 2.20 and Table 2.11). Indeed, Taxonomer's accuracy was intermediate between Sailfish's and Cufflinks', demonstrating state-of-the-art performance. To highlight utility of simultaneous pathogen detection and transcript expression profiling, we compared human mRNA expression profiles directly from respiratory samples of patients with influenza A virus infection (17) (cases, n=4) and asymptomatic controls (n=40, see Figure 2.21). Influenza A virus was detected in all case samples (see Figure 2.22). Expression profiles for 17 human genes were significantly higher in cases, and clearly differentiated cases from controls (see Figure 2.23 and 2.24 and Table 2.12). As expected, Gene Ontology (44) assignments for the top 50 genes demonstrated their involvement in recognizing pathogen-associated molecular patterns and in the antiviral host response (see Table 2.13). Most but not all of these genes are known players in the host response to viral infections (www.ncbi.nlm.nih.gov/biosystems/217173). Together, these results demonstrate the accuracy and power for discovery and diagnostic application of Taxonomer's combined pathogen detection and host response profiling.

2.3.9 Microbial Detection in Real-World Scenarios

In Figure 2.25, we show that Taxonomer can be used to detect previously unrecognized infections, to identify microbial contamination in RNA-seq data, and to analyze data from commonly used next-generation sequencers. In RNA-seq data from

test-negative patients with suspected Ebola virus disease, Taxonomer detected a range of other reported infections (21) (HIV, Lassa virus, Enterovirus - typed by Taxonomer as Coxsackievirus, GB virus C). However, Taxonomer also identified previously unrecognized bacterial infections (*Chlamydophila psittaci*, *Elizabethkingia meningoseptica*) that may have caused the patients' symptoms (see Figure 2.25A, and 2.26). Taxonomer's power for virus discovery was demonstrated by analyzing RNA-seq data from an NP/OP sample (17) that contained a novel anellovirus with only 44%-60% predicted protein sequence identities to the most similar sequenced strain (see Figure 2.27). While 44 of 239 anellovirus reads were classified to the family Anelloviridae at the read-level (see Figure 2.25B), analysis of contigs assembled from all reads binned by Taxonomer as 'viral' and 'unknown' could be leveraged to further boost sensitivity, which resulted in detection of 4 contigs (representing all 239 reads) to the family Anelloviridae (data not shown). Taxonomer can also be leveraged to quality control next-generation sequencing data (45-50). To demonstrate this, we analyzed RNA-seq data from induced pluripotent stem cell cultures with and without *Mycoplasma* contamination (see Figure 2.25B). Taxonomer identified 56% of reads as bacterial and classified the contaminant as *M. yeatsii*. Lastly, Taxonomer produced highly comparable results when the same two respiratory samples were sequenced on three popular instruments (MiSeq, HiSeq, Ion Proton). In all three cases, similar proportions of reads were classified to known viral (influenza A) and bacterial (*Mycoplasma pneumoniae*) pathogens (see Figure 2.25D and Table 2.14) demonstrating that Taxonomer is compatible with the different read lengths and error profiles of these sequencing platforms.

2.4 Discussion

In Taxonomer, we have created a tool that is fast, accurate, and capable of the gamut of analyses required to take full advantage of large and complex DNA and RNA-seq datasets for metagenomics. Taxonomer provides means for nucleotide and protein based homology searches, phylogenetic classification at the read and contig level, and host transcriptional profiling. As a result, Taxonomer provides greater accuracy and comprehensive taxonomic profiling than fast alignment-free tools (e.g. Kraken), while providing 10-100X faster classification and greater accuracy than comprehensive alignment-based tools (e.g. SURPI). In addition, Taxonomer achieves accuracies on 16S amplicon data that closely approach the current standard, RDP. This is made possible by Taxonomer's comprehensive databases and its novel kmer weighting approach, which combine to reliable bacterial community profiling from RNA-seq data in which 16S sequences are highly abundant. Moreover, Taxonomer is very fast, requiring only a few minutes to carry out its broad array of analyses. On the same typical HiSeq 2500 datasets, Taxonomer is days faster than RDP, hours faster than SURPI, and within minutes of the fastest published tool, Kraken, which only provides nucleotide classification.

Taxonomer provides maximum flexibility for detection of known and unknown bacteria, fungi, and viruses. As the vast majority of bacteria, fungi, and viruses remain unknown (25,51-53), reference databases are inevitably incomplete. As we demonstrated, Taxonomer's marker gene-based approach for bacterial and fungal identification and discovery leverages large databases that provide maximum taxonomic information, which helps avoid misclassifications pitfalls (54).

Taxonmer's integrated means for protein-based classification further improves its sensitivity, especially for virus detection where nucleotide-based classification is of limited utility due to high mutation rates and high sequence diversity in many viral phyla. Moreover, our results demonstrate the power of Taxonomer in real-world scenarios by identifying known viruses (respiratory viruses, HIV, Lassa virus, Coxsackievirus, GB virus C, Bas Congo Virus, avian influenza A virus H7N9) and unrecognized bacteria and viruses in previously test-negative patients (*Anellovirus*, *Chlamydophila psittaci*, *Elizabethkingia meningoseptica*).

Host gene expression profiling, part of Taxonomer's integrated analysis architecture, is of growing interest for infectious diseases testing (55). While host gene expression profiles can differentiate viral from bacterial infections using blood samples (14-16), Taxonomer enables simultaneous pathogen detection and gene expression profiling from routinely collected respiratory samples. This may eliminate the need for a blood draw, improve diagnosis and discovery, and enable novel applications such as differentiating true infections from asymptomatic carriage, characterizing chronic infections in immunocompromised patients, and monitoring antimicrobial treatment success.

As we demonstrated, Taxonomer can also be used to rapidly identify microbial contamination in RNA-seq studies, which can confound transcriptional response profiles (48) or lead to unsafe biological interpretations (56). Contamination by exogenous sequences directly or through commonly used laboratory reagents have led to erroneous disease associations and genome assemblies, further highlighting quality control applications for Taxonomer (45-50). This is of particular concern when source DNA or

RNA is of low concentration, as with single-cell sequencing studies (57). Lastly, metagenomic sequencing data are usually purged of host sequences prior to deposition in public databases to guarantee anonymity of patients. During analysis of some such sequences (21), varying numbers of human reads were detected, suggesting that Taxonomer is more effective at detecting (and removing) host-derived sequences than currently used tools. Therefore, screening of metagenomics datasets with Taxonomer prior to submission could improve protection of study subjects' privacy. Taxonomer is the only ultrafast metagenomics tool that combines all analytical modalities necessary for these applications.

Finally, with Taxonomer we have sought to democratize these analyses by providing a fast, interactive web service based upon the iobio (58) visualization toolkit. The ability to conveniently upload and rapidly analyze RNA-seq data from patient samples using personal computers and mobile devices means that results can be quickly shared and reviewed by experts, even across great geographic distances enhancing collaborations and facilitating public health responses. As costs and turn-around times for high-throughput sequencing continue to fall and mobile sequencers become available (59), Taxonomer will enable diagnostic laboratories to analyze high-throughput sequencing data in meaningful timeframes without costly computational infrastructure or specialized bioinformatics expertise.

2.5 References

1. Garrett WS. Cancer and the microbiota. *Science* (New York, NY). American Association for the Advancement of Science; **2015**; 348(6230):80–86.

2. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*. Nature Publishing Group; **2014**; 12(10):661–672.
3. Cox LM, Blaser MJ. Antibiotics in early life and obesity. *Nat Rev Endocrinol*. Nature Publishing Group; **2015**; 11(3):182–190.
4. Collins SM. A role for the gut microbiota in IBS. *Nat Rev Gastroenterol Hepatol*. **2014**; 11(8):497–505.
5. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat Med*. Nature Publishing Group; **2013**; 19(5):576–585.
6. Yurkovetskiy LA, Pickard JM, Chervonsky AV. Microbiota and autoimmunity: exploring new avenues. *Cell Host Microbe*. **2015**; 17(5):548–552.
7. Shakya M, Quince C, Campbell JH, Yang ZK, Schadt CW, Podar M. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ Microbiol*. **2013**; 15(6):1882–1899.
8. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*. Nature Publishing Group; **2013**; 499(7457):219–222.
9. Norman JM, Handley SA, Baldrige MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*. **2015**; 160(3):447–460.
10. Hunter P. The secret garden's gardeners. Research increasingly appreciates the crucial role of gut viruses for human health and disease. *EMBO Rep*. EMBO Press; **2013**; 14(8):683–685.
11. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med*. **2014**; 370(25):2408–2417.
12. Lipkin WI. The changing face of pathogen discovery and surveillance. *Nature Publishing Group*. Nature Publishing Group; **2013**; 11(2):133–141.
13. Chiu CY. Viral pathogen discovery. *Curr Opin Microbiol*. **2013**; 16(4):468–478.
14. Zaas AK, Burke T, Chen M, McClain M, Nicholson B, Veldman T, et al. A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. *Sci Transl Med*. American Association for the Advancement of Science; **2013**; 5(203):203ra126–203ra126.

15. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*. **2009**; 6(3):207–217.
16. Hu X, Yu J, Crosby SD, Storch GA. Gene expression profiles in febrile children with defined viral and bacterial infection. *Proc Natl Acad Sci USA*. *National Acad Sciences*; **2013**; 110(31):12792–12797.
17. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. **2015**; 372(9):835–845.
18. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods*. *Nature Publishing Group*; **2014**; 11(12):1189–1189.
19. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. Jordan IK, editor. *PLOS ONE*. *Public Library of Science*; **2013**; 8(10):e76935.
20. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*. *Cold Spring Harbor Lab*; **2014**; 24(7):1180–1192.
21. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (New York, NY)*. *American Association for the Advancement of Science*; **2014**; 345(6202):1369–1372.
22. Audano P, Vannberg F. KAnalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics*. *Oxford University Press*; **2014**; 30(14):2070–2072.
23. Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatrica*. *Blackwell Publishing Ltd*; **2007**; 96(5):644–647.
24. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and more accurate sequence alignment with SNAP. 2011.
25. Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrel CM, Solovyov A, et al. A strategy to estimate unknown viral diversity in mammals. *mBio*. *American Society for Microbiology*; **2013**; 4(5):e00598–13–e00598–13.
26. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*. *Oxford University Press*; **2012**; 28(1):125–126.

27. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. **2015**; 12(1):59–60.
28. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. Nature Publishing Group; **2010**; 28(5):511–515.
29. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*. Nature Publishing Group; **2006**; 24(9):1151–1161.
30. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*. Oxford University Press; **2012**; 40(20):10073–10083.
31. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. Oxford University Press; **2009**; 25(9):1105–1111.
32. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. American Society for Microbiology; **2006**; 72(7):5069–5072.
33. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et al. Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol*. **2013**; 22(21):5271–5277.
34. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*. Oxford University Press; **2007**; 23(10):1282–1288.
35. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. BioMed Central Ltd; **2014**; 15(3):R46.
36. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Priesse E, Quast C, et al. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*. Oxford University Press; **2014**; 42(Database issue):D643–8.
37. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science (New York, NY)*. American Association for the Advancement of Science; **2014**; 345(6200):1048–1052.
38. Subramanian S, Huq S, Yatsunenkov T, Haque R, Mahfuz M, Alam MA, et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*. Nature Publishing Group; **2014**; 510(7505):417–421.

39. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*. Nature Publishing Group; **2014**; 32(5):462–464.
40. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. Wang D, editor. *PLoS Pathog*. Public Library of Science; **2012**; 8(9):e1002924.
41. Hu Y, Lu S, Song Z, Wang W, Hao P, Li J, et al. Association between adverse clinical outcome in human disease caused by novel influenza A H7N9 virus and sustained viral shedding and emergence of antiviral resistance. *Lancet*. **2013**; 381(9885):2273–2279.
42. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. Nature Publishing Group; **2012**; 6(3):610–618.
43. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*. Oxford University Press; **2014**; 42(Database issue):D633–42.
44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*. Nature Publishing Group; **2000**; 25(1):25–29.
45. Cantalupo PG, Katz JP, Pipas JM. HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18. Beemon KL, editor. *J Virol*. American Society for Microbiology; **2015**; 89(8):4051–4057.
46. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. PeerJ Inc; **2014**; 2(7):e675.
47. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol*. American Society for Microbiology; **2013**; 87(22):11966–11977.
48. Olarerin-George AO, Hogenesch JB. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Research*. Oxford University Press; **2015**; 43(5):2535–2542.
49. Smuts H, Kew M, Khan A, Korsman S. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *J Virol*. American Society for Microbiology; **2014**; 88(2):1398–1398.

50. Strong MJ, Xu G, Morici L, Bon-Durant SS, Baddoo M, Lin Z, et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. Rall GF, editor. PLoS Pathog. Public Library of Science; **2014**; 10(11):e1004437.
51. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. Nature Publishing Group; **2013**; 499(7459):431–437.
52. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology. Nature Publishing Group; **2014**; 12(9):635–645.
53. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et al. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol. **2013**; 22(21):5271–5277.
54. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Systems. **2015**; 1(1):72–87.
55. Hudson LL, Woods CW, Ginsburg GS. A novel diagnostic approach may reduce inappropriate antibiotic use for acute respiratory infections. Expert Review of Anti-infective Therapy. Taylor & Francis; **2014**; 12(3):279–282.
56. Mariotti E, D'Alessio F, Mirabelli P, Di Noto R, Fortunato G, Del Vecchio L. Mollicutes contamination: a new strategy for an effective rescue of cancer cell lines. Biologicals. **2012**; 40(1):88–91.
57. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. Gilbert T, editor. PLOS ONE. Public Library of Science; **2014**; 9(10):e110808.
58. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. Nature Methods. Nature Publishing Group; **2014**; 11(12):1189–1189.
59. E CH. Pint-sized DNA sequencer impresses first users. Nature. Nature; **2015**; 521(7550):15–16.

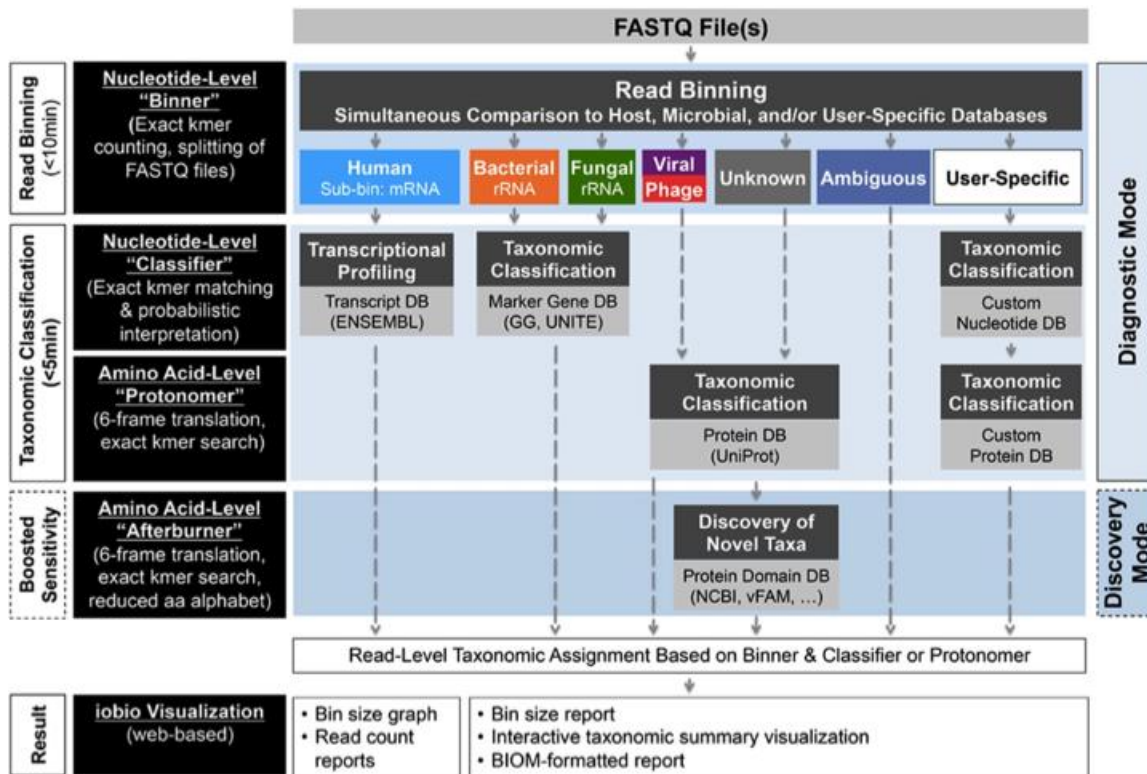


Figure 2.1. Taxonomer's architecture. Raw FASTA, FASTQ, or SRA files (with or without gzip compression) are the input for Taxonomer. For paired-end data, mate pairs are analyzed jointly. Taxonomer consists of four main modules. The 'Binner' module categorizes ('bins') reads into broad taxonomic groups (host and microbial) followed by comprehensive microbial and host gene expression profiling at the nucleotide ('Classifier' module) or amino acid-level ('Protonomer' and 'Afterburner' modules). Normalized host gene expression (gene-level read counts) and microbial profiles can be downloaded. Read subsets can be downloaded for custom downstream analyses.

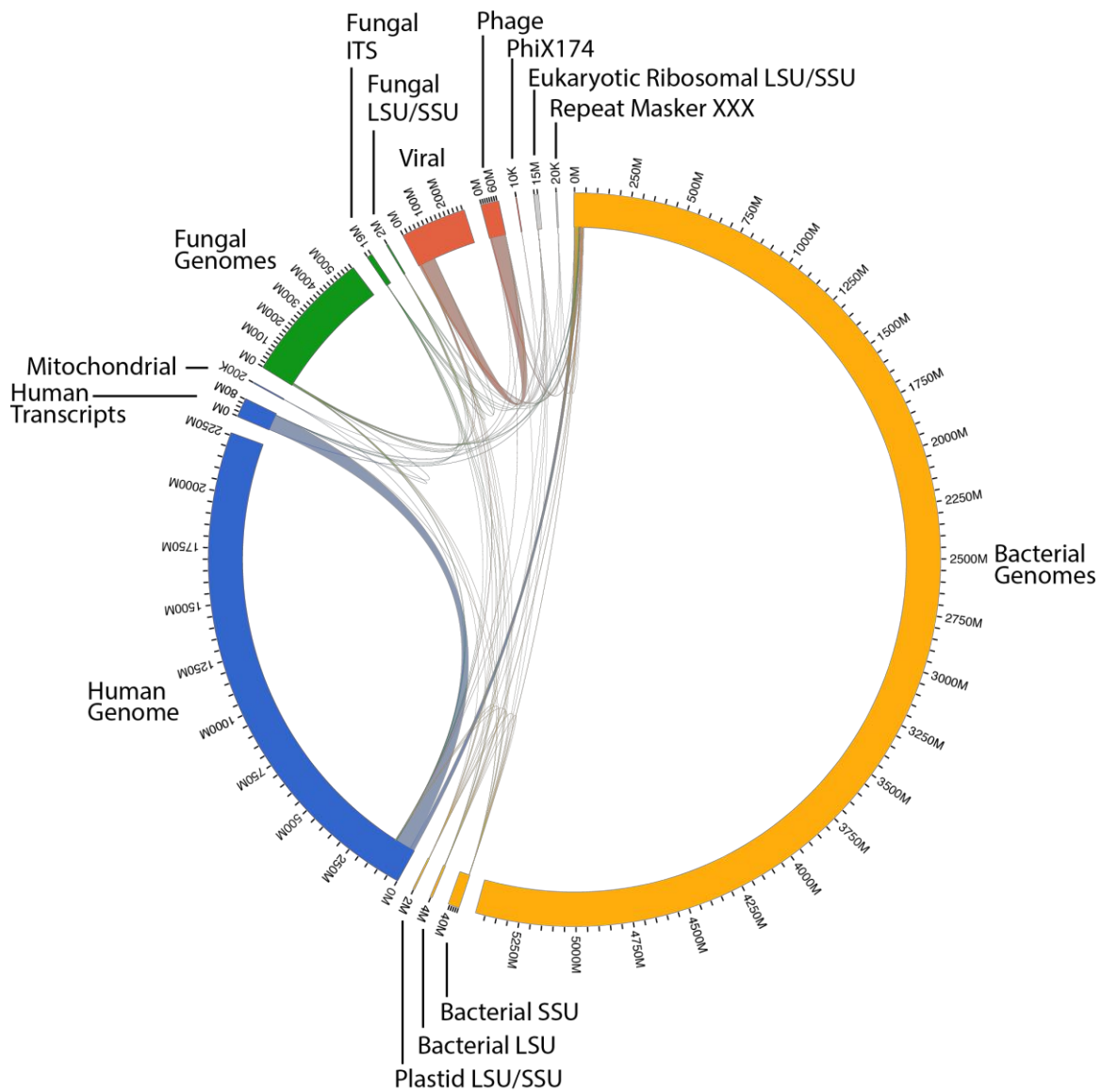


Figure 2.2. Intersections of 21-mers in Binner module databases. The widths of the cords that connect different sections indicate the number of intersecting kmers.

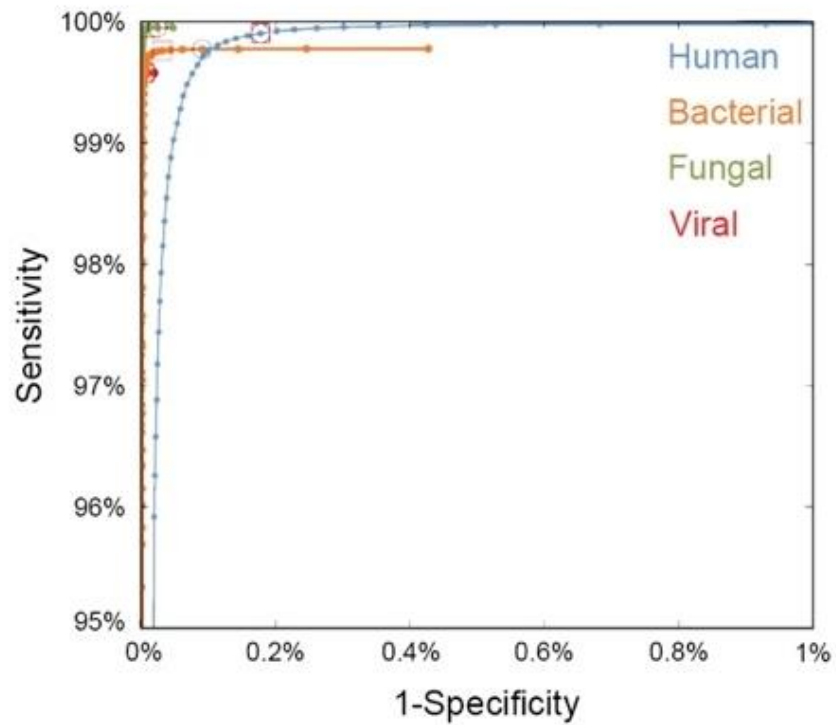


Figure 2.3. Receiver operator characteristics curves for classification of human and microbial sequences by the Binner module. Boxed and circled thresholds represent optimal cutoffs as determined by F1 score and Youden's index, respectively.

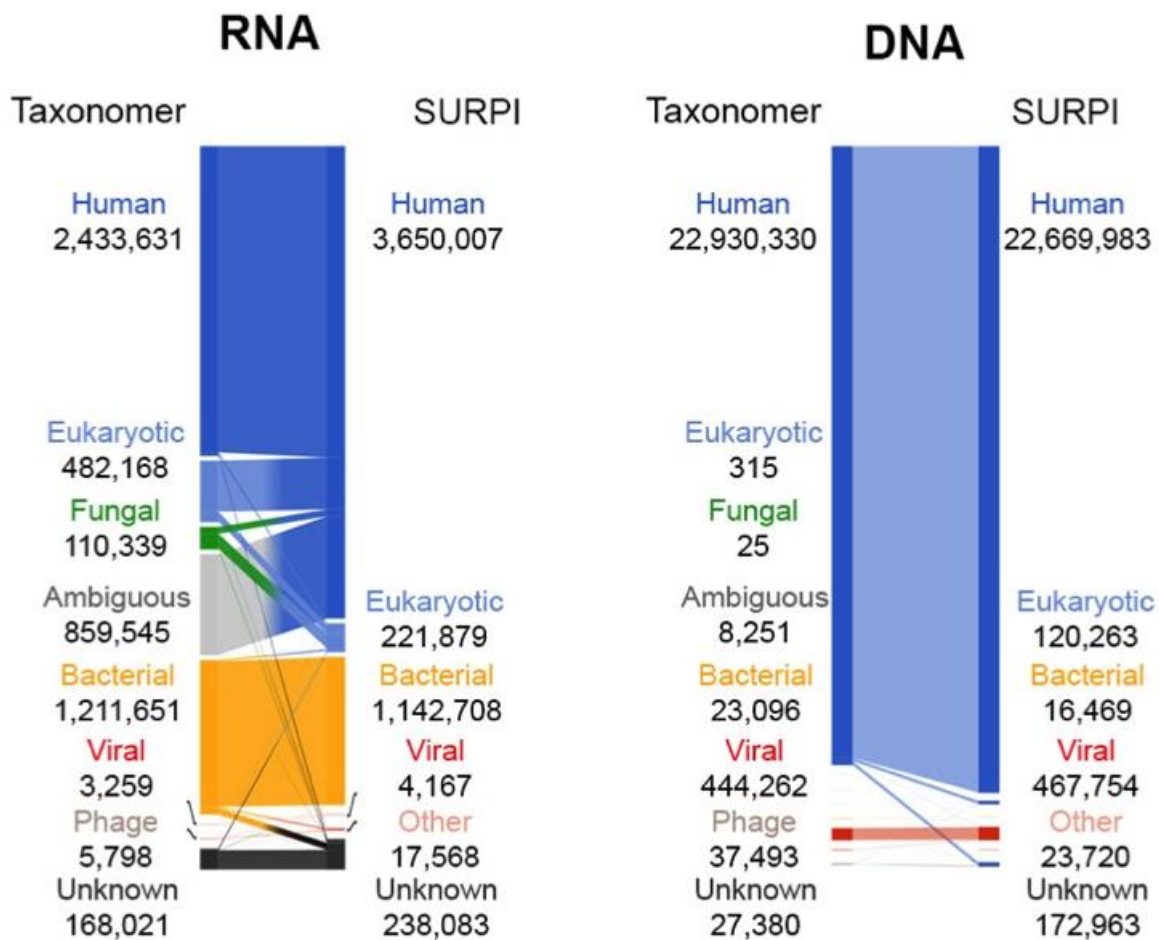


Figure 2.4. Agreement between read binning by the Binner module versus SURPI's assignments. SURPI's assignments are based upon sequential subtraction.

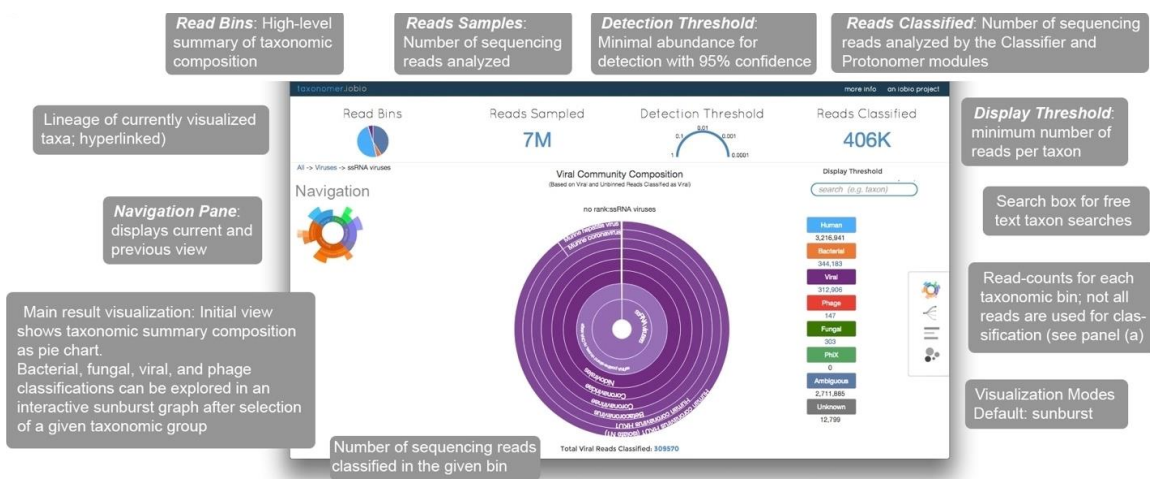


Figure 2.5. Taxonomer web-service. Taxonomic classification of bacteria, fungi, and viruses is visualized as a sunburst graph (center), in which the size of a given slice represents the relative abundance at the read level. Taxonomic ranks are shown hierarchically with the highest rank in the center of the graph. Sequences that cannot be classified to the species level, either because they are shared between taxa or represent novel microorganisms, are collapsed to the lowest common ancestor and shown as part of slices that terminate at higher taxonomic ranks (e.g. genus, family).

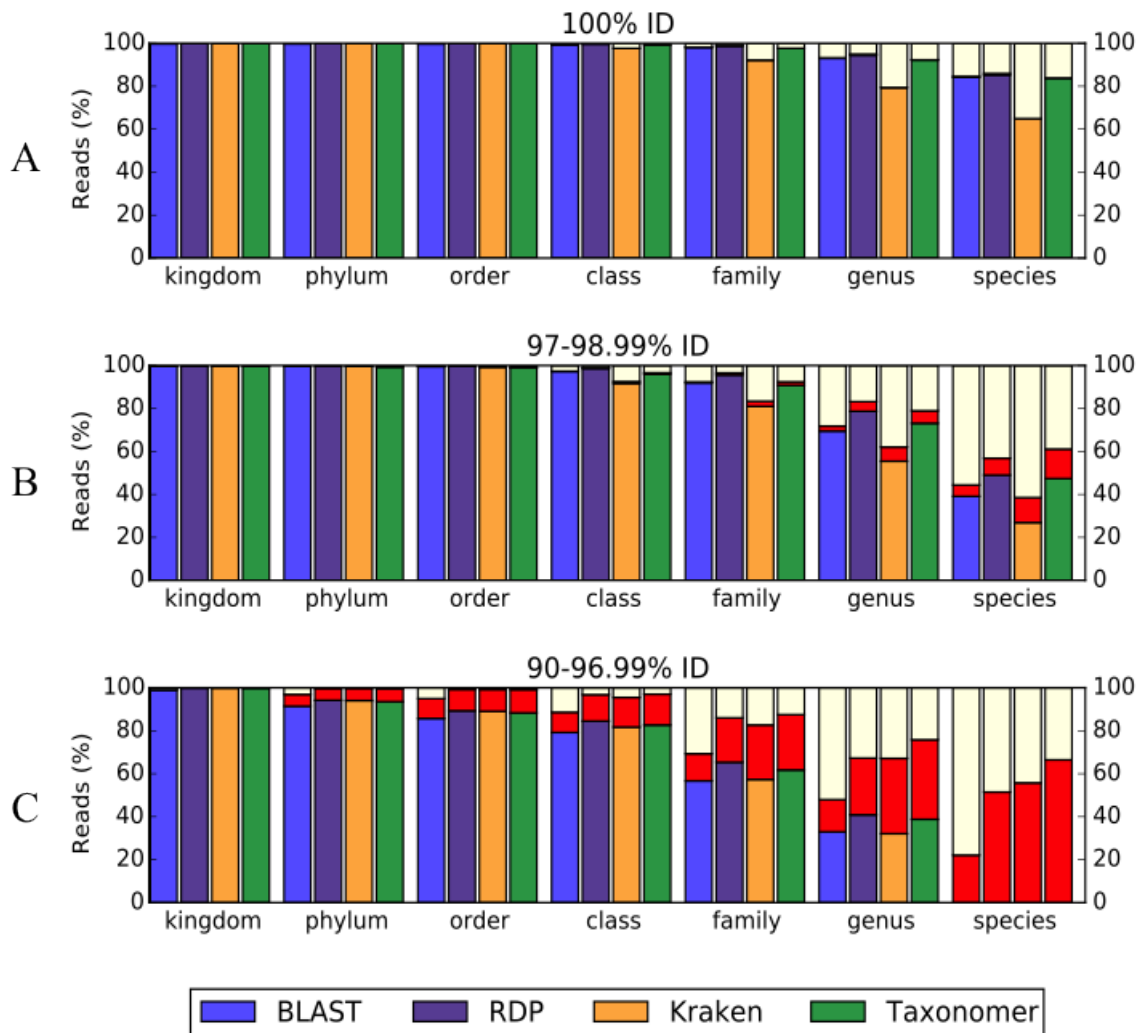


Figure 2.6. False-positive and false-negative classifications of query sequences not represented in the reference database. (A) Read-level classification accuracy for synthetic reads simulated (20X coverage) from SILVA references (n=10,000) with identical representation in the reference database as classified by BLAST, the RDP Classifier, Kraken, and Taxonomer. (B) Panel b shows the same analysis with SILVA references (n=10,000) for whom highly similar, but non-identical references (97% to 98.99% pairwise sequence identity based on full-length MegaBLAST) are present in the reference database. (C) This effect is even more pronounced for synthetic reads simulated from SILVA references (n=10,000) that only share 90% to 96.99% pairwise sequence identity with the closest match in the reference database (based on full-length MegaBLAST). All studies were performed with 250bp paired-end 16S rDNA reads simulated at 20X coverage from randomly selected SILVA references with no error.

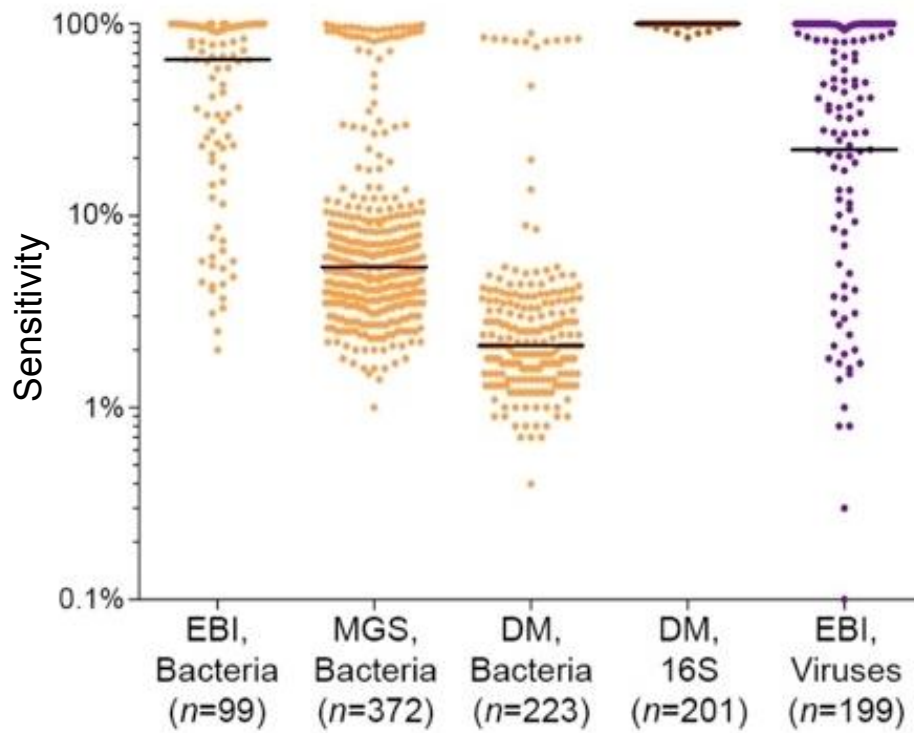


Figure 2.7. Sensitivity for binning of bacterial and viral reads can be low for phylogenetically distant species. Synthetic bacterial and viral reads were generated from single-cell sequencing-based draft bacterial genomes, bacterial genome scaffolds derived from metagenomic sequencing data, and recently published genome sequences.

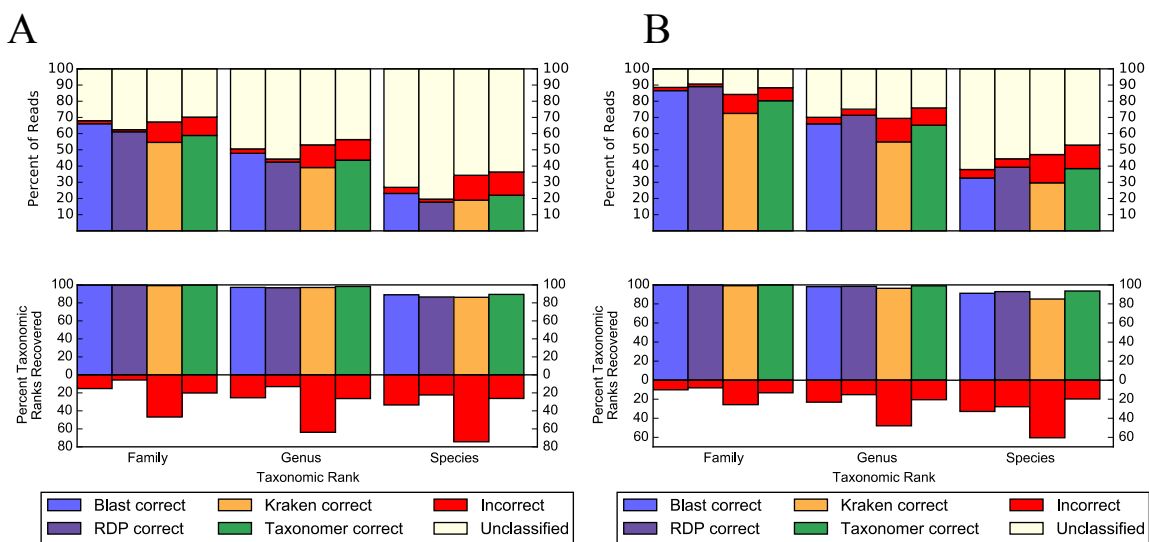


Figure 2.8. Read-level (top) and taxon-level (bottom) bacterial classification accuracy of BLAST, the RDP Classifier, Kraken, and Taxonomer. All tools with the Greengenes 99% OTU database using (A) 100bp single-end and (B) 100bp paired-end 16S rDNA reads simulated at 5X coverage from 1,013 randomly selected SILVA references with $\geq 97\%$ sequence identity to reference sequences.

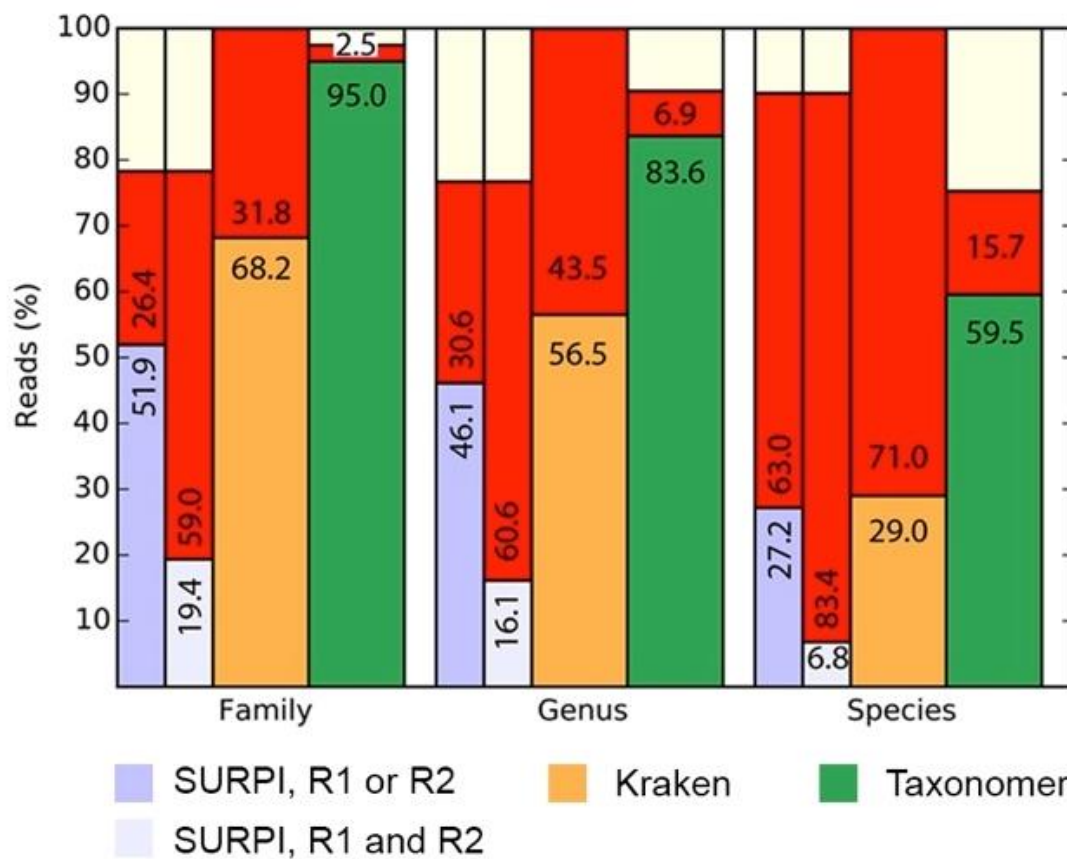


Figure 2.9. Taxonomer sensitivity and specificity for read-level bacterial classification compared to two other rapid classification tools SURPI and Kraken. Analysis performed with each tool's default settings and databases: nt (www.ncbi.nlm.nih.gov/nucleotide, SURPI), RefSeq (Kraken), and Greengenes 99% OTU (Taxonomer). Results for SURPI are based on correct identification by either (dark bar) or both (light bar) read mates.

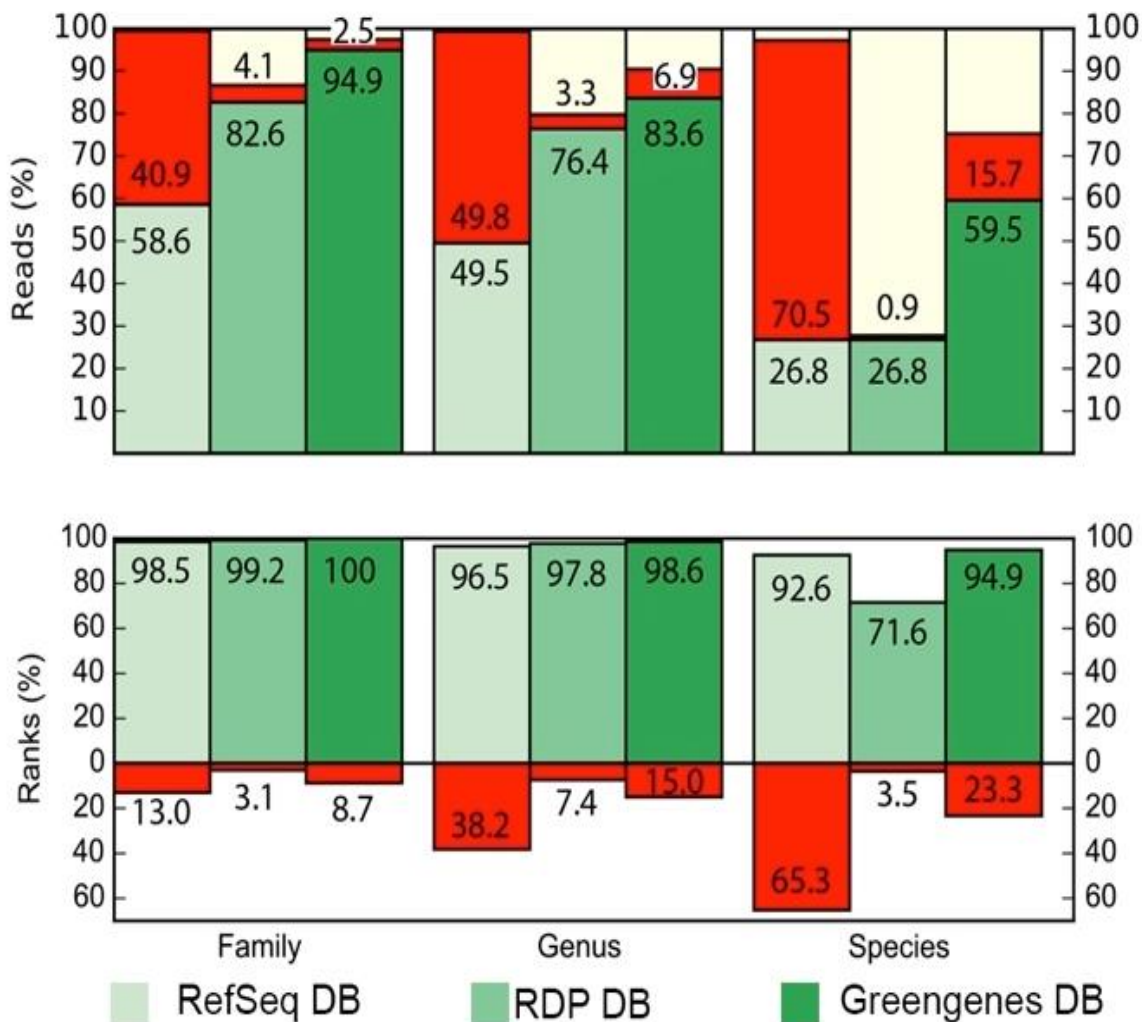


Figure 2.10. Comparison of three commonly used reference databases in Taxonomer. RefSeq (n=210,627; 5,242 bacterial genomes), Greengenes 99% OTU (n=203,452), and RDP (n= 2,929,433). Taxonomer provides greatest read-level (top) and taxon-level (bottom, i.e. percentage of bacterial species identified) sensitivity for bacterial classification at only a moderate decrease in specificity when using the Greengenes database compared to the RDP and RefSeq databases (simulated 16S rDNA as in panel a). Because of its large size and greater completeness, the RDP database provides the greatest species-level specificity at the tradeoff of sensitivity. For ease of reference, the top right-most column is repeated from panel a.

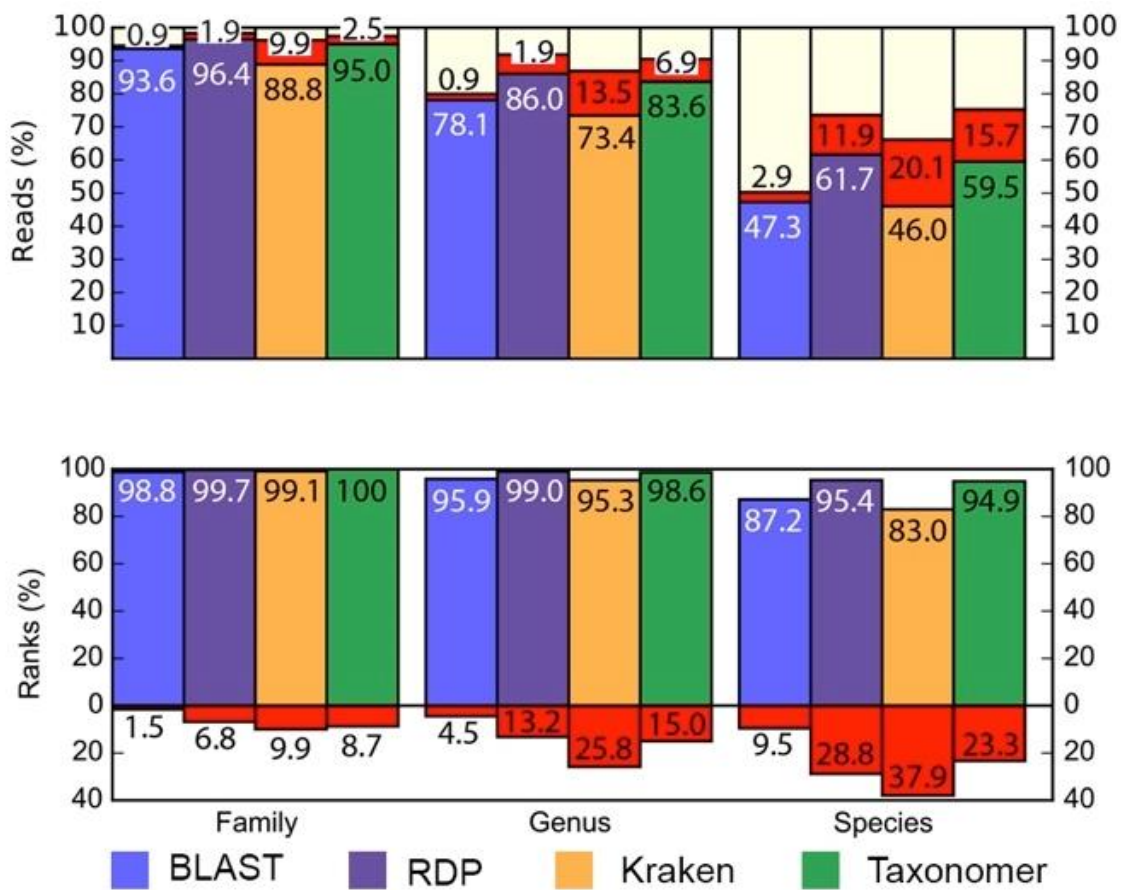


Figure 2.11. Bacterial 16S rRNA classification accuracy. Taxonomer classification is similar to the RDP Classifier and superior to Kraken at the read-level (top) and taxon-level (bottom, all using the Greengenes database). Given the applied criteria, BLAST is less sensitive but more specific.

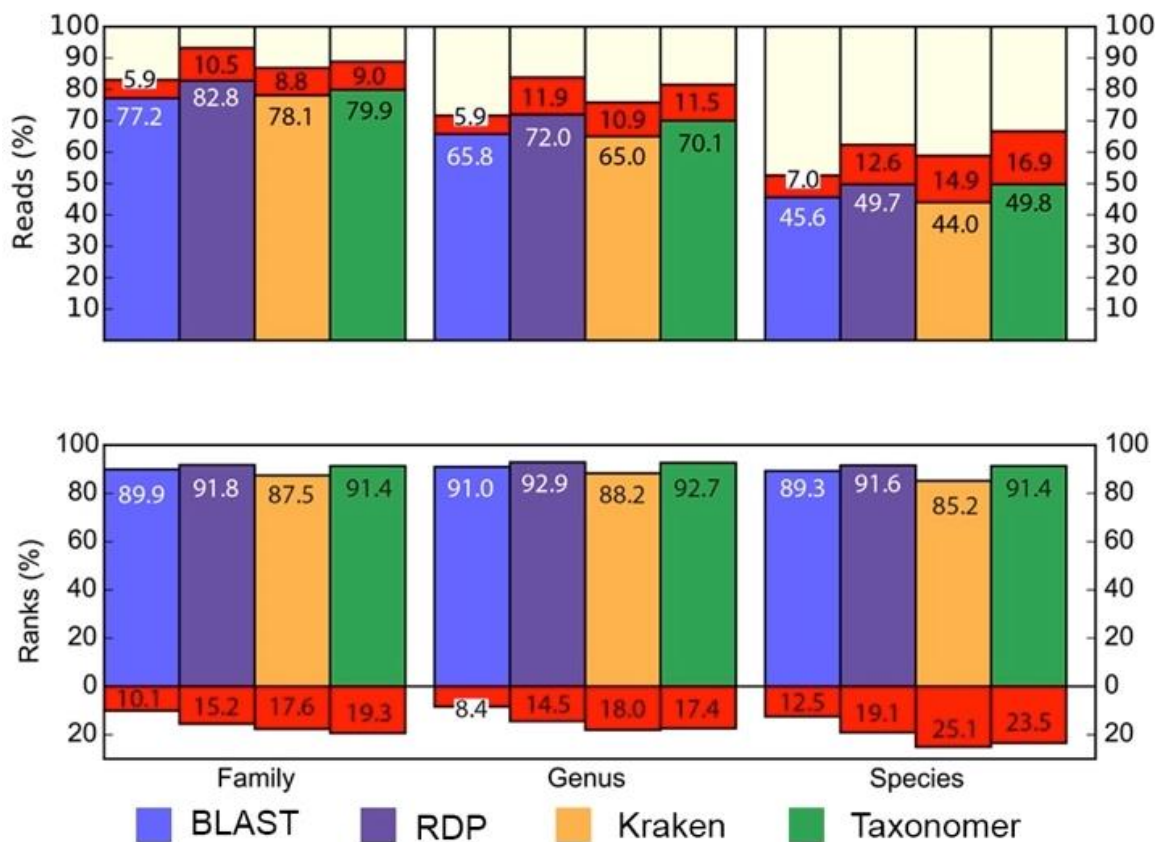


Figure 2.12. Fungal ITS classification accuracy. Taxonomer performs similar to the RDP Classifier and better than Kraken for classification of synthetic fungal internal transcribed spacer (ITS) sequences at the read-level (top) and taxon-level (bottom).

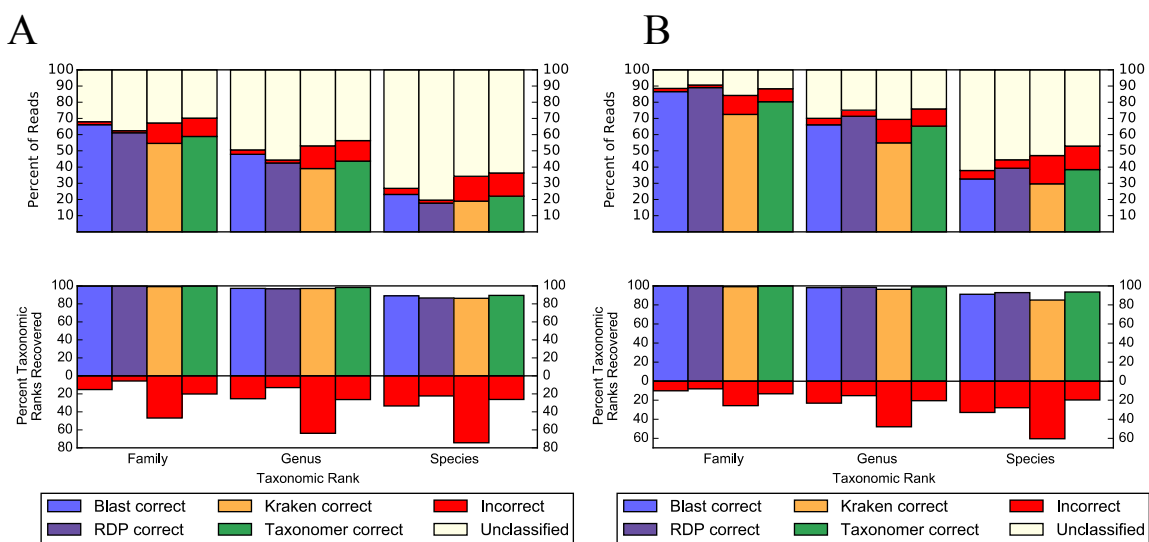


Figure 2.13. Effect of Sequence length on classification 16SrRNA reads. Read-level (top) and taxon-level (bottom) bacterial classification accuracy of BLAST, the RDP Classifier, Kraken, and Taxonomer (all tools with the Greengenes 99% OTU database) using (A) 100bp single-end and (B) 100bp paired-end 16S rDNA reads simulated at 5X coverage from 1,013 randomly selected SILVA references with $\geq 97\%$ sequence identity to reference sequences (see methods). Performance of Taxonomer is comparable to the RDP Classifier and superior to Kraken; given the applied criteria, BLAST is less sensitive but more specific.

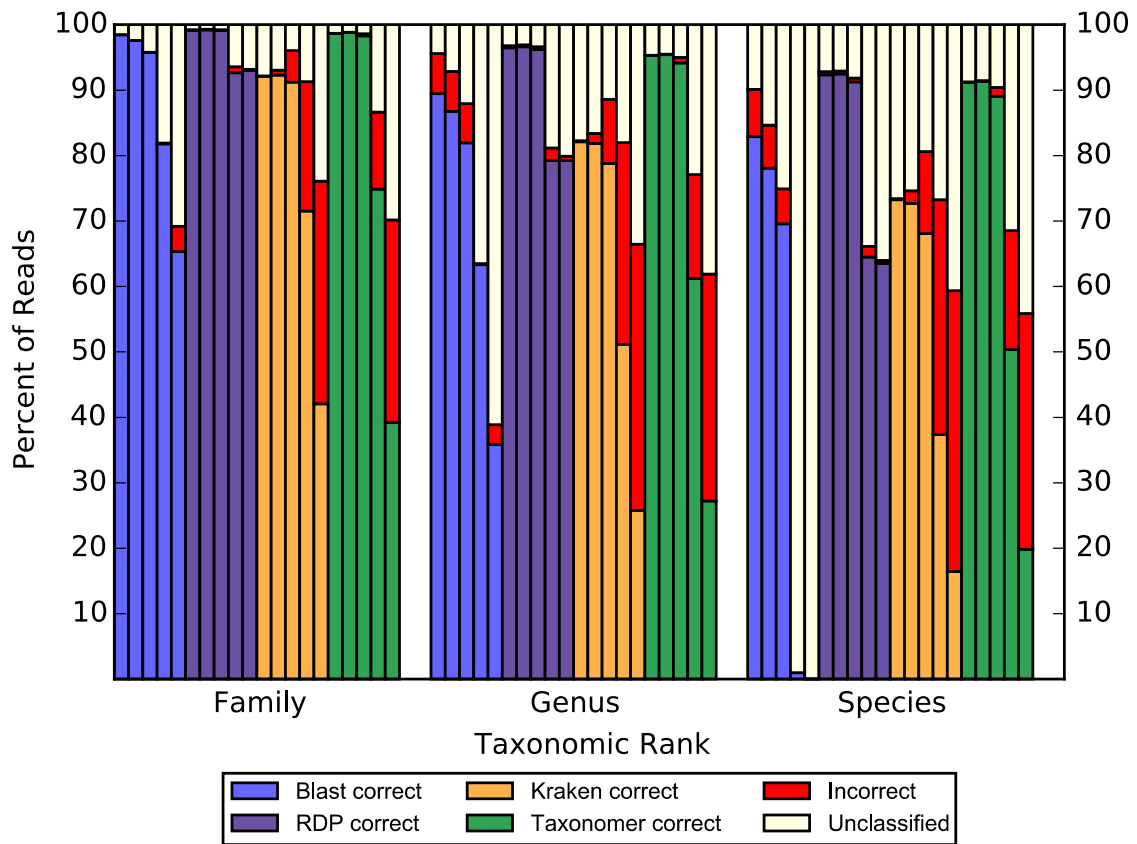


Figure 2.14. Impact of sequencing error rates. Family, genus, and species level classification accuracy for BLAST, the RDP Classifier, Kraken, and Taxonomer using the same read-length and database across error rates of 0.01%, 0.1%, 1%, 5%, and 10%.

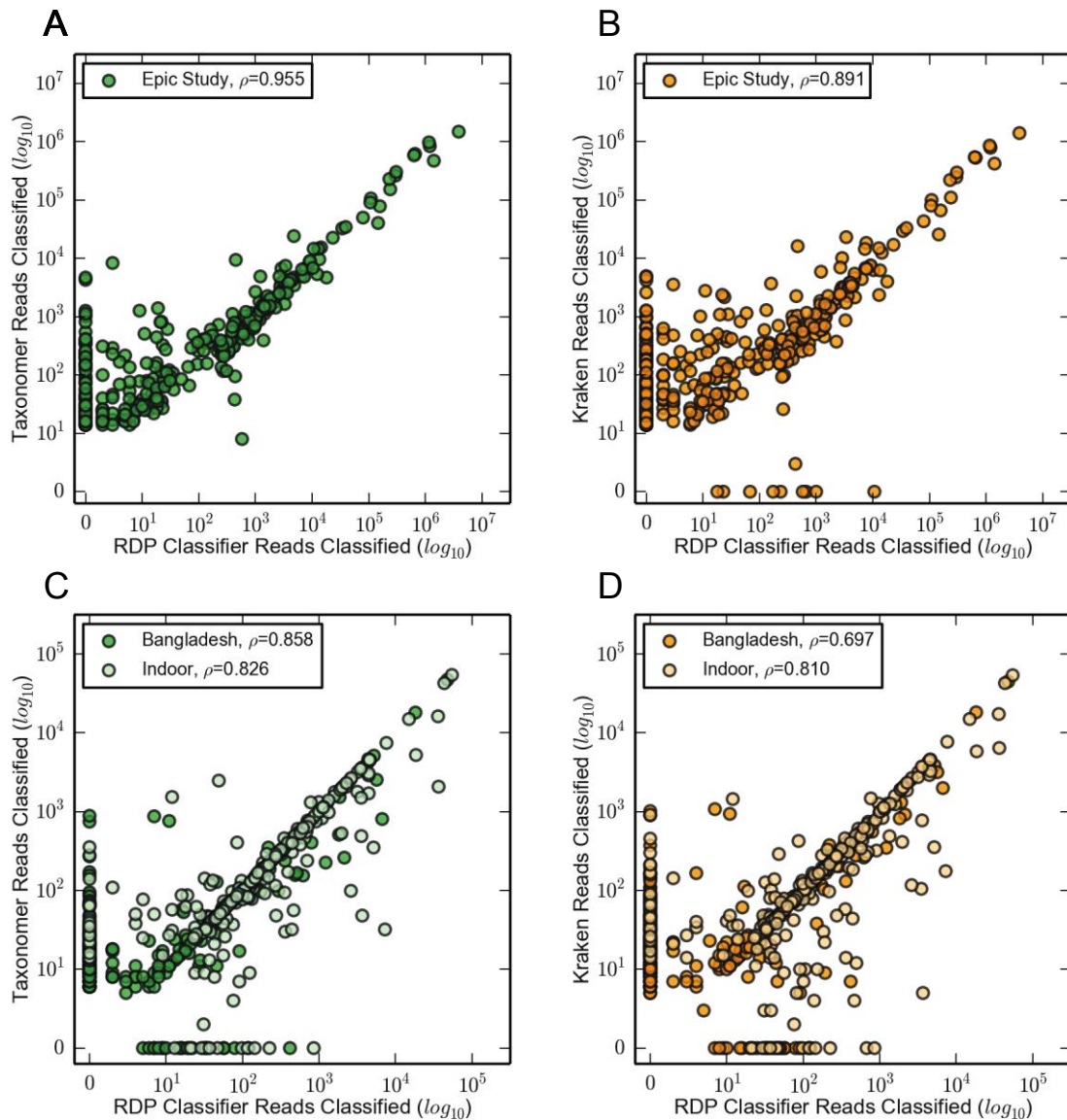


Figure 2.15. Classification of metagenomics data compared to the RDP classifier (A) Bacterial community profiling using RNA-Seq-based shotgun metagenomics with pediatric nasopharyngeal and oropharyngeal swab samples ($n=20$) with Taxonomer and the RDP Classifier at the genus-level. (B) RNA-Seq metagenomics results (as in panel a) were also analyzed by Kraken using the Greengenes 99% OTU reference database. (C) 16S rRNA gene amplicon sequences of variable region 4 from 2 published data sets generated on HiSeq2000 (dark green, 1x150bp reads) and MiSeq instruments (light green, 2x150 reads). (D) 16S rRNA gene amplicon sequences (as in panel c) were also analyzed by Kraken using the Greengenes 99% OTU reference database.

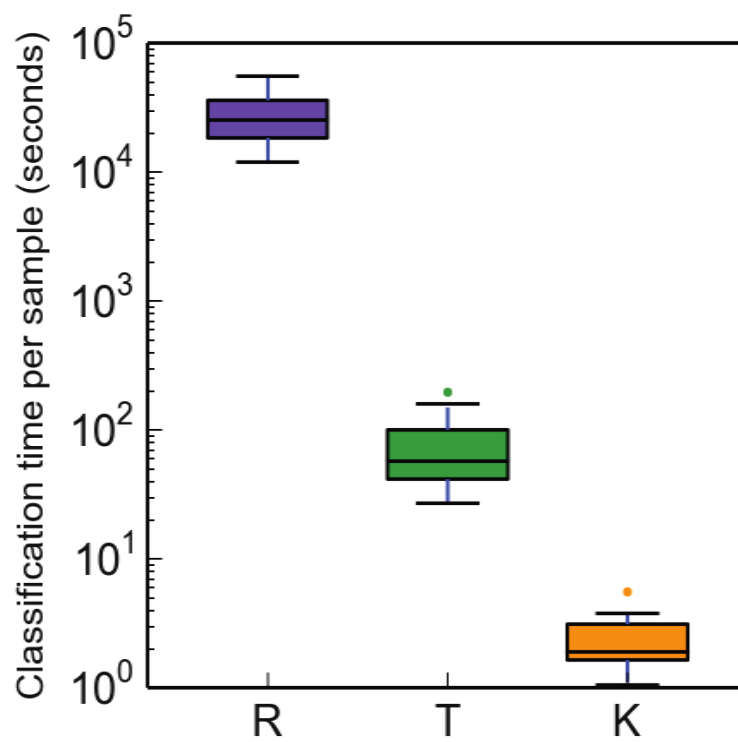


Figure 2.16. Analysis times for the RDP Classifier (R), Taxonomer (T), and Kraken (K). Time for classification of samples shown in Figure 2.15.

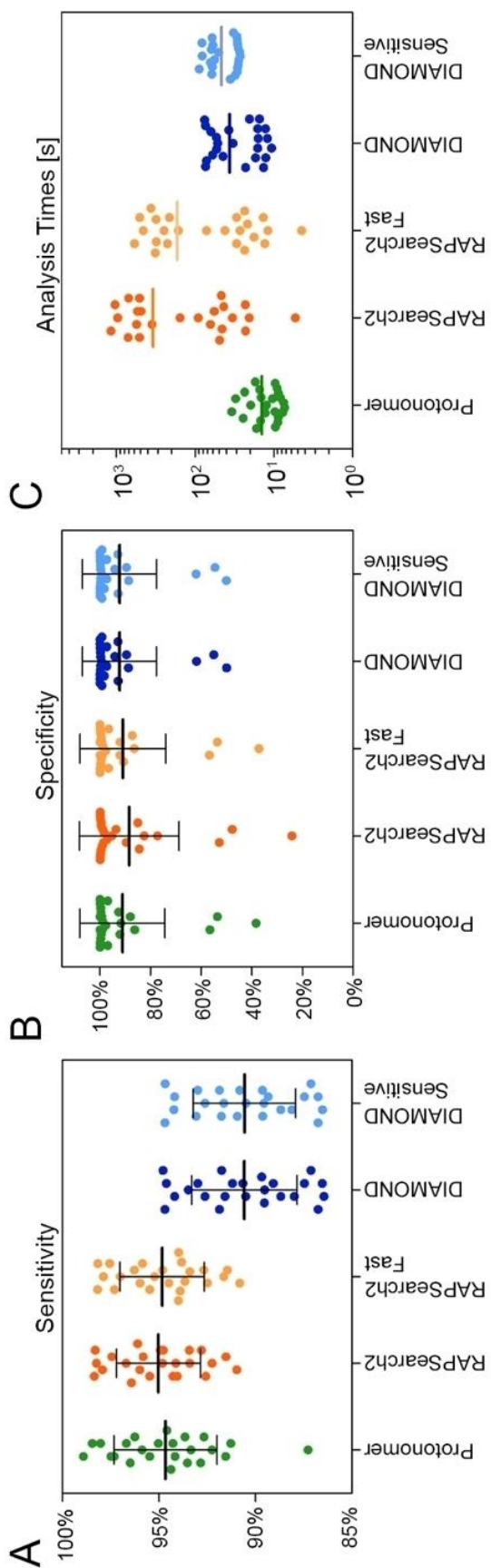


Figure 2.17. Performance of the Protonomer module for virus detection. (A) Sensitivity. Protonomer ($94.6 \pm 2.7\%$), and RAPSearch2 (default, $95.0 \pm 2.2\%$; fast, $94.8 \pm 2.2\%$) were more sensitive than DIAMOND (default, $90.5 \pm 2.7\%$; sensitive, $90.5 \pm 2.7\%$). (B) Specificity. Conversely, Protonomer ($90.7 \pm 17.1\%$) and DIAMOND (default: $92.0 \pm 17.1\%$, sensitive: $91.9 \pm 14.9\%$) provided higher specificity than RAPSearch2 in default mode ($88.0 \pm 20.0\%$). (C) Analysis times for Protonomer, RAPSearch2, and DIAMOND were tested on the same samples, which required analysis of between $\sim 2,000$ and $\sim 1,000,000$ ‘viral’ + ‘unknown’ sequences.

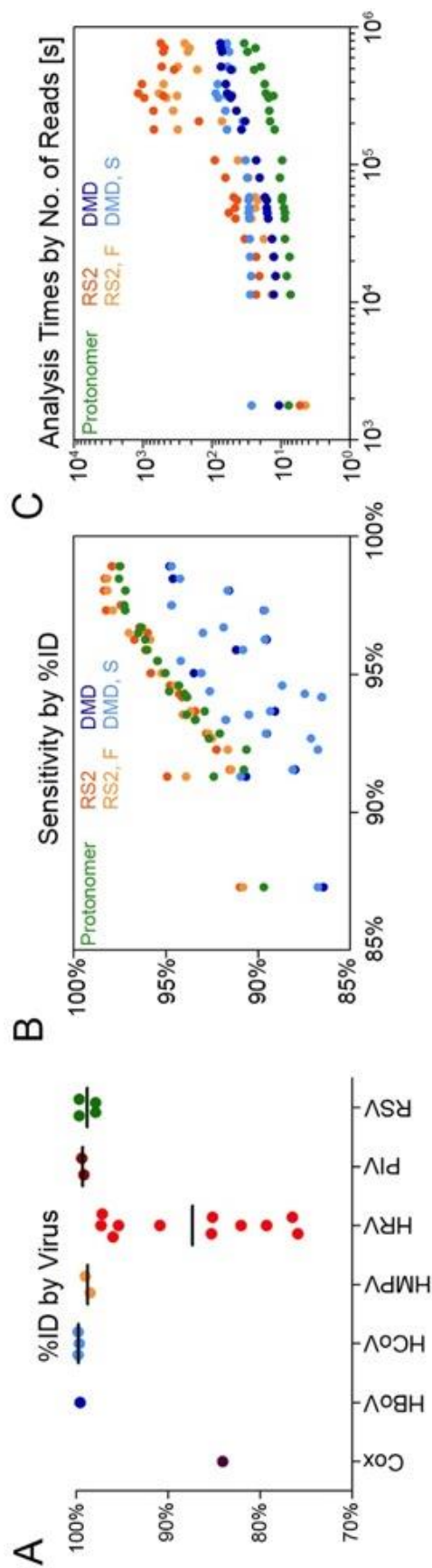


Figure 2.18. Sensitivity of tools correlated with phylogenetic distance of viral strains to reference sequences (A) Nucleotide sequence identity of viral consensus sequences from 23 respiratory samples used for Protonomer benchmarking to the best match in the NCBI nt database. Virus-positive samples were selected to represent a range of pairwise identities. (B) Sequencing reads were binned and the ‘viral’ and ‘unclassified’ bins were taxonomically classified by Protonomer, RAPSearch2 (default and fast settings), and DIAMOND (default and sensitive settings, see Figure 2.17). (C) Analysis times for Protonomer, RAPSearch2, and DIAMOND were tested on the same samples, which required analysis of between ~2,000 and ~1,000,000 ‘viral’ + ‘unknown’ sequences.

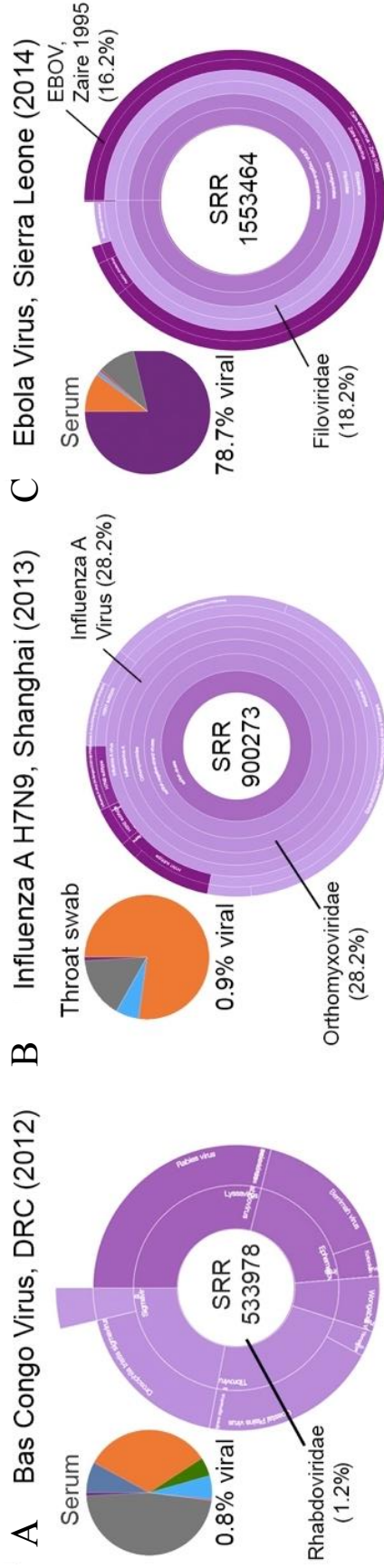


Figure 2.19. Detection of highly pathogenic viruses. (A) Novel Rhabdovirus in RNA-Seq data (SRR533978) from serum of a patient with hemorrhagic fever in the Democratic Republic of Congo (DRC), now known as Bas Congo Virus (40); approximately 13% of target reads from this highly divergent virus were classified at the family level (Rhabdoviridae) with genus-level assignments of Lyssavirus (1), Ephemerovirus (2), unassigned Rhabdoviridae (3), Tibrovirus (4), Sigmavirus (5); (B) avian influenza virus H7N9 in RNA-Seq data (SRR900273) from a throat swab of a patient in Shanghai with H7N9 infection (41); (C) Ebola virus, strain Zaire 1995, in RNA-Seq data (SRR1553464) from serum of a patient with suspected ebola virus disease in Sierra Leone (21).

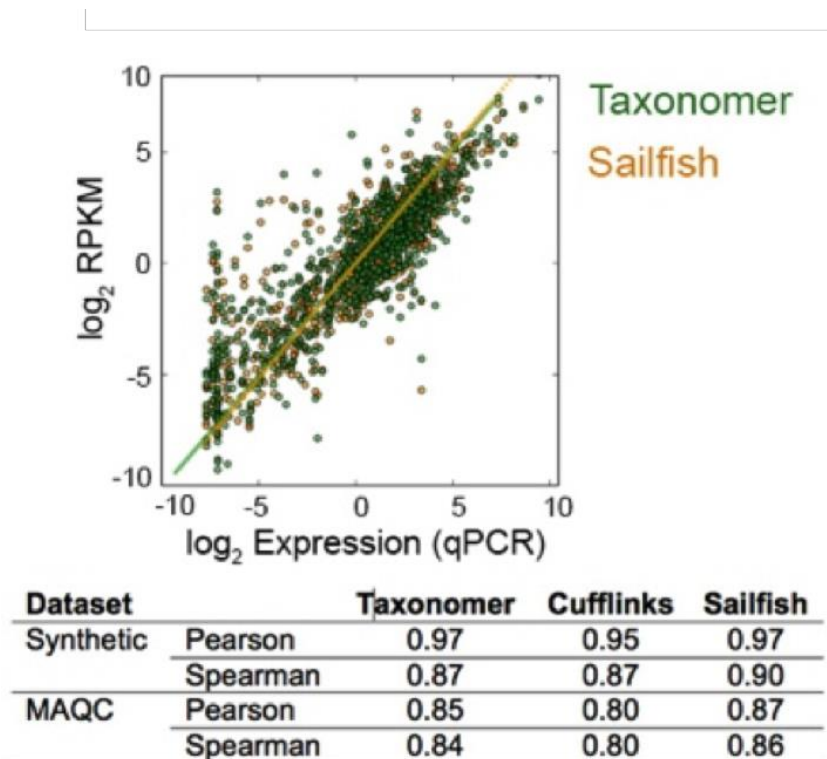


Figure 2.20. Published RNA-seq data from a commercially available RNA standard (see Table 2.11) analyzed by Taxonomer, Sailfish, and Cufflinks. Estimated transcript expression was compared to data obtained by quantitative PCR (qPCR).

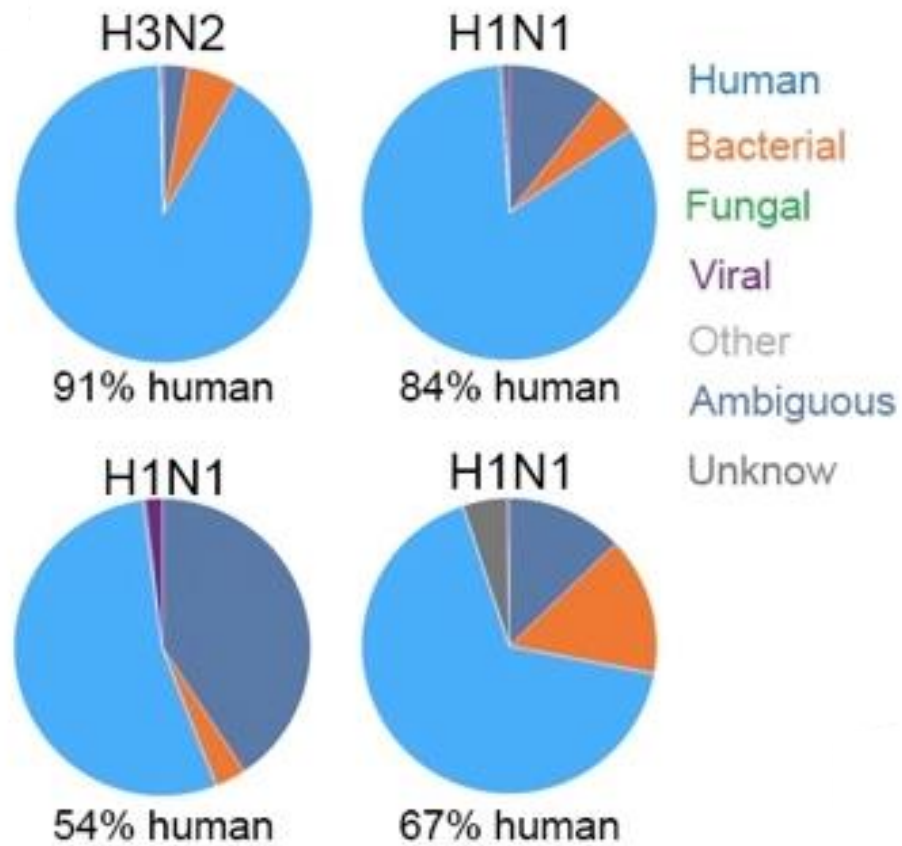


Figure 2.21 Application of Taxonomer to metagenomic RNA-seq data from routine respiratory samples from patients with influenza infection.

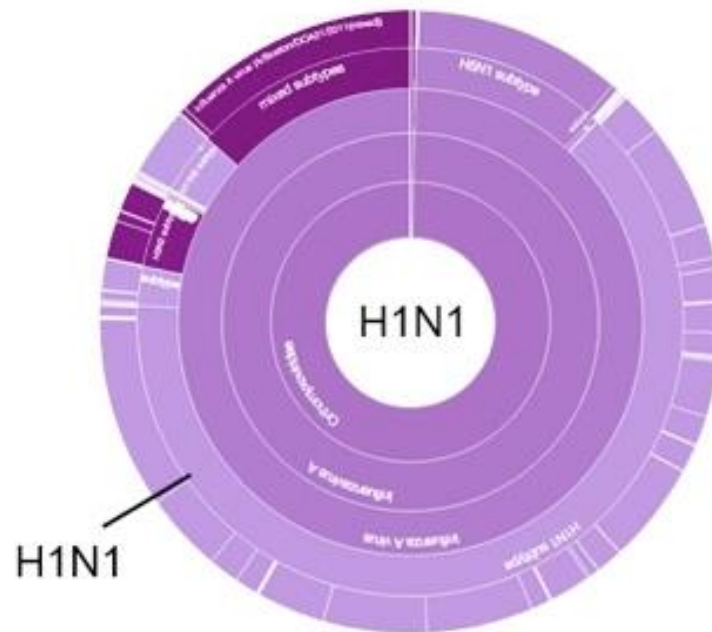


Figure 2.22. Classification of viral sequencing reads by Protonomer and typing of this strain as influenza A (H1N1).

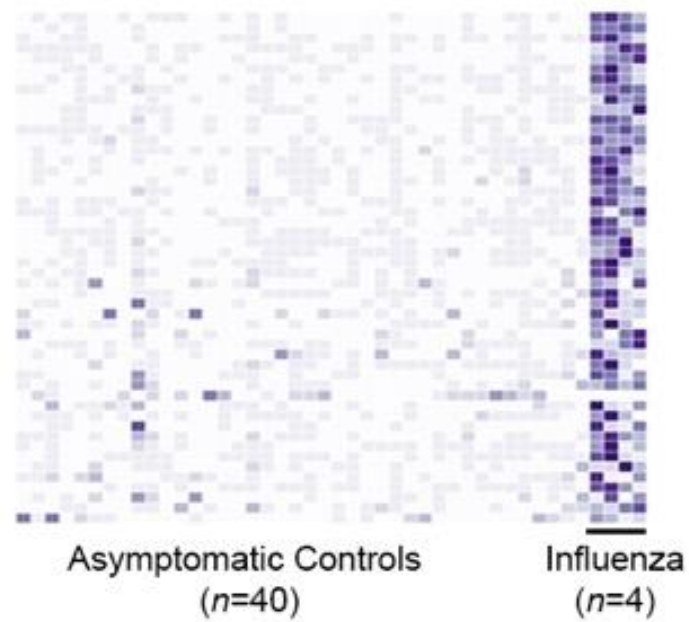


Figure 2.23. Differential gene-level mRNA expression profiles from 4 patients with influenza A virus compared to asymptomatic controls ($n=40$; top 50 differentially expressed genes are shown).

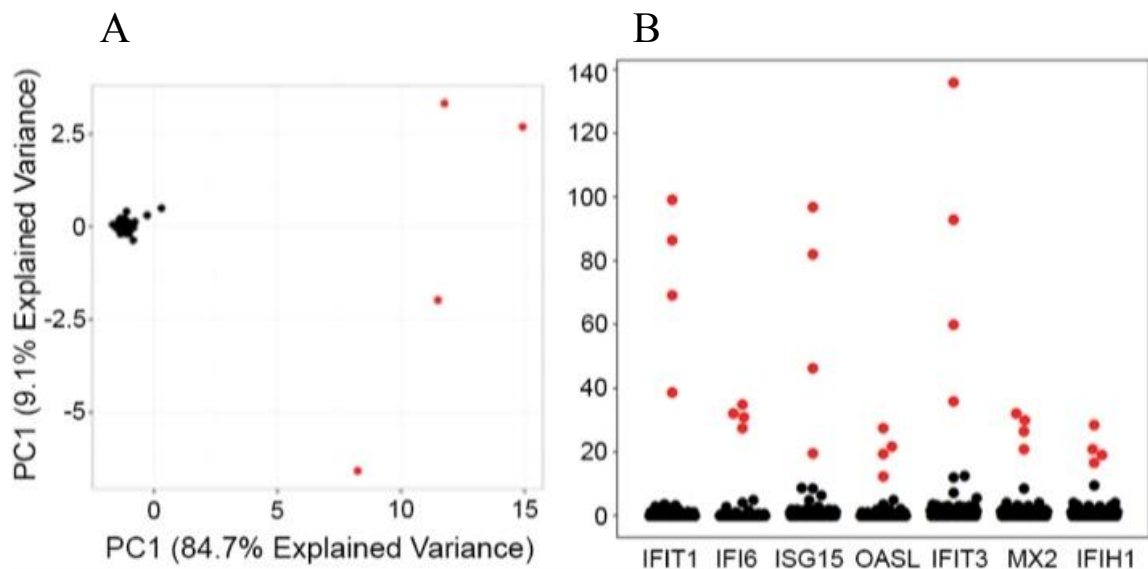
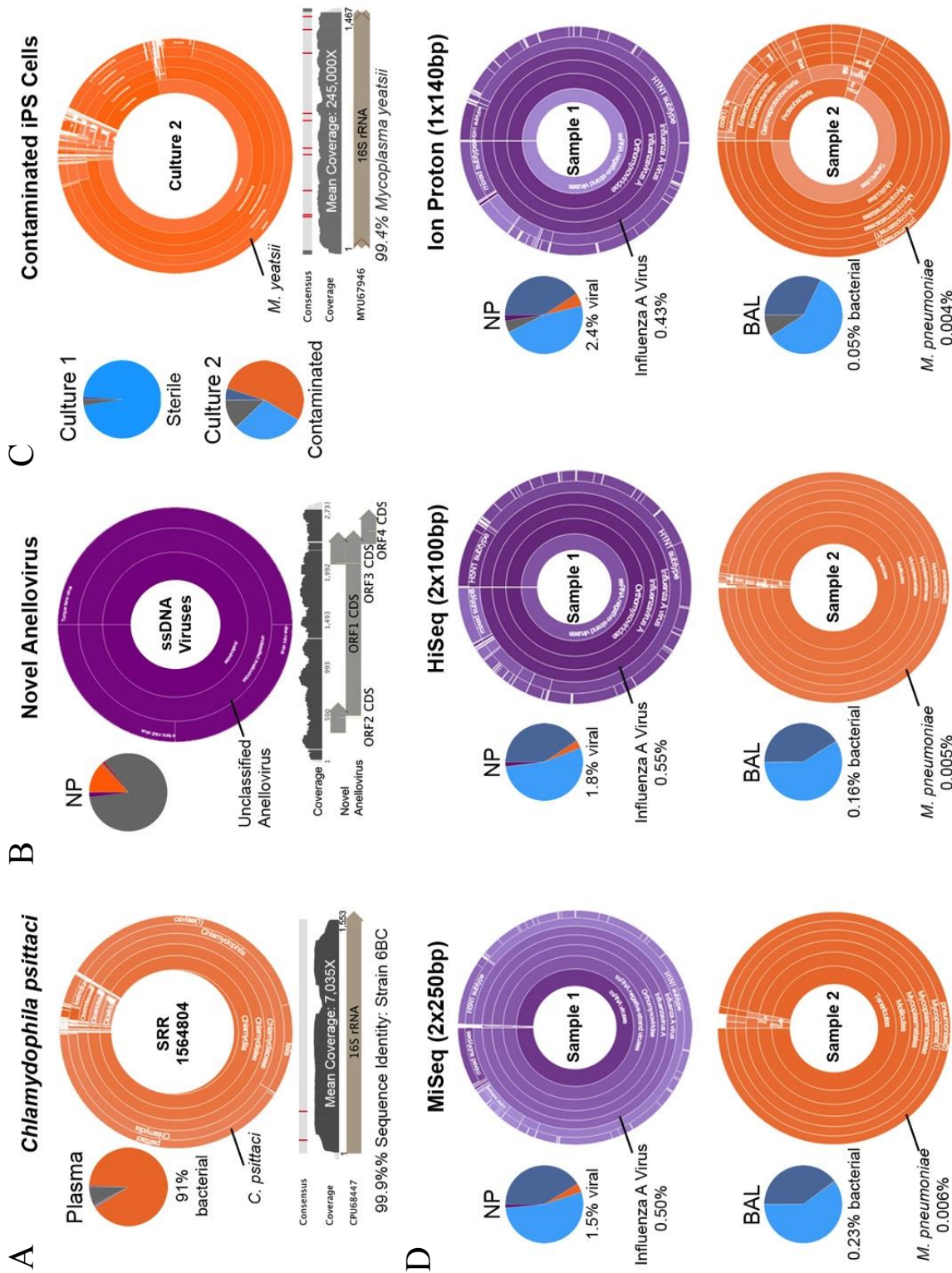


Figure 2.24. Sub-analysis of differential gene-level mRNA expression profiles from 4 patients with influenza A virus compared to asymptomatic controls infection (A) Expression profiles for the 17 most differentially expressed genes differentiate cases from controls (principal component analysis, PC1 and PC2 explaining 93.8% of the total variance). (B) Normalized expression levels for individual patients of seven of the top 17 genes.

Figure 2.25. Sample applications of Taxonomer. (A) Taxonomer detected a previously unrecognized *Chlamydomphila psittaci* infection (psittacosis), in plasma from a patient with suspected Ebola virus disease in Sierra Leone (SRR1564804). The 16S rRNA gene was covered a mean of 7,035-fold with the consensus 16S rRNA sequence from this isolate sharing 99.9% identity with the type strain (6BC, ATCC VR-125, CPU68447) enabling reliable identification. Positions of 2 single nucleotide polymorphisms are highlighted in red. (B) Taxonomer detected a novel Anellovirus in a nasopharyngeal swab. Forty-four reads were classified at the family level (Anelloviridae) or below. Mapping reads back to a manually-constructed viral consensus genome sequence showed 14-fold mean coverage, 68.5% pairwise nucleotide-level identity, and 44%-60% predicted protein identity with TTV-like mini virus isolate LIL-y1 (EF538880.1). (C) Identification of *Mycoplasma yeatsii* contamination in RNA-seq data from cultured iPS cell (right) compared to non-contaminated iPS cell culture (left) based on read binning (top). High expression of rRNA is demonstrated by 32% of RNA-Seq reads mapping to the *M. yeatsii* 16S rRNA gene (245,000X coverage, 99.4% sequence identity with type strain GIH (MYU67946). (D) Taxonomer is compatible with different sequencing protocols, recovering similar proportions of viral (influenza A, 0.43% to 0.55% of all reads) and bacterial (*Mycoplasma pneumoniae*, 16S rRNA sequences representing 0.004% to 0.006% of all reads) pathogen sequences when sequencing samples on 3 commonly-used sequencers with 2 different library preparation methods. Samples were known to be positive for influenza A (H1N1) pdm09 and *M. pneumoniae* based on diagnostic PCR tests.



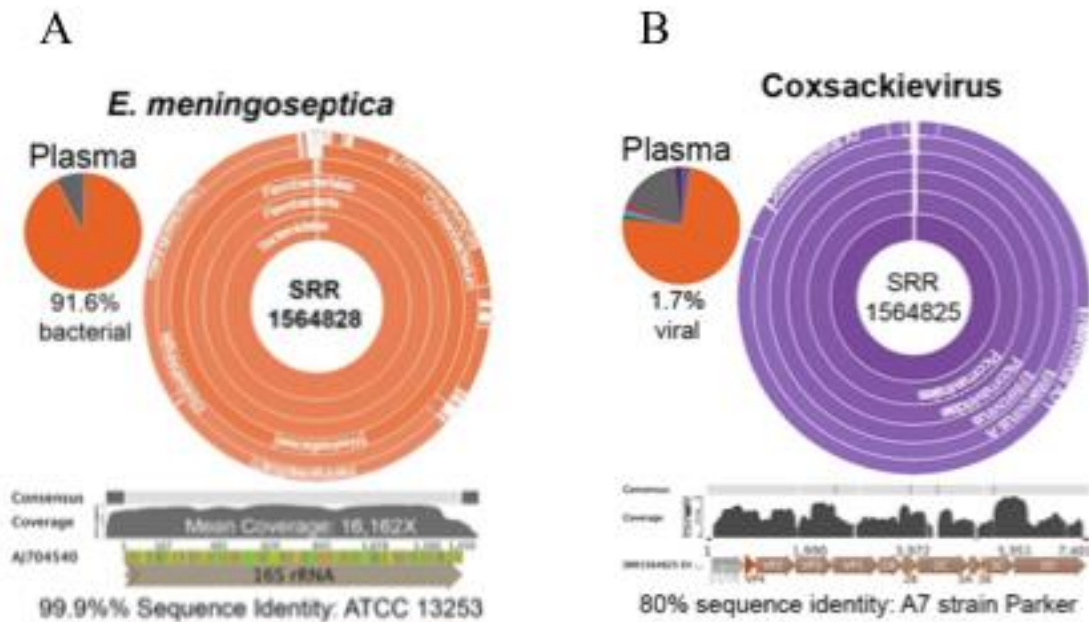


Figure 2.26. Taxonomer in clinical samples. (A) Taxonomer detected *Elizabethkingia meningoseptica* in sample SAMN03015718 (SRR1564828). Mean coverage of the 16S rRNA gene was 16,162-fold and the consensus sequence shared 99.9% nucleotide sequence identity with the type strain of *E. meningoseptica* (AJ704540, ATCC 13253). *E. meningoseptica* is a ubiquitous gram-negative bacterium that characteristically causes meningitis or sepsis in newborns but also immunocompromized adults. (B) Taxonomer classified a reported Enterovirus as Enterovirus A in plasma from a patient with suspected Ebola virus disease in Sierra Leone (SRR1564825). Mean sequencing depth was 162X covering 96% of the reference sequence (AY421765). Analysis of a manually constructed viral consensus genome sequences identified the strain as sharing 80% nucleotide sequence identity with Coxsackievirus A7, strain Parker.

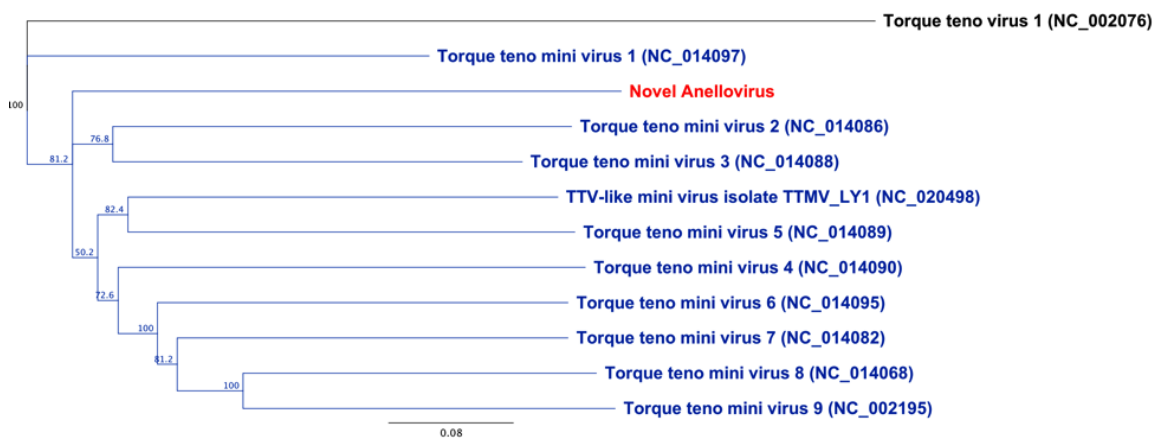


Figure 2.27. Phylogenetic tree of consensus sequence of novel Anellovirus (see Figure 2.25B) with reference sequences for Torque teno mini viruses. Torque teno virus 1 is shown as outgroup.

Table 2.1 Reference databases for the Binner module, source, version, number of reference sequences, and kmers.

Binner Database	Flag	Source	Version	Sequences	kmers	Source
Human genome	2	NCBI SILVA	GRCh38 ¹ 4/10/2014	455 +8 LSU + 158 SSU	2,257,262,659	ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/Assembled_chromosomes/seq/ http://www.arb-silva.de/documentation/release-119/
Human transcripts	32	NCBI	4/10/2014	98,746	79,809,821	ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/RNA/
Human mitochondria	512	Mitomap	4/10/2014	26,829	248,056	http://mitomap.org/bin/view.pl/MITOMAP AP/Mitobank
Bacterial genomes	1	NCBI	4/10/2014	5,189	5,433,934,380	ftp.ncbi.nlm.nih.gov/genomes/
Bacterial SSU	128	SILVA	119_ref	1,473,625	45,354,161	http://www.arb-silva.de/documentation/release-119/
Bacterial LSU	256	SILVA	119_ref	41,505	4,817,477	http://www.arb-silva.de/documentation/release-119/
Plastids LSU/SSU	16384	SILVA	119_ref	2,946 LSU 10,458 SSU	2,148,466	http://www.arb-silva.de/documentation/release-119/
Fungal genomes	4	NCBI	4/10/2014	1,146	583,739,768	ftp.ncbi.nlm.nih.gov/genomes/
Fungal LSU/SSU	64	SILVA	119_ref	2,317 LSU 21,106 SSU	2,306,754	http://www.arb-silva.de/documentation/release-119/
Fungal ITS	1024	UNITE	07/04/2014	409,493	19,398,519	https://unite.ut.ee/index.php
Viruses (NCBI)	8	NCBI	8/7/2014	1,668,565	279,444,799	entrez search with query txid10239[Organism] and not txid131567[Organism] and not gbdiv_pat[PROP]

Table 2.1 continued.

Binner Database	Flag	Source	Version	Sequences	kmers	Source
Phage	16	NCBI	4/10/2014	14,440	69,802,836	Phage taxID's per viralzone: http://viralzone.expasy.org/all_by_protein/256.html and http://viralzone.expasy.org/all_by_species/663.html
Phi X174	4096	NCBI		1	5,366	NC_001422
Other Eukaryotes LSU/SSU	2048	SILVA	119_ref	86,633	16,062,966	http://www.arb-silva.de/documentation/release-119/
Repeat masker	8192	Repeatmasker		383	16,419	http://www.repeatmasker.org/
Adapter sequences						
ERCC controls						

SSU – small subunit rRNA

LSU – large subunit rRNA

ITS – internal transcribed spacer

Table 2.2 Bin assignment for reads with equal numbers of kmer matches to multiple Binner databases and kmer matches below threshold. Some reference sequence databases are subsets or overlap with others (e.g. ‘Human transcripts’ and ‘Human genome’) and some sequences may be assigned varying taxID’s (e.g. phage sequences may be annotated as viruses or as bacteria, if integrated as prophages). As a result, query sequences may share an equal number of kmers with more than one reference database. The ‘Binner’ module assigns these query sequences as outlined below.

Equal kmer count of...	And...	Assignment
‘Human transcripts’	‘Human genome’ and/or ‘Mitochondrial genomes’	‘Human transcripts’
‘Bacterial 16S’	‘Bacterial LSU’ and/or ‘Bacterial genomes’ and/or ‘Plastids LSU/SSU’	‘Bacterial 16S’
‘Fungal ITS’	‘Fungal genomes’ and/or ‘Fungal LSU/SSU’	‘Fungal ITS’
‘Phage’	‘Viruses (NCBI)’ and/or ‘Bacterial genomes’	‘Phage’
All other ties		‘Ambiguous’
Kmer count < threshold		‘Unknown’

Table 2.3 Contents of visualized pie charts in the web portal. Sub-bin assignments are summarized for interactive visualization at Taxonomer.iobio.io as indicated.

Bin	Sub-bins
Human	'Human genome', 'Human transcripts', 'Mitochondrial genomes'
Bacterial	'Bacterial genomes', 'Bacterial SSU', 'Bacterial LSU', 'Plastids LSU/SSU'
Fungal	'Fungal genomes', 'Fungal LSU/SSU', 'Fungal ITS'
Viral	'Viruses (NCBI)', 'Phage'
Other	'Other Eukaryotes LSU/SSU'
Ambiguous	Any database combination not specified above

Table 2.4 Optimal kmer cutoffs for bin assignments based on the Youden's Index and F1 Score. Optimal kmer cutoffs determined by receiver operator characteristics analysis using the Youden's Index and F1 Score (23) are shown. The default cutoff used by the 'Binner' module is 11.

Bin	Youden's Index	F1 Score
Human	13	13
Bacteria	5	8
Fungal	3	4
Virus	3	4
Parasite*	22	21

*Parasites are not present in the binner databases, reads from parasites are considered true positives if they remain unbinned

Table 2.5 Viruses, percent nucleotide-level identity to reference sequences in the NCBI nt database, as well as numbers of total and viral reads for pediatric upper respiratory tract specimens used to compare ‘Protonomer’, RAPSearch2, and DIAMOND for protein-level classification of viral sequences.

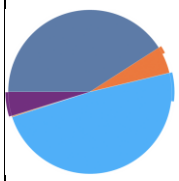
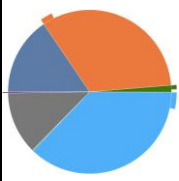
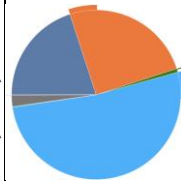
Virus	Nucleotide ID	GenBank Accession	Total Reads	Target Reads (<i>n</i>)	Target Reads (%)
HCoV (HKU1)	99.8%	KF686344	317,354	305,544	96.3%
HCoV (NL43)	99.8%	JQ765567	44,825	20,800	46.4%
HCoV (OC43)	99.7%	AY903460	15,515	6,919	44.6%
Coxsackie Virus B4	84.1%	KF878966	21,399	1,027	4.8%
HBoV	99.6%	JQ923422	206,869	1,119	0.5%
HMPV	98.5%	GQ153651	80,362	7,059	8.8%
HMPV	99.0%	EF535506	55,240	2,683	4.9%
HRV-A	90.9%	EF173415	11,369	2,413	21.2%
HRV-C	85.2%	DQ875932.2	490,829	491	0.10%
HRV-C	85.3%	DQ875932.2	704,819	394	0.06%
HRV-C	79.3%	JF436925.1	662,784	200	0.03%
HRV-C	97.3%	JX074056	385,808	208,446	54.0%
HRV-C	82.1%	JF317017	306,436	232,451	75.9%
HRV-C	97.2%	JX074056	246,973	35,474	14.4%
HRV-C	75.9%	KF958311	28,862	2,657	9.2%
HRV-C	96.0%	JN990702	330,157	252,416	76.5%
HRV-C	76.5%	GQ223228	179,888	153,429	85.3%
HRV-C	95.4%	GQ323774	58,005	1,369	2.4%
PIV-1	99.2%	JQ901989	107,818	9,392	8.7%
PIV-3	99.4%	KF530232	48,547	15,651	32.2%
RSV-A	99.7%	KF826849.1	762,085	2,218	0.29%
RSV-B	97.9%	JQ582843	1,784	1,035	58.0%
RSV-B	97.9%	JQ582843	40,707	32,047	78.7%
RSV-B	99.7%	JN032120.1	516,693	495,469	95.9%

HCoV – human coronavirus
HBoV – human bocavirus
HMPV – human metapneumovirus
HRV – rhinovirus
PIV – parainfluenza virus
RSV – respiratory syncytial virus.

Table 2.6 Flux Simulator parameters used to generate simulated RNAseq reads for benchmarking transcript assignment. Following the benchmarks used for Sailfish we filtered the transcript GTF using the gffread utility with the flags -C -M -E and -T, as well as any transcripts consisting solely of Ns. The GTF was sorted using the FluxSimulator sortGTF command and used to generate the synthetic data for benchmarking.

Stage	Parameters
Expression	NB_MOLECULES 5000000 REF_FILE_NAME Homo_sapiens_ENSEMBL_37.75.gtf TSS_MEAN 50 POLYA_SCALE NaN POLYA_SHAPE NaN
Fragmentation	FRAG_SUBSTRATE RNA FRAG_METHOD UR FRAG_UR_ETA NaN FRAG_UR_D0 1
Reverse Transcription	RTRANSCRIPTION YES RT_PRIMER RH RT_LOSSLESS YES RT_MIN 500 RT_MAX 5500
Filtering & Amplification	FILTERING YES GC_MEAN NaN PCR_PROBABILITY 0.05
Sequencing	READ_NUMBER 150000000 READ_LENGTH 76 PAIRED_END YES ERR_FILE 76 FASTA YES UNIQUE_IDS NO

Table 2.7 Processing time of Taxonomer compared to rapid classification pipelines SURPI and Kraken. Five RNA-Seq samples generated from nasal specimens with varying degrees of taxonomic composition illustrate the effect on pipeline speeds.

Sample Composition, Total Reads	Pathogen	Application	Subtraction	Binning	Classification	Protein Search	Total Time	% Reads Classified
 <p>HCov 6,599,164</p>	Taxonomer	-	5m	22s	10s	5.5m	99.9%	
	Kraken	-	-	1.5m	-	1.5m	99.6%	
	SURPI	3.3m	-	74m	15m	92m	99.9%	
 <p>Influenza A virus 7,542,552</p>	Taxonomer	-	8m	40s	30s	9.2m	88%	
	Kraken	-	-	1.5m	-	1.5m	66%	
	SURPI	9.8m	-	208m	18m	236m	78%	
 <p>HMPV 6,252,311</p>	Taxonomer	-	5.2m	56s	10s	6.3m	98%	
	Kraken	-	-	1.3m	-	1.3m	93%	
	SURPI	56m	-	648m	24m	728m	95%	

Human-blue; Bacteria-orange; Fungal-green; Virus-red; other-yellow; unclassified-grey.

Table 2.8 Broad taxonomic classification of read 1 versus read 2 by SURPI differs for 2-9% of mate pairs. Broad taxonomic classification by SURPI¹⁴ (see Figure 2.9) was determined for read 1 and read 2 of paired synthetic reads (SILVA 119, see methods) and RNA-Seq data (samples from Table 2.7, limited to pairs passing quality filters, see methods). Broad taxonomic assignments were compared for concordance. Discordance ranged between 2-3% for synthetic 16S read pairs and from 3-9% for RNA-Seq data. Discordance was greatest for samples with higher abundance of bacterial reads (samples 2 and 3, see Table 2.7), presumably due to database incompleteness, inconsistent annotations, and because SURPI's assignment is based on the single reference sequence with the highest score.

Sample	Read Length	Read pairs with discordant		Total	%
		assignment, R1 vs. R2	(<i>n</i>)		
Synthetic 16S	2x100bp	6,984	300,128	2.3	
Synthetic 16S	2x250bp	2,888	119,009	2.4	
Sample 1 (Table 2.7)	2x100bp	172,759	5,916,921	2.9	
Sample 2 (Table 2.7)	2x100bp	586,486	6,261,301	9.4	
Sample 3 (Table 2.7)	2x100bp	326,263	5,536,276	5.9	

Table 2.9 Accessions for published 16S amplicon data used in for bacterial abundance estimates, numbers of reads, and analysis times for the RDP Classifier and Taxonomer. Number of reads for reference is based on mate pairs.

Sample	Source	Ref.	Reads	RDP Classifier [min]	Taxonomer [min]
ERR498444	Human gut	(38)	20,469	285	0.91
ERR498459	Human gut	(38)	8,413	303	0.45
ERR498467	Human gut	(38)	16,864	426	0.62
ERR498476	Human gut	(38)	19,066	402	0.96
ERR498532	Human gut	(38)	20,458	315	1.02
ERR498541	Human gut	(38)	18,803	354	0.78
ERR498566	Human gut	(38)	12,612	200	0.60
ERR498576	Human gut	(38)	10,070	258	0.49
ERR498611	Human gut	(38)	19,506	342	0.96
ERR498653	Human gut	(38)	14,311	225	0.61
ERR502969	Dog nose	(37)	62,836	492	1.26
ERR502989	Human nose	(37)	79,093	930	2.37
ERR503004	Kitchen floor	(37)	74,061	594	2.00
ERR503007	Human hand	(37)	67,144	615	1.41
ERR503052	Human nose	(37)	54,569	468	0.87
ERR503054	Human hand	(37)	77,382	822	2.67
ERR503166	Bedroom floor	(37)	57,718	498	1.10
ERR503209	Kitchen floor	(37)	64,996	534	1.91
ERR503211	Bathroom door knob	(37)	70,964	630	1.44
ERR503212	Human nose	(37)	124,363	852	3.27

Table 2.10 Taxonomer bacterial abundance estimates compared to those of the RDP Classifier using recently published 16S amplicon sequencing metagenomics studies.

Sample	Approach	Platform, Reads	Reads/ Sample	Time/Sample [s]			Rank	ρ
				Taxonomer	Reads	RDP		
Stool (38)	16S (V4) amplicon	HiSeq (1x150bp)	1.6x10 ⁴	0.7 min	311 min	Order	0.960	
Environment, human, pets (37)	16S (V4) amplicon	MiSeq (2x150bp)	7.3x10 ⁴	1.8 min	644 min	Order	0.997	
Respiratory	RNA-Seq	HiSeq (2x100 bp)	1.6x10 ⁶	27.4 min	7,245 min	Order	0.992	
						Genus	0.955	

Spearman correlation coefficients (ρ) of abundance estimates are shown for Taxonomer and the RDP Classifier at the order and genus-levels using the Greengenes 99% OTU reference database. * - 2.5%; ** - 1.9%; *** - 2.5%

Table 2.11 Accession numbers for human brain RNAseq data used to compare with MAQC qPCR data.

Sample	Source	Reads
SRR037452	Human brain	11,712,885
SRR037453	Human brain	11,413,794
SRR037454	Human brain	11,816,021
SRR037455	Human brain	11,244,980
SRR037456	Human brain	12,081,324
SRR037457	Human brain	11,365,146
SRR037458	Human brain	11,616,331

Table 2.12 Genes (n=17) that are differentially regulated in nasopharyngeal and oropharyngeal swabs from children with pneumonia who tested positive for influenza virus (n=4) compared to asymptomatic controls (n=40). Read counts and *P*-values (raw and adjusted) are shown. A – controls; B – influenza (see Figure 2.23).

Gene ID	Base Mean A	Base Mean B	Fold Change	<i>p</i>	<i>p</i> (adj)
IFIT1	0.7	73.4	104.5	7.1E-19	1.5E-14
IFI6	0.5	31.4	64.8	6.3E-13	6.7E-09
IFIT2	2.1	135.5	63.8	7.8E-09	5.5E-05
ISG15	1.4	61.2	43.3	1.4E-08	6.4E-05
OASL	0.6	20.3	33.3	1.5E-08	6.4E-05
IFIT3	2.1	81.2	38.7	5.4E-08	1.9E-04
NT5C3A	0.7	20.1	30.7	3.3E-07	9.9E-04
MX2	1.4	27.4	19.2	4.0E-07	1.1E-03
IFITM1	2.4	32.8	14.0	6.4E-07	1.5E-03
CXCL10	0.6	37.3	64.6	9.0E-07	1.9E-03
IFI44L	1.5	26.6	17.8	1.6E-06	3.1E-03
MX1	4.2	56.5	13.5	1.8E-06	3.2E-03
IFIH1	1.4	21.3	15.0	9.7E-06	1.6E-02
OAS2	2.8	37.5	13.2	1.3E-05	1.9E-02
SAMD9	2.8	61.9	22.5	2.6E-05	3.7E-02
RSAD2	1.4	47.0	33.7	2.9E-05	3.8E-02
DDX58	1.1	16.6	15.3	3.9E-05	4.8E-02

Table 2.13 Gene ontology assignments for enrichment of biological processes (g) and molecular functions (h) are shown.

Term	Go Accession	Background frequency	Sample frequency	P-value	Up/Down	Biological processes (B) or Molecular functions (M) Term
Defense response to virus	0051607	148	22	3.3E-31	Up	B
Response to virus	0009615	252	23	4.7E-28	Up	B
Defense response to other organism	0098542	328	23	1.9E-25	Up	B
Response to other organism	0051707	645	26	1.9E-23	Up	B
Response to external biotic stimulus	0043207	645	26	1.9E-23	Up	B
Immune effector process	0002252	413	23	3.4E-23	Up	B
Response to biotic stimulus	0009607	674	26	5.7E-23	Up	B
Type I interferon signaling pathway	0060337	68	15	1.9E-22	Up	B
Cellular response to type I interferon	0071357	68	15	1.9E-22	Up	B
Response to type I interferon	0034340	69	15	2.3E-22	Up	B
2'-5'-oligoadenylate synthetase activity	0001730	4	4	5.15E-07	Up	M
Double-stranded RNA binding	0003725	53	6	5.45E-06	Up	M
Adenylyltransferase activity	0070566	23	4	5.46E-04	Up	M

Table 2.14 Taxonomer is compatible with different sequencing protocols. Two samples known to be positive for influenza A virus H1N1 and (nasopharyngeal swab) and *Mycoplasma pneumoniae* (bronchoalveolar lavage) based on diagnostic PCR test were analyzed by 3 commonly-used next-generation sequencers (illumina MiSeq, illumine HiSeq, Life Technologies Ion Proton).

Bin	MiSeq (2 x 250 bp)		HiSeq (2 x 100 bp)		Ion Proton (1 x ~140 bp)	
	Reads (n)	%	Reads (n)	%	Reads (n)	%
Influenza A						
Human	2,383,619	78.3	23,365,714	80.8	16,004,966	86.0
Fungal	2,824	0.1	57,141	0.2	88,237	0.5
ITS	101	0.0	3,065	0.0	1,088	0.0
Bacteria	105,307	3.5	922,407	3.2	783,684	4.2
16S	21,776	0.7	215,163	0.7	501,958	2.7
Phage	1,328	0.0	2,943	0.0	3	0.0
Viral	44,670	1.5	538,290	1.9	467,567	2.5
Other	476,256	15.7	3,786,107	13.1	719,070	3.9
Unknown	3,958	0.1	26,939	0.1	36,934	0.2
Mycoplasma						
Human	3,015,081	88.2	26,438,464	89.9	15,464,350	94.4
Fungal	3,290	0.1	67,372	0.2	42,548	0.3
ITS	48	0.0	3,636	0.0	459	0.0
Bacteria	1,749	0.1	9,454	0.0	7,917	0.1
16S	309	0.0	2,951	0.0	2,650	0.0
Phage	1,270	0.0	3,158	0.0	0	0.0
Viral	64	0.0	509	0.0	295	0.0
Other	391,584	11.5	2,834,912	5.0	825,431	5.0

CHAPTER 3

UNBIASED DETECTION OF RESPIRATORY VIRUSES BY NEXT-GENERATION SEQUENCING AND TAXONOMER, A RAPID, INTERACTIVE, WEB-BASED DATA ANALYSIS TOOL²

3.1 Introduction

Laboratory diagnosis of infectious diseases has historically taken a syndrome-based approach. Culture of appropriate specimens on a combination of relevant media or cell lines enables detection of many common bacterial, viral, and fungal pathogens. However, culture requires experienced personnel, several days to weeks to yield a definitive answer, depends on viability and appropriate culture conditions, and has limited sensitivity. Molecular tests have superior turnaround times, sensitivity, and taxonomic resolution. However, only targeted pathogens can be detected and differentiation of clinically or epidemiologically relevant strains or genotypes is limited. In addition, designs of molecular tests needs to be updated when new species or strains are recognized and to ensure that newly identified genetic variants can be detected.

² A version of this chapter has been accepted for publication in the Journal of Clinical Microbiology. Co-authors include Dr. Mark Yandell, Dr. Karen Eilbeck, Dr. Robert Schlaberg, and Keith Simmon.

In contrast, next-generation sequencing-based metagenomic testing combines many advantages of molecular tests and culture-based methods. Host and pathogen-derived nucleic acids are sequenced without a priori knowledge of expected pathogens allowing simultaneous detection of a virtually unlimited number of microorganisms, provided they possess sufficient sequence homology with reference sequences to enable classification. Unbiased, sequence-based pathogen detection will provide the greatest benefit when many diverse pathogens may cause overlapping symptoms, when an etiologic diagnosis influences treatment, and when molecular markers for drug resistance are known. One such application is the detection of respiratory pathogens. Even with state-of-the-art, multiplex molecular tests, identifying the etiology of respiratory tract infections is often unsuccessful; e.g. respiratory pathogens can only be detected in ~40-80% of patients with community-acquired pneumonia (CAP) with current tests (1-5). In addition, respiratory viruses of unclear pathogenicity (e.g. rhinovirus) are often found as the sole pathogen in respiratory samples, leaving doubt about the true etiology (6-9). With unbiased pathogen detection, alternative causes can be excluded with much greater confidence. Lastly, metagenomics-based testing provides sequence information on detected strains, often enabling genotyping, assessment of molecular markers for drug resistance, or molecular epidemiologic studies.

While several recent studies have demonstrated the power of next-generation sequencing-based metagenomics for pathogen detection (10-18), its performance compared to commercially available molecular tests is incompletely understood. Equally important, it remains to be demonstrated whether these approaches can be implemented in diagnostic laboratories and employed within a clinically meaningful timeframe using

computational resources and data analysis expertise available in diagnostic laboratories. Complexities of laboratory workflow, speed of sequence analysis, and expertise required for result analysis and interpretation are chief concerns.

We evaluate the analytical performance of metagenomics for detection of respiratory viruses using kit-based RNA-seq analysis of total RNA extracted from pediatric nasopharyngeal (NP) swabs. Resulting sequencing data were analyzed with a rapid, interactive, web-based data analysis tool, Taxonomer, eliminating the need for expensive computational hardware and bioinformatics expertise (19). We compared results to those of an FDA-cleared, multiplex PCR panel, the GenMark eSensor RVP (RVP), and demonstrated the utility of viral sequence information.

3.2 Materials and Methods

3.2.1 Samples

Nasopharyngeal (NP) swabs from children less than 5 years of age tested by the GenMark eSensor Respiratory Virus Panel (GenMark Dx, Carlsbad, CA) between April 2013 and March 2014 were de-identified using standard institutional procedures (University of Utah IRB number 56504) and stored at -80°C . Specimens positive for RNA viruses tested by the GenMark assay were retrospectively collected with preference given to dual infections (human metapneumovirus $n=5$, human rhinovirus $n=10$, Influenza A $n=5$, Influenza B $n=5$, parainfluenza 1 $n=5$, parainfluenza 2 $n=1$, parainfluenza 3 $n=4$, respiratory syncytial virus $n=8$). In addition, 67 samples were selected at random for inclusion in a direct side-by-side comparison.

3.2.2 GenMark eSensor Respiratory Virus Panel

Nucleic acid was extracted from 200uL of each sample, plus 10uL of internal control, on the NucliSENS easyMAG (BioMerieux, Durham, NC) and eluted into 60uL, 5uL of which was reverse transcribed and amplified with the eSensor Respiratory Virus Panel reagents following the manufacturer's instructions (GenMark). The following 14 viral targets are reported the eSensor XT-8™ system (GenMark): adenovirus B/E, adenovirus C, influenza A, influenza A H1, influenza A H3, influenza A 2009 H1N1, influenza B, respiratory syncytial virus subtype A, respiratory syncytial virus subtype B, parainfluenza virus 1, parainfluenza virus 2, parainfluenza virus 3, human metapneumovirus, and human rhinovirus.

3.2.3 Library Preparation and RNA Sequencing

NP swabs were thawed, vortexed, and 160uL of the transport media was transferred for extraction with the QIAamp Viral RNA mini kit, following the manufacturer's instructions (Qiagen, Valencia, CA). Eluted RNA was vacuum dried and stored at -80°C overnight. RNA-seq libraries were prepared with the TruSeq RNA Sample Prep Kit, following the manufacturer's instructions (Illumina, San Diego, CA). Libraries were quantified with the Illumina Universal Library Quantification Kit (Kapa Biosystems, Inc., Wilmington, MA). Library quality was assessed with a High Sensitivity DNA Analysis Kit on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries from 24 samples were combined in equimolar ratios for a final concentration of 9.6nM and sequenced in batches of 24 samples per lane on a HiSeq 2500 instrument (Illumina, San Diego, CA).

3.2.4 Data Analysis

RNA-seq data were analyzed with Taxonomer, a kmer-based, rapid, interactive metagenomic sequence analysis tool accessed through a web interface on the iobio framework, (<http://taxonomer.iobio.io>) (19,20). Taxonomer classifies each read to the highest taxonomic rank possible given a comprehensive sequence database. Relevant human pathogens detected by Taxonomer were manually confirmed with using Geneious (Biomatters, Ltd., Auckland, New Zealand) by mapping reads against a curated list of full-length viral reference sequences downloaded from NCBI.

Sequence-based typing of viral strains was performed by manual alignment to a reference genome and BLAST analysis of the largest contig of the appropriate genomic segment (e.g. VP1/3 for rhinovirus) or whole viral genome, if possible. Strains were considered typed if the references with the highest sequence identity over the entire contig all belonged to the same genotype (e.g. RSV-B was the highest match with no RSV-A at the same percent identity).

3.2.5 Monoplex, Real-Time PCR for Respiratory Viruses

Total nucleic acid was extracted on the Chemagic MSM I platform (Perkin Elmer) using 200uL of the NP swab transport media. Nucleic acid was eluted into 80uL and 10uL of the elution was used for amplification on an ABI 7900 instrument (Life Technologies, Foster City, CA). Real-time PCR assays for human metapneumovirus, respiratory syncytial virus, influenza A and B virus, parainfluenza virus types 1-4, and enterovirus validated for diagnostic testing at ARUP Laboratories were used for comparison (see Table 3.1) (21,22). For human rhinovirus and coronavirus species

(HKU1, NL63, OC43), research tests were used (23). All assays employ the QuantiTect RT-PCR kit (Qiagen), which includes an internal control and UNG, as well as primer/probe sets from Epoch Biosciences (ELITech Group, Bothell, WA). The following amplification conditions were used in all assays: 1 cycle for 10 min at 20C (UNG step), 1 cycle for 30 min at 50C (RT), 1 cycle for 15 min at 95C (Taq activation), then 50 cycles of 15 sec at 95C, 30 sec at 56C, and 30 sec at 76C, then 1 final melt cycle of 15 sec at 95C, 15 sec at 45C, and 15 sec at 95C. RNA standards specific for each virus were used to generate standard curves. Monoplex, real-time PCR was performed for each of the viruses detected by the RVP on the respective samples.

3.2.6 Statistics

Linear and Spearman's rank correlations were performed with Prism version 5.04 software (GraphPad Software, Inc., La Jolla, CA) and *P*-values less than 0.05 were considered significant.

3.3 Results

3.3.1 Agreement with RVP-Positive Samples

Archived nasopharyngeal swabs positive for one or more viruses by the RVP were retrospectively selected to cover each RNA virus target on the RVP. Preference was given to samples with co-detection of >1 virus. Agreement between RNA viruses detected by the RVP and RNA-seq in 42 samples was 86% (see Figure 3.1). Six respiratory viruses detected by the RVP were not detected by RNA-seq (blue bars). Four of these (one each of rhinovirus, influenza B virus, parainfluenza type 2, and respiratory

syncytial virus) were also not detected by monoplex, real-time PCR (hashed blue bars). Considering these false-positive by RVP, adjusted positive agreement between the RVP and metagenomics was 95%. Both of the remaining RVP-positive/PCR-positive/metagenomics-negative samples were low-positive for rhinovirus with monoplex PCR threshold cycles of 33 and 35.

3.3.2 Side-by-Side Comparison

Between April 2013 and March 2014, all NP swabs from children less than 5 years of age submitted for the RVP were banked after testing. From these, 67 were selected at random for testing by RNA-seq. Of these, 36 (53.7%) were positive by RVP for one or more respiratory virus resulting in detection of 37 respiratory viruses (Adenovirus, n=2; HMPV, n=4; influenza A virus, n=3; PIV-1, n=1; HRV, n=20; RSV, n=7). Of these 37 respiratory virus detections, 34 91.9% were also detected by RNA-seq. The RVP detected 3 additional, two of which tested negative by monoplex, real-time PCR while the third one could not be tested due to limited sample volume. RNA-seq and Taxonomer analysis detected 12 additional respiratory viruses, 3 of which were targeted by the RVP. The overall positivity rate for RNA-seq was 63%. Seven of the 12 additional viruses (58.3%) were confirmed by monoplex, real-time PCR, 3 were PCR-negative, and 2 could not be tested due to due to limited sample volume. For the 3 viruses detected by RNA-seq but not monoplex, real-time PCR had low viral read counts (4-31 reads). However, these reads were located across unique regions of the viral genome and sequences differed by several nucleotides from those of other viral strains detected at higher read counts on the same runs, suggesting they were not misclassifications and

unlikely to be contaminants. Rhinovirus was the most frequently detected pathogen by either method (n=22), RSV was second (n=6), followed by coronavirus (not included on the RVP, n=5) and HMPV (n=5). Manual confirmation of Taxonomer results showed 100% qualitative agreement for detection of respiratory viruses.

Co-detection of ≥ 2 respiratory viruses was more common with RNA-seq (14%) than with the RVP (3%). The majority of co-detections involved rhinoviruses and human bocavirus. When possible, these co-detections were confirmed with lab developed real-time PCR. There were several samples that were positive by both the RVP and RNA-seq but negative by monoplex, real-time PCR, suggesting that RNA-seq has at least comparable sensitivity compared to the RVP and monoplex, real-time PCR. Bocavirus was a common partner in co-detections by RNA-seq (n=4), but is not targeted by the RVP.

3.3.3 Taxonomic Composition of Total RNA from NP Swabs and Abundance of Viral RNA

Unlike with PCR, the sensitivity of RNA-seq-based metagenomics for pathogen detection is heavily influenced by the nucleotide composition of the sample. To characterize the relative contribution of host and microbial RNA in routinely collected NP swabs, we used Taxonomer to determine the high-level taxonomic classification of all reads (median of 15 million reads per sample, IQR 19-8 million reads). Since total RNA was sequenced, the majority of RNA-seq reads were of human origin (median 56.1%, IQR 47.7-62.3%), a median of 33.3% were not unambiguously classifiable (mostly ribosomal or mitochondrial RNA, data not shown, IQR 16.3-42.9%), a median of 3.4% of

reads were bacterial (IQR 0.9-14.8%), and only a median of 0.01% of reads was of viral origin (IQR 0.002-0.07%, see Figure 3.2). However, the number of viral reads spanned >5 orders of magnitude, with as few as 2 and as many as 3.2×10^5 reads detected by Taxonomer in RVP and monoplex, real-time PCR-positive sample. This highlights the importance of sample preparation and sequencing methods that minimize the likelihood of sample-to-sample contamination or limit methods to separate contamination from pathogen detection.

3.3.4 Correlation of Viral Read Counts with Viral Loads

Determined by PCR

Semi-quantitative detection of respiratory viruses has been shown to correlate with disease severity, at least in some studies (24-26). Thus, we compared viral loads in NP swab samples determined by real-time PCR with normalized viral reads counts using the large number of viruses detected by metagenomics (n=68) using the normalization scheme described in (27). Briefly, numbers of viral reads were divided by the number of total reads and the size of the respective viral genome in kilobases and then multiplied by 1 million to generate an RPKM value. In addition, monoplex, real-time lab-developed PCR was used with standards of known concentration to determine viral copies per mL of viral transport media. Correlation of RPKM and viral copies per mL was highly significant with a *P*-value <0.0001 (see Figure 3.3). This suggests that normalized viral read counts can be used for semiquantitative measurement of the viral burden in clinical samples.

3.3.5 Reproducibility

Three samples positive by RNA-seq each with a distinct fractional abundance of viral reads were selected to evaluate within-run and between-run variability. Two of the viruses were also detected by the RVP; the third one was a sample positive by PCR for coronavirus. Each sample was processed from start to finish (extraction to analysis) a total of five (HRV and HMPV) or 14 (HCoV) times. Libraries were sequenced on the same (within run) and on different (between-run) HiSeq lanes (see Figure 3.1). The fraction of viral reads is graphed as a proportion of the total RNA reads for each repeat and the coefficient of variation was calculated from these values. Given the complexity of the workflow, RNA-seq and Taxonomer analysis demonstrated excellent reproducibility, with coefficients of variation of 65% (HMPV, lowest fractional abundance), 16% (HCoV), and 47% (HRV, greatest fractional abundance).

3.3.6 Sequence-Based Characterization of Viral Strains and Antiviral Drug Resistance Determination

As metagenomics provides sequence information in addition to mere determination of presence or absence of pathogens, we studied available viral sequences to demonstrate utility. Even though viral reads were a very small proportion of the total reads, sufficient sequence was obtained for 84% of positive specimens to enable high-resolution, sequence-based genotyping. Consistent with the RVP results, all of the influenza A virus-positive specimens were typed as 2009 H1N1 strains. By RNA-seq, we were able to examine the oseltamivir resistance mutation at amino acid position 275 (H275Y) of the neuraminidase gene in six out of eight positive specimens (1.6-200-fold

median coverage). None of the six isolates had the H275Y amino acid substitution. RSV-B was far more common than RSV-A (nine vs. three of twelve RSV-positive samples, respectively). These results were consistent with RVP-based typing. Most rhinoviruses belonged to rhinovirus species C (62%) with only 21% belonging to rhinovirus species A, and 3% to rhinovirus B. Fourteen percent of rhinoviruses were untypeable (see Figure 3.1). For 14 (52%) of the rhinovirus-positive samples, coverage of the viral genome was sufficient to generate full-length viral consensus sequences. Genetic diversity was greatest for strains that belonged to rhinovirus species C. Strains B and N that clustered closely together, were collected during the same month from patients from the same state. The most divergent sample from any full-length sequence in the NCBI nt database was sample I, which had only 75% sequence identity with the closest match, HRV-C3 (strain HRV-QPM, EF186077.2). This sample was missed by the RVP but tested positive by the monoplex, real-time PCR with a Ct of 20. The one enterovirus sequence was most similar to coxsackievirus B4 strain E2 (NCBI accession number AF311939, 84% overall nucleotide identity). The alpha coronavirus NL63 was detected in 3 samples and beta coronaviruses HKU1 and OC43 were detected in 2 samples, each. All human bocavirus detected (n=4) belonged to genotype 1.

3.3.7 Detection of RNA from DNA viruses

RNA-seq was able to detect only 2 of 6 adenovirus-positive samples (see Figure 3.2a). Only very few adenovirus reads were generated in the 2 RNA-seq-positive samples (see Figure 3.2b). However, human bocavirus RNA was detected at high read counts in four samples. Additionally, high levels of RNA reads from a number of non-respiratory

DNA viruses were detected by metagenomics including HSV-1, CMV, EBV, and anellovirus (data not shown). Optimized nucleic acid extraction methods or simultaneous preparation of cDNA and DNA libraries may enable more complete characterization of the DNA virome in clinical samples.

3.3.8 Reagent Contamination

Contamination from reagents employed during extraction, library preparation, and sequencing has been previously described (28). To assess the contamination generated by our approach, we extracted and sequenced 3 molecular grade water samples alongside clinical samples. The reads generated by these samples were largely bacterial. No respiratory viruses or known human pathogenic viruses were detected (data not shown).

3.4 Discussion

We showed that RNA-seq-based metagenomics combined with a rapid, user-friendly data analysis tool has accuracy and sensitivity that compared favorably with a commercial multiplex PCR test. The unbiased approach of RNA-seq allowed us to query a theoretically unlimited number of pathogens in parallel, resulting in detection of more human viruses and a higher positivity rate. These included well-known respiratory viruses with clinical relevance when detected in the upper respiratory tract as well as potential pathogens that may only be relevant in the appropriate (e.g. immunocompromized) host and when detected from the lower respiratory tract (e.g. HSV, CMV). Interestingly, even though we used RNA-seq and included a DNase treatment step, DNA viruses were detected in some but not all PCR-positive samples. It is

possible that detection of mRNA from DNA viruses may serve as a marker of active replication. This is of relevance as several DNA viral respiratory pathogens can become latent (e.g. HSV, CMV) or persist for extended periods (e.g. HBoV) so that detection of their genomic DNA may not be a sufficient indication for acute infections.

The sensitivity of RNA-seq-based metagenomics is a function of sample composition and sequencing depth. When sequenced to the same depth, samples with an abundance of non-pathogen RNA (e.g. highly cellular samples or samples with abundant normal flora) result in lower analytical sensitivity than sample in which the pathogen RNA is more abundant (e.g. less cellular samples, higher pathogen load, absence of normal flora). As ribosomal RNA (rRNA) represents a large proportion of host RNA, rRNA depletion strategies have been used to mitigate this effect. We decided not to use this approach as it may have off-target effects (e.g. depleting microbial rRNA or other sequences with sufficient homology), which limits the unbiased nature of metagenomics. In addition, rRNA depletion or target enrichment steps add complexity and cost to the workflow. Samples were sequenced to a depth of 5-10 x 10⁶ reads/sample to limit sequencing costs. This sequencing depth resulted in comparable positivity rates and agreement with targeted PCR of >90%. When clinically relevant, samples can be sequenced deeper resulting in proportionally increases of analytical sensitivity.

When approaching the limit of detection, small numbers of viral reads pose challenges to result interpretation as they can represent results that are true-positive, low-level detection or artifacts. False-positive detections can be due to contamination during library preparation (e.g. sample-to-sample contamination prior to indexing), may be a result of sequencing artifacts (e.g. run-to-run carry-over or de-multiplexing errors), or

may be caused by erroneous classification during data analysis (e.g. due to highly homologous or low-complexity regions) (29). Thus, the confidence of viral detection depends on the number of viral reads and evenness of coverage. Given the testing complexity, read counts may vary between analytical replicates. To determine the within-run and between-run variability, we tested multiple aliquots of 3 virus-positive samples from sample extraction through data analysis. Respiratory viral read counts across a wide range of fractional abundance were highly reproducible within and between runs (CV ranging from 16% to 63%). Taken together, our results indicate that metagenomics combined with rapid, interactive, and user-friendly data analysis has value in supplementing current, PCR-based tests and may replace pathogen-specific tests in the future.

Other than its broad scope, another distinct advantage of metagenomics-based pathogen detection is the ability to determine the molecular subtype of a particular virus and query it quickly for genotypic markers of drug resistance or pathogenicity. In our study, molecular typing was possible for 84% of all viral strains. Relevant information derived from typing included (1) almost $\frac{2}{3}$ of rhinoviruses belonged to the more pathogenic species C including one highly divergent strain missed by the RVP (30,31); (2) all influenza A viruses were 2009 H1N1 strains but none contained the H275Y mutation conferring oseltamivir resistance; (3) RSV-B was 3-times more prevalent than RSV-A, which may be relevant as strain-specific differences in pathogenicity have been suggested; (4) high-resolution typing of an enterovirus as Coxsackievirus B4; (5) typing of 7 coronaviruses as NL63, HKU1, and OC43; (6) and genotyping of 4 bocavirus strains as HBoV-1. As genotype-phenotype correlations become better understood, genotypic

strain characterization will gain importance. This will also facilitate epidemiologic investigations or studies of vaccine effectiveness. Particularly in the case of influenza, real-time sequence information will improve surveillance studies, enable early detection of antiviral drug resistance, and inform vaccination strategies.

Respiratory viral burden correlates with disease severity and may help differentiate asymptomatic shedding from active infection (24-26,32). Published studies correlating viral read counts with quantitative PCR had limited sample sizes (13-15). Thus, we tested whether normalized read counts could be used for quantification of the viral burden by comparison to viral loads determined by pathogen-specific, laboratory-developed, quantitative real-time PCRs. While viral reads always represented a small fraction of total reads (see Figure 3.3a), normalized counts correlated highly significantly with viral loads. RNA-seq could therefore also be used to measure viral burden.

While we demonstrated analytical performance comparable to an FDA-cleared multiplex PCR, there are several barriers for routine diagnostic deployment of metagenomics-based testing. These include lengthy turn-around times, costs, and complexity of data analysis. First, the library preparation method used in this study required ~14 hours. We performed sequencing on an illumina HiSeq 2500 instrument in high output run mode, which took an additional ~11 days. At the time of writing, partially automated solutions for RNA-seq library preparation within ~8 hours and sequencing within ≤ 1 day for comparable per-base costs have become available (11,33). These advances are starting to enable diagnostic laboratories to provide results in a clinically meaningful timeframe and with a workflow that can be implemented in diagnostic

laboratories. However, for wide adoption, rapid, automated, closed system library preparation methods and quicker sample-to-data times are needed.

Second, cost is a great concern regarding the use of next-generation sequencing in infectious disease diagnostics. For the present study, RNA-seq reagent costs per sample were within \$10-20 of reagent costs for RVP. This was in part due to multiplexing 24 samples per sequencing lane. At the time of writing, cheaper library preparation kits and sequencing platforms have further decreased costs, quickly eliminating the cost differential. Enrichment of viral sequences and depletion of uninformative host RNA can reduce sequencing costs by increasing coverage but introduces complexity and costs of library preparations (34,35). We analyzed RNA-seq data solely for the presence of respiratory viruses, ignoring bacterial respiratory pathogens, as most of those can be part of the normal upper respiratory tract flora and only NP swabs were tested. However, when used with lower respiratory tract samples, RNA-seq has the potential to also replace a large number of commonly performed culture and PCR-based tests, at which point costs for unbiased pathogen detection will be even more competitive.

Finally, data analysis needs to be rapid, user-friendly, and reliable enough so it can be implemented without large investments in highly trained personnel and computational infrastructure. We used our recently published metagenomics data analysis tool, Taxonomer, to screen for the presence of respiratory viruses (19). Taxonomer analyzed $\sim 1 \times 10^6$ reads/minute, requiring <10 minutes per sample. For diagnostic applications, data analysis solutions are needed that minimize the time users spend reviewing results. We confirmed all respiratory virus detections manually to ensure accuracy. However, this was only informative in samples with low viral read counts

given concern of false-positive results due to misclassification or sequencing artifacts. Several RVP and multiplex PCR-positive samples only produced <10 reads for that virus, making detections unreliable at this low end (see Figure 3.3b). Deeper sequencing or target enrichment depletion approaches could alleviate the problem but increase costs and/or workflow complexity. For highly variable viruses (e.g. Picornaviridae), suspicious reads can be mapped back to viral consensus sequences of source strains to identify reads that likely represent artifacts. For diagnostic adoption, interpretive criteria similar to those being established for genomics laboratories will need to be developed and incorporated in diagnostic data analysis tools to enable consistent and rapid analyses (36,37).

In summary, we showed that metagenomics-based detection of respiratory viruses holds promise as a diagnostics tool enabling unbiased pathogen detection, molecular typing, and genotypic assessment of drug resistance or pathogenicity. Barriers to adoption, including turnaround time, cost, and complex data analysis are rapidly being removed. Initial adoption may be for testing of immunocompromized or otherwise predisposed patients, when routine therapeutic approaches fail, during clusters of infections of unknown etiology, or when molecularly characterize of pathogens is sought. As highlighted by a diverse HRV-A strain missed by the RVP, the unbiased nature of metagenomics can also assist with detection of novel viruses or variant strains

3.5 References

1. Choi S-H, Hong S-B, Ko G-B, Lee Y, Park HJ, Park S-Y, et al. Viral infection in patients with severe pneumonia requiring intensive care unit admission. *Am J Respir Crit Care Med.* **2012**; 186(4):325–332.

2. Karhu J, Ala-Kokko TI, Vuorinen T, Ohtonen P, Syrjälä H. Lower respiratory tract virus findings in mechanically ventilated patients with severe community-acquired pneumonia. *Clin Infect Dis*. Oxford University Press; **2014**; 59(1):62–70.
3. Honkinen M, Lahti E, Österback R, Ruuskanen O, Waris M. Viruses and bacteria in sputum samples of children with community-acquired pneumonia. *Clin Microbiol Infect*. Blackwell Publishing Ltd; **2012**; 18(3):300–307.
4. Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM, et al. Community-acquired pneumonia requiring hospitalization among U.S. adults. *N Engl J Med*. **2015**; 373(5):415–427.
5. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. **2015**; 372(9):835–845.
6. Pavia AT. Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. *Clin Infect Dis*. Oxford University Press; **2011**; 52 Suppl 4(Supplement 4):S284–9.
7. Ruuskanen O, Lahti E, Jennings LC, Murdoch DR. Viral pneumonia. *The Lancet*. **2011**; 377(9773):1264–1275.
8. Ruuskanen O, Järvinen A. What is the real role of respiratory viruses in severe community-acquired pneumonia? *Clin Infect Dis*. Oxford University Press; **2014**; 59(1):ciu242–73.
9. Self WH, Williams DJ, Zhu Y, Ampofo K, Pavia AT, Chappell JD, et al. Respiratory viral detection in children and adults: comparing asymptomatic controls and patients with community-acquired pneumonia. *J Infect Dis*. Oxford University Press; **2015**; :jiv323.
10. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. A Novel Rhabdovirus associated with acute hemorrhagic fever in central Africa. Wang D, editor. *PLoS Pathog*. Public Library of Science; **2012**; 8(9):e1002924.
11. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med*. **2014**; 370(25):2408–2417.
12. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. Sequence analysis of the human virome in febrile and afebrile children. Zhang C, editor. *PLOS ONE*. Public Library of Science; **2012**; 7(6):e27735.
13. Prachayangprecha S, Schapendonk CME, Koopmans MP, Osterhaus ADME, Schürch AC, Pas SD, et al. Exploring the potential of next-generation sequencing in detection of respiratory viruses. *J Clin Microbiol*. American Society for

Microbiology; **2014**; 52(10):3722–3730.

14. Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol. American Society for Microbiology*; **2011**; 49(10):3463–3469.
15. Fischer N, Indenbirken D, Meyer T, Lütgehetmann M, Lellek H, Spohn M, et al. Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. Tang Y-W, editor. *J Clin Microbiol. American Society for Microbiology*; **2015**; 53(7):2238–2250.
16. Nakamura S, Yang C-S, Sakon N, Ueda M, Tougan T, Yamashita A, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. Sommer P, editor. *PLOS ONE. Public Library of Science*; **2009**; 4(1):e4219.
17. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, et al. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol. American Society for Microbiology*; **2014**; 52(1):139–146.
18. Seo S, Renaud C, Kuypers JM, Chiu CY, Huang M-L, Samayoa E, et al. Idiopathic pneumonia syndrome after hematopoietic cell transplantation: evidence of occult infectious etiologies. *Blood. American Society of Hematology*; **2015**; 125(24):3789–3797.
19. Flygare S, Simmon KE, Miller C, Qiao Y, Kennedy B, Di Sera T, et al. Interactive web-based metagenomics analysis portal for universal pathogen detection and host response-based diagnosis and discovery. *Genome Biology. In Press* **2016**
20. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods*. **2014**; 11(12):1189–1189.
21. Hymas WC, Aldous WK, Taggart EW, Stevenson JB, Hillyard DR. Description and validation of a novel real-time RT-PCR enterovirus assay. *Clin Chem. American Association for Clinical Chemistry*; **2008**; 54(2):406–413.
22. Hymas WC, Mills A, Ferguson S, Langer J, She RC, Mahoney W, et al. Development of a multiplex real-time RT-PCR assay for detection of influenza A, influenza B, RSV and typing of the 2009-H1N1 influenza virus. *J Virol Methods*. **2010**; 167(2):113–118.
23. Dare RK, Fry AM, Chittaganpitch M, Sawanpanyalert P, Olsen SJ, Erdman DD. Human coronavirus infections in rural Thailand: a comprehensive study using real-time reverse-transcription polymerase chain reaction assays. *J Infect Dis. Oxford University Press*; **2007**; 196(9):1321–1328.

24. Saleeby El CM, Bush AJ, Harrison LM, Aitken JA, Devincenzo JP. Respiratory syncytial virus load, viral dynamics, and disease severity in previously healthy naturally infected children. *J Infect Dis.* Oxford University Press; **2011**; 204(7):996–1002.
25. Houben ML, Coenjaerts FEJ, Rossen JWA, Belderbos ME, Hofland RW, Kimpen JLL, et al. Disease severity and viral load are correlated in infants with primary respiratory syncytial virus infection in the community. *J Med Virol.* Wiley Subscription Services, Inc., A Wiley Company; **2010**; 82(7):1266–1271.
26. Zhao B, Yu X, Wang C, Teng Z, Wang C, Shen J, et al. High human bocavirus viral load is associated with disease severity in children under five years of age. Costa C, editor. *PLOS ONE.* Public Library of Science; **2013**; 8(4):e62318.
27. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods.* **2008**; 5(7):621–628.
28. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* BioMed Central Ltd; **2014**; 12(1):87.
29. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. Heimesaat MM, editor. *PLOS ONE.* Public Library of Science; **2014**; 9(4):e94249.
30. Lau SKP, Yip CCY, Lin AWC, Lee RA, So L-Y, Lau Y-L, et al. Clinical and molecular epidemiology of human rhinovirus C in children and adults in Hong Kong reveals a possible distinct human rhinovirus C subgroup. *J Infect Dis.* Oxford University Press; **2009**; 200(7):1096–1103.
31. Bochkov YA, Gern JE. Clinical and molecular features of human rhinovirus C. *Microbes Infect.* **2012**; 14(6):485–494.
32. Christensen A, Nordbø SA, Krokstad S, Rognlien AGW, Døllner H. Human bocavirus in children: mono-detection, high viral load and viraemia are associated with respiratory tract infection. *Journal of Clinical Virology.* **2010**; 49(3):158–162.
33. Lefterova MI, Suarez CJ, Banaei N, Pinsky BA. Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. *J Mol Diagn.* Elsevier; **2015**; 17(6):623–634.
34. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing through solution-based capture enrichment. *Genome Research.* Cold Spring Harbor Lab; **2015**; :gr.191049.115.

35. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio. American Society for Microbiology*; **2015**; 6(5):e01491–15.
36. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology*. **2012**; 30(11):1033–1036.
37. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nature Biotechnology*. **2015**; 33(7):689–693.
38. Kuroda M, Niwa S, Sekizuka T, Tsukagoshi H, Yokoyama M, Ryo A, et al. Molecular evolution of the VP1, VP2, and VP3 genes in human rhinovirus species C. *Sci Rep. Nature Publishing Group*; **2015**; 5:8185.

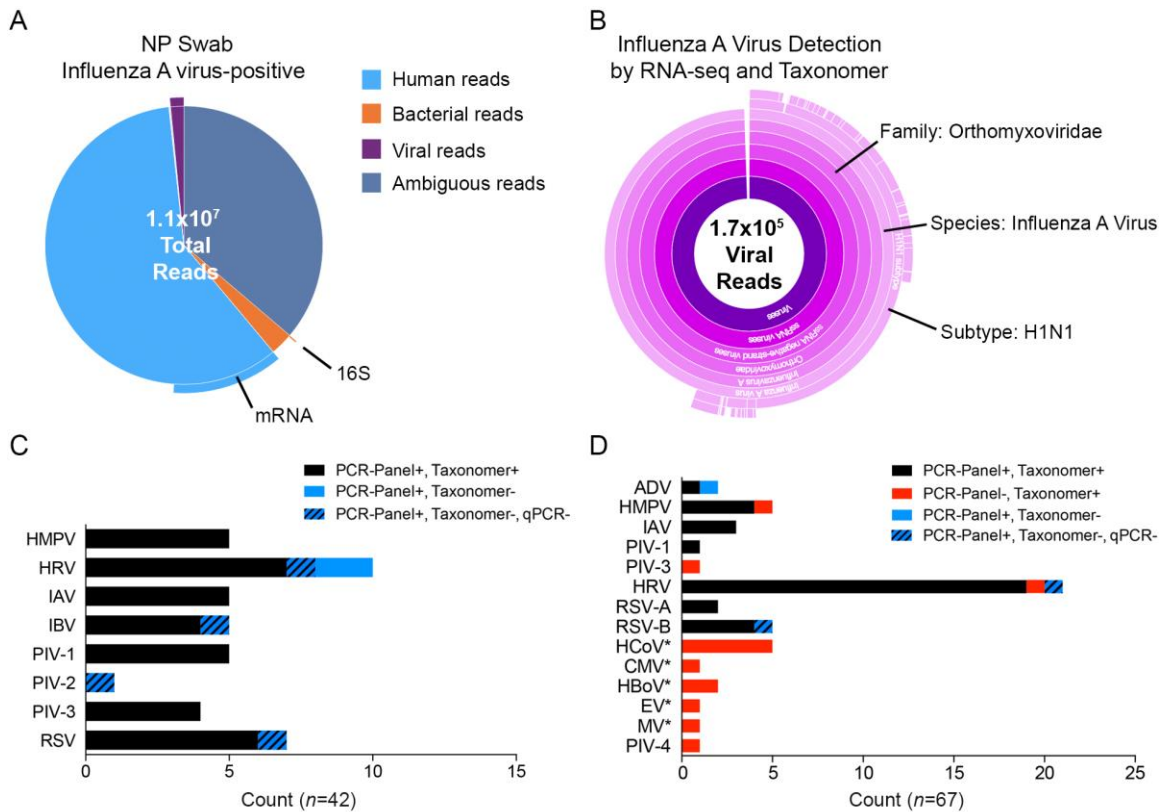


Figure 3.1. Respiratory Virus detection by RNA-seq plus Taxonomer and performance comparison with a commercial multiplex PCR panel. (A) Fractional abundance of human, bacterial, and viral sequences in RNA-seq results from an NP swab of a female infant who tested positive for influenza A virus, determined by Taxonomer and visualized through Taxonomer’s IOBIO interface (19, 20). In this sample, ~1.5% of reads were of viral and 2.7% of bacterial origin. (B) Taxonomer identified 1.74×10^5 reads as viral, of which 1.73×10^5 (99.4%) could be classified to the species level (influenza A virus), and 7.9×10^4 (45.3%) to the subtype H1N1. (C) RNA-seq plus Taxonomer identified 36 of 42 (86%) of the respiratory viruses detected by a commercial, FDA-cleared PCR panel. Four of the 6 respiratory viruses missed by metagenomics (1 each of HRV, IBV, PIV-2, and RSV) could not be detected by quantitative, multiplex real-time PCR (hashed bar). Eliminating these 4 samples from the analysis, RNA-seq plus Taxonomer identified 36 of 38 (95%) of the respiratory viruses detected by the PCR panel. (D) Using NP swab samples ($n=67$) collected during a 12-month period, the commercial PCR panel detected 37 and RNA-seq plus Taxonomer detected 48 viral infections (37 targeted by the PCR panel, 11 not targeted by the PCR panel, asterisk). Among the viruses targeted by the PCR panel, three were only identified by RNA-seq plus Taxonomer and three were only detected by the PCR panel.

ADV – adenovirus, HMPV – human metapneumovirus, IAV – influenza A virus, PIV – parainfluenza virus, HRV – human rhinovirus, RSV – respiratory syncytial virus, HCoV – human coronavirus, CMV – cytomegalovirus, HBoV – human bocavirus, EV – enterovirus, MV – measles virus, ITS – internal transcribed spacer

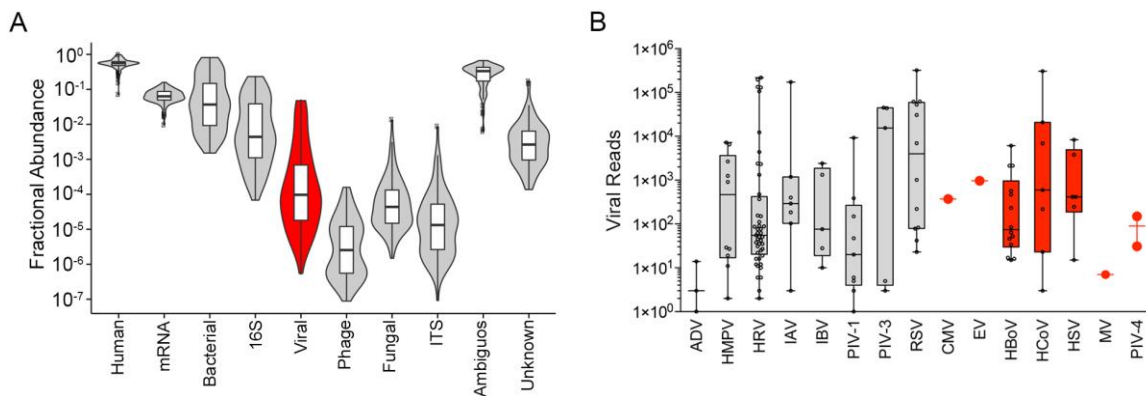


Figure 3.2. Overall taxonomic composition of RNA-seq reads and numbers of viral reads by respiratory virus. (A) Fractional abundance of reads binned as human, human mRNA, bacterial, bacterial 16S, viral, phage, fungal, fungal ITS, ambiguous, and unknown is shown as median and interquartile range (box plots) and as violin plots. Only reads identified as viral (red, median $\sim 1:10^{-4}$ reads) were used for this analysis. (B) Viral read counts differed across 5 orders of magnitude.

ADV – adenovirus, HMPV – human metapneumovirus, IAV – influenza A virus, PIV – parainfluenza virus, HRV – human rhinovirus, RSV – respiratory syncytial virus, HCoV – human coronavirus, CMV – cytomegalovirus, HBoV – human bocavirus, EV – enterovirus, MV – measles virus, ITS – internal transcribed spacer

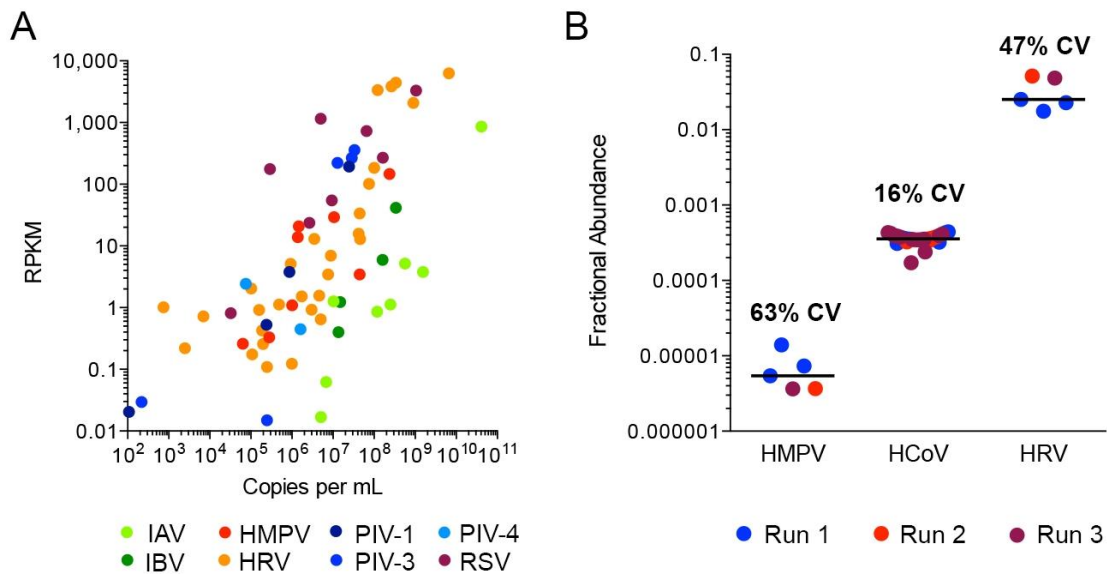


Figure 3.3. Correlation of normalized read counts with viral burden and precision of viral read abundance within and between sequencing runs. (A) The correlation between viral copies per mL of viral transport media and determined by quantitative PCR and normalized viral reads (viral reads per kb viral genome size per million total reads, RPKM) detected by RNA-seq and Taxonomer was assessed by a Spearman correlation test ($\rho=0.7$ $P<0.0001$). (B) Reproducibility was evaluated by extracting and sequencing the same sample 5 (human rhinovirus, HRV; and human metapneumovirus, HMPV) or 14 (human coronavirus, HCoV) times. Replicate libraries were prepared independently and sequenced on the same lane (within run) or different lanes (between run). Fractional abundance (viral reads per total reads) is shown for within run replicates (same color) and between-run replicates (different colors). Precision is shown as percent coefficient of variation (%CV).

RPKM)

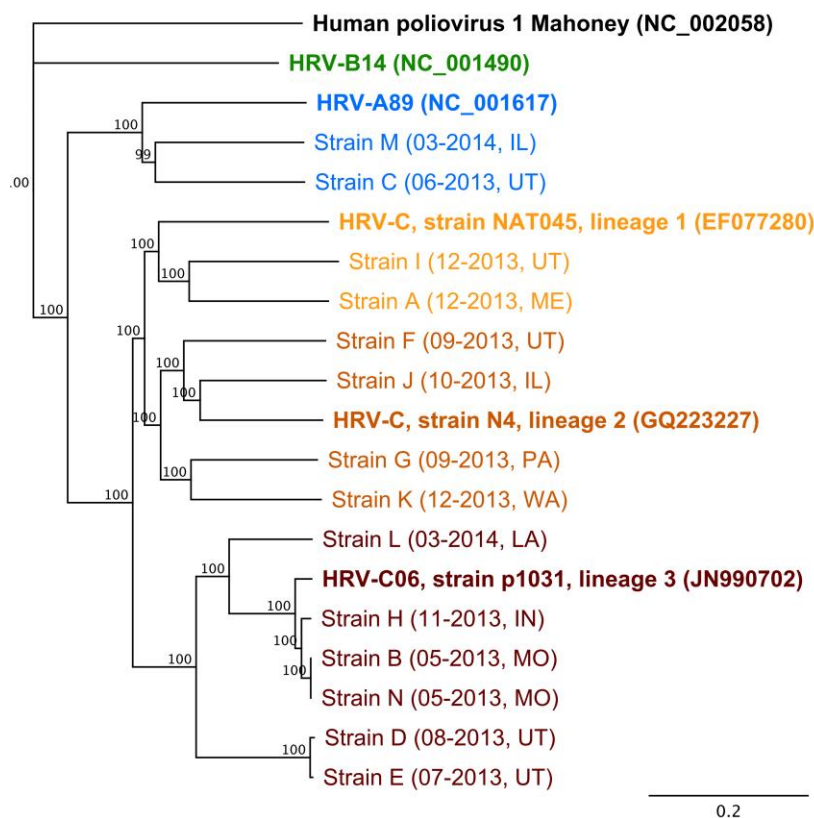


Figure 3.4. High-resolution, sequence-based typing of 14 human rhinovirus strains based on RNA-seq directly from NP swabs. Most strains belonged to rhinovirus species C ($n=12$, 86%) with 2 strains (14%) belonging to lineage 1, 4 strains (29%) belonging to lineage 2, and 6 strains (43%) belonging to lineage 3; 2 strains belonged to rhinovirus species B and no rhinovirus species A strains were detected. Near full-length sequences of 14 human rhinovirus strains (strains A through N) were aligned (MUSCLE) and a neighbor-joining consensus tree (1,000 replicates) is shown. Full-length reference sequences for Rhinovirus A (HRV-A89), B (HRV-B14), and representative full-length genome sequences from each of the Rhinovirus C lineages (EF077280, GQ223227, and JN990702, (38)) were included for comparison. Poliovirus 1 was used as outgroup. For strains sequenced as part of this study, month, year, and state of sample collection are indicated in parentheses. Colors represent species and lineage-level clades.

Table 3.1 Primer and probe sequences for the unpublished monoplex, real-time PCR assays for respiratory viruses. Asterisks indicate primer and probe sequences incorporating chemically modified bases to increase the thermodynamic stability: A* = super A base, T* = neutral base.

Virus	Name	Sequence
Parainfluenza 1	PI1-L8	AATAAATCATAAGCAATGAAAGATGGATGAA TCACTC
	PI1-E11	AATAAATCATAACTTCA*AGTTCTTCTGCA*CC A
	PI1-FAM8	MGB-FAM-ACAATCA*CGATTCA*TGG-NFQ
	PI1-FAM11	MGB-FAM-ACAATTA*CGA*TT*AGTGG-NFQ
Parainfluenza 2	PI2-L1	AATAAATCATAAGGTATAGCAGTGACTGAAC
	PI2-E1	AATAAATCATAACCATTTA*CCTAAGTGATG
	PI2-AP593-1	MGB-AP593-1-CTTTTGCGATTGATTCCA-NFQ
Parainfluenza 3	PI3-L3	AATAAATCATAAGTATATCAACTGTGTTCAAC TCC
	PI3-L4	AATAAATCATAAGTATATCAACTGTGTTTCGAC TCC
	PI3-E4	AATAAATCATAACAAGTA*CAATA*TCTTCTAT GCC
	PI3-AP525-1	MGB-AP525-1-CTTTCATCAACTTTGG-NFQ
Parainfluenza 4	PI4-L2	AATAAATCATAAAACCTCGCAGTAGTGGTCTG
	PI4-E3	AATAAATCATAACTAGATTACCATCAACAGGA AAG
	PI4-AP642-1	MGB-AP642-1-AT*T*T*ACCTAATCT*T*T*C- NFQ
hMPV	Forward	AATAAATCATAAGAGARAAYTATTTCCATG
	Reverse	AATAAATCATAATGYT*CTGT*TA*ATAT*YCM A*CAC
	Probe	MGB-FAM-G*CAT*GZ*CA*Z*T*GGT*GT*GG-Q
ADV	ADV-E1	AATAAATCATAAGATGGCCACCCCATCGA
	ADV-E2	AATAAATCATAAGATGGCTACCCCTTCGA
	ADV-E3	AATAAATCATAAGATGGCCACCCCTTCGA
	ADV-E4	AATAAATCATAAGATGGCCACTCCCTTCGA
	ADV-L1	AATAAATCATAAGGCCCGAGATGTGCATGTA
	ADV-L2	AATAAATCATAAGTCCGGCGATGTGCATGTA
	ADV-L3	AATAAATCATAAGCCCGGCGATGTGCATGTA
	ADV-FAM1	MGB-FAM-CCATTGG*GGCAG*CATCGA-NFQ
	ADV-FAM5	MGB-FAM-ACTGCGGCA*TCA*TCGA-NFQ

CHAPTER 4

VIRAL PATHOGEN DETECTION BY METAGENOMICS AND PANVIRAL PCR IN CHILDREN WITH PNEUMONIA WITH NO IDENTIFIABLE ETIOLOGY, RESULTS FROM THE CDC ETIOLOGY OF PNEUMONIA IN THE COMMUNITY (EPIC) STUDY³

4.1 Introduction

Pneumonia is the leading cause of childhood death globally and more than 2 million children die of pneumonia every year (1). In the United States, up to 50% of children ≤ 5 years with community-acquired pneumonia (CAP) require hospitalization, accounting for 110,000 admissions annually (2,3). Pathogens vary by age (4-6) but viruses are the most common causes of CAP in children ≤ 5 years, especially in the absence of lobar pneumonia and pleural effusion (6,7). However, a pathogen cannot be identified in 14-23% of children with CAP, even with extensive, state-of-the-art diagnostic testing (8-15). This situation may be due to viruses that are not part of the test panel, strains escaping detection due to genetic variation, unrecognized bacterial

³ A version of this chapter has been prepared for Emerging Infectious Diseases. Co-authors include Dr. Mark Yandell, Dr. Karen Eilbeck, Dr. Robert Schlaberg, and Keith Simmon.

infections, novel and emerging pathogens, or inadequate sample collection. Most of these limitations of current diagnostic tests could be overcome by the use of unbiased pathogen detection methods, such as high-throughput sequencing and broad-range PCR. In contrast to currently-used pathogen-specific tests, unbiased methods do not require a priori knowledge of likely pathogens; can detect previously unrecognized or unsuspected pathogens, be they viruses, bacteria, fungi, or parasites; and are tolerant to sequence polymorphism that may interfere with PCR-based detection (16,17). Two popular methods for unbiased pathogen detection are shotgun metagenomic sequencing of DNA or RNA extracted directly from patient samples and broad-range PCR amplification targeting conserved genomic regions.

In a recent national pediatric pneumonia study (Etiology of Pneumonia In the Community, EPIC), no clear etiology could be identified for ~19% of children with CAP despite use of state-of-the-art methods (8). Pathogen identification, however, is critical in order to tailor therapy appropriately, treat bacterial infections with antibiotics, and discontinue unnecessary antibiotics in cases of viral pneumonia. In addition, accurate pathogen identification is also required for effective infection-control and to examine vaccine effectiveness. The aim of the present study was to (1) reduce the proportion of children with CAP but no identifiable causative agent and to (2) assess whether unbiased pathogen detection methods may be more efficient than ever larger panels of pathogen-specific tests to determine the etiology of CAP in young children.

We applied unbiased pathogen detection using two independent approaches, RNA-seq and panviral genus/family PCR, to detect respiratory pathogens in nasopharyngeal (NP)/oropharyngeal (OP) swabs from children <5 years hospitalized with

CAP of unclear etiology (n=70). Banked samples from asymptomatic, age and season-matched children (n=90) were used as controls. RNA-seq-based metagenomics (next-generation shotgun sequencing of any RNA in a patient sample) enables sequence-independent detection of any pathogens with sufficient sequence homology to known viruses, bacteria, fungi, or parasites to allow their classification using existing reference databases [16]. This method requires no a priori knowledge of likely pathogens and represents the most unbiased method available. Panviral genus/family PCR uses broad-range PCR primers designed to detect members of viral genera and families that contain known human pathogens. As such, it allows detection of known pathogens as well as novel viruses that are close relatives of targeted viruses (18).

4.2 Materials and Methods

4.2.1 Study Population

The EPIC study population, enrollment criteria, and specimen collection are described in detail elsewhere (7). In brief, children with CAP younger than 5 years of age hospitalized at Primary Children's Hospital in Salt Lake City between January 1, 2010 and June 30, 2012 were included in this study based on pneumonia with negative pathogen detection tests per EPIC protocol (8). Inclusion criteria included acute infection, acute respiratory illness, and radiographically confirmed pneumonia. Control patients younger than 5 years of age without pneumonia were enrolled at Primary Children's Hospital between February 1, 2011 and June 30, 2012. Patients that received live attenuated influenza vaccine or underwent otolaryngologic surgery were excluded. Control patients underwent same-day elective surgery and NP/OP swabs were collected

in the operating room. To assess occurrence of respiratory symptoms or fever after enrollment as a control, a follow-up telephone interview was conducted with controls 14 days after sample collection. General exclusion criteria were hospitalization within 7 (immunocompetent children) or 90 (immunosuppressed children) days, alternative diagnosis of a respiratory disorder, newborns in continuous inpatient care, children with tracheostomy tube, cystic fibrosis, cancer with neutropenia, children who had received a solid-organ or hematopoietic stem-cell transplant (<90 days), active graft-versus-host disease, or bronchiolitis obliterans.

4.2.2 Sample Collection and Pathogen Detection per

EPIC Protocol

Combined nasopharyngeal (NP) and oropharyngeal (OP) swabs were collected as soon as possible after presentation but no more than 72 hours before or after hospital admission. Swabs were transferred into 3 ml Universal Transport Media (Diagnostic Hybrids, Inc.), refrigerated, and stored at -80°C within 24 hours. During the EPIC study, bacterial pathogens (*Haemophilus influenzae* or other gram-negative bacteria, *Staphylococcus aureus*, *Streptococcus anginosus*, *Streptococcus mitis*, *Streptococcus pneumoniae*, or *Streptococcus pyogenes*) were sought by culture (blood, endotracheal aspirate, bronchoalveolar-lavage specimen, pleural fluid) or PCR (whole blood, pleural fluid); *Chlamydomphila pneumoniae* and *Mycoplasma pneumoniae* were also sought by PCR from NP/OP swabs. Viral pathogens (adenovirus, coronavirus, HMPV, human rhinovirus, influenza, parainfluenza virus, RSV) were detected by PCR from NP/OP swabs and by serology (except human rhinovirus and coronaviruses).

4.2.3 RNA-Seq

4.2.3.1 Nucleic Acid Extraction

Combined NP/OP swab samples (75-200 μ L) in universal transport media were extracted using the QIAamp Viral RNA extraction kit (Qiagen). Extraction was carried out as described by the manufacturer with the addition of on-column DNase treatment: after AW1 wash 80 μ L of DNase I mix (Qiagen), containing 10 μ L of RNase-free DNase I and 70 μ L of Buffer RDD, was added to the QIAamp Mini column, incubated at room temperature for 15 minutes, and an additional wash step with 250 μ L AW1 was performed. This study was approved by the University of Utah (IRB_00035409) and CDC (5827) IRBs.

4.2.3.2 Library Generation

Indexed cDNA libraries were prepared with the TruSeq RNA Sample Prep Kit, following the manufacturer's instructions (Illumina, San Diego, CA). Libraries were quantified by qPCR with the Illumina Universal Library Quantification Kit (Kapa Biosystems, Inc., Wilmington, MA) and the Applied Biosystems 7900HT Fast Real-Time PCR System (Applied Biosciences). Library size and quality was assessed with a High Sensitivity DNA Analysis Kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries from 24 samples were combined in equimolar ratios for a final concentration of 9.6nM and sequenced in batches of 24 samples per lane on a HiSeq 2500 instrument (Illumina, San Diego, CA) generating 2x100bp sequencing reads.

4.2.3.3 Analysis of Metagenomics Data

Matching paired-end reads were concatenated adding a '-' between read 1 and read 2. The resulting sequences were analyzed by Taxonomer and results visualized through taxonomer.iobio.io (19). Taxa with only 1 read assigned to them were ignored. Viral taxa (other than phages) were confirmed manually by mapping the sequencing files against the relevant reference sequences in Geneious (version 8.1, Biomatters). Viral taxa identified based on <100 reads were only considered if reads were not an identical match to any other sample within the same batch by manual analysis.

4.2.4 Panviral Family/Genus PCR Panel

4.2.4.1 Nucleic Acid Extraction

Combined NP/OP swab samples (200 μ L) in universal transport media were extracted either using a manual method by the QIAamp Viral RNA extraction kit (Qiagen) or using an automatic method by the BioSprint 96 One For All kit (Qiagen) on a Kingfisher 96 platform (Thermo) according to the manufacturer instruction.

4.2.4.2 Panviral Family/Genus PCR

Panviral family/genus PCRs were designed to amplify known and potentially novel members of the viral families/genera listed below. They were designed using the CODEHOP principle to conserved genes and regions (20,21). Samples were tested with broadly reactive reverse transcriptase PCR (RT-PCR) or PCRs for the following viral families/genera: Adenoviridae, Togaviridae (Alphavirus), Anelloviridae, Arenaviridae, Astroviridae, Bornaviridae, Bunyaviridae, Caliciviridae, Circoviridae, Coronaviridae,

Flaviviridae (Flavivirus), Herpesviridae, Orthomyxoviridae (Influenzaviruses A, B and C), Paramyxoviridae, Parvoviridae, Picornaviridae (Enterovirus and Parechovirus), Polyomaviridae, Reoviridae (Aquareovirus, Orthoreovirus, Orbivirus, Rotavirus, and Seadornavirus) and Rhabdoviridae (22-29). First round RT-PCR for RNA viruses was performed with Superscript III/Platinum Taq One Step kits (Invitrogen) and Titanium Taq (Clontech) kits for the second round PCR. First and second round PCR for DNA viruses was performed with Hot Start Ex Taq kits (Takara). Positive and negative PCR controls containing mutation-engineered synthetic RNA transcript or DNA amplicon and nuclease-free water, respectively, were included in each run. PCR products were visualized on 2% agarose gels.

4.2.4.3 Sequence Confirmation

Positive bands of the expected size that had strong signal and without additional bands were cleaned up using Exonuclease I (New England Biolabs) and Shrimp Alkaline Phosphatase (Roche). Samples were incubated at 37°C for 15 minutes followed by 80°C for 15 minutes to inactivate the Exonuclease and Shrimp Alkaline Phosphatase. Positive bands of the expected size with additional bands present in the PCR products were purified using QIAquick Gel Extraction kits (Qiagen). Purified PCR amplicons were sequenced with the PCR primers in both directions on an ABI Prism 3130 Automated Capillary Sequencer (Applied Biosystems) using Big Dye 3.1 cycle sequencing kits (Life Technologies).

4.3 Results

4.3.1 Study Population and Seasonal Trends of

Test-Negative CAP

A total of 70 children with no identifiable etiology CAP and 90 asymptomatic children were included in this study (see Table 4.1). The proportion of children enrolled across seasons, in different age groups (<1 year, 1-2 years, 2-4 years), and with underlying conditions (asthma or reactive airway disease; preterm birth) was similar among cases and controls. Among the cases, 96% of children presented with fever, 83% with cough, 46% with dyspnea, 45% had radiographical evidence of consolidation, 65% of alveolar or interstitial infiltrates, and of 28% pleural effusion. The median length of hospital stay was 3 days, 28% of cases required ICU admission, but all of the cases survived. All control patients were asymptomatic at the time of enrollment. In a questionnaire administered 14 days after enrollment, 62% reported no respiratory symptoms since enrollment, 10% reported cough, and 9% reported fever. Overall, 19% of children below age 5 had CAP with no identifiable etiology (see Figure 4.1). This proportion varied by season ranging from 0-20% during most winter and spring months to 30-60% during several summer and fall months.

4.3.2 Detection of Known Respiratory Pathogens by

RNA-Seq and Panviral PCR

Banked NP/OP samples from children hospitalized with pneumonia (n=63) and asymptomatic controls (n=52) in which respiratory pathogens were detected per EPIC protocol were used to assess the ability of RNA-seq and panviral PCR to detect known

pathogens. Samples were selected to contain RNA viruses (human coronavirus, n=5; human rhinovirus, n=50; human metapneumovirus, n=12; influenza A virus, n=1; influenza B virus, n=1; parainfluenza virus 1-3, n=5; respiratory syncytial virus, n=29), DNA viruses (adenovirus, n=8), and bacterial pathogens (*M. pneumoniae*, n=7). Overall, 56 of 60 (93%) and 43 of 59 (73%) of known pathogens were detected by RNA-seq or panviral PCR in cases and controls (see Table 4.2), respectively. In cases, combined RNA-seq or panviral PCR detected 56 of 60 (93%), RNA-seq alone 54 of 60 (90%), and panviral PCR alone 34 of 60 (57%) of known pathogens (see Table 4.2). In controls, these numbers were 43 of 59 (73%) by combined RNA-seq and panviral PCR, 38 of 59 (64%) by RNA-seq, and 13 of 59 (22%) by panviral PCR. Overall, RNA-seq was at least as sensitive as panviral PCR in identifying known respiratory pathogens. Sensitivity for each of the known pathogens for detection by RNA-seq, panviral PCR, and the combination of both methods is shown in Table 4.2. Of note, RNA-seq detected none of the 8 ADV-positive samples (cases n=3, controls n=5), whereas panviral PCR detected 1 of 3 (33%) cases and 3 of 5 controls (60%). Detection of HRV was more sensitive by RNA-seq than panviral PCR. All known *M. pneumoniae* infections were detected by RNA-seq.

In addition, RNA-seq and panviral PCR detected 58 cases of previously undetected human viral infections in children with CAP and 61 in asymptomatic controls (see Figure 4.2). The largest number of previously unrecognized viral infections were caused by anelloviruses (cases n=45, controls n=38), HHV6 (cases n=8, controls n=9), and HHV7 (cases n=6, controls n=10). While these undetected viruses were unlikely pathogenic, others may have contributed to patients' symptoms (e.g. astrovirus, human

parechovirus, human bocavirus). These results demonstrated the ability of RNA-seq and panviral PCR to detect known viral pathogens and identify additional viruses.

4.3.3 Pathogen Detection in Children with Pneumonia with No Identifiable Etiology

Human viruses were detected in 53 (76%) children with pneumonia and 55 (61%) controls (see Figure 4.3). Among likely respiratory pathogens, odds ratios (OR) were highest for HBoV (OR 10.0, 95% confidence interval (CI) 2.2-46), Coxsackieviruses (OR 9.4, 95% CI 0.5-185), HRV-A (OR 4.0, 95% CI 0.4-39), adenovirus (OR 3.9, 95% CI 0.2-97), and HPIV-4 (OR 2.6, 95% CI 0.2-29). In addition, measles virus (cases, n=2), polyomaviruses (cases, MW polyomavirus, n=1; Merkel cell polyomavirus, n=1), Epstein-Barr virus (EBV, cases n=4, controls, n=2), β -papillomavirus (cases, HPV type 5, n=1), herpes simplex virus (HSV, cases, n=1), and rotavirus (cases, n=1) had odds ratios >2. Of the potential pathogens, only ADV and HRV were targeted in the EPIC study.

The most prevalent viruses were anelloviruses (49% in cases, 36% in controls), HHV6 (13% in cases, 10% in controls), and HHV7 (9% in cases, 10% in controls), all of which had odds ratios between 0.5 and 2. Parvovirus B19 (cases, n=1; controls, n=1) and Echovirus (controls, n=1) also had odds ratios of 0.5 to 2. Cytomegalovirus (CMV, cases, n=1; controls, n=3), human parechovirus (HPeV, cases, n=1; controls, n=3), and cardioviruses (controls, n=2) were more commonly detected in controls than in cases.

Figure 4.3 also shows the proportion of pathogens co-detected with other viruses. Reflecting the greater prevalence of viruses in cases, a total of 68% of detection in cases

and 33% of detection in controls were co-detections. In cases, co-detections affected 50% or more of detections for all pathogens except HPeV, HSV, beta-papillomavirus, polyomavirus, and *C. trachomatis*.

HBoV was the only pathogen with a statistically significant association ($P < 0.001$). As distribution of age and season of enrollment differed between cases and controls (see Table 4.1), we included both as categorical variables in a multivariable model. HBoV remained strongly associated with CAP in these multivariable analyses. HBoV was co-detected with other pathogens in 9 of 11 cases (82%). Co-detected viruses were anelloviruses (n=7), EBV (n=2), HRV-A (n=2), ADV (n=1), Coxsackievirus (n=1), HHV6 (n=1), and HHV7 (n=1).

4.3.4 Bacterial Pathogens in Children with Pneumonia with No Identifiable Etiology

Overall, there was great variability in the overall taxonomic composition of the RNA extracted from NP/OP samples as determined by RNA-seq and Taxonomer analysis. In at least two patients with negative results per EPIC protocol, potential bacterial pathogens were identified in great abundance (see Figure 4.4). Both samples were from 1-year-old children and in both, >95% of all sequencing reads were derived from bacteria. In patient J13, almost 95% of the bacterial reads mapped to *Pseudomonas fluorescens* (best match: strain LBUM223) and in patient M4, almost 90% of bacterial reads mapped to *Serratia marcescens* (best match: strain FGI94). Across all cases, a mean of $37 \pm 26\%$ of reads were of human origin and $46 \pm 31\%$ of bacterial origin with the

remaining reads being of viral, fungal, or not unambiguously classifiable. For controls, these numbers were $32\pm 20\%$ and $43\pm 28\%$, respectively.

4.3.5 Comparison of Viral Detection by Metagenomics and Panviral PCR

In cases, 32% of viral detections were by both methods, 22% by RNA-seq only, and 47% by panviral PCR only (see Figure 4.5). In controls, these numbers were 19%, 20%, and 61%, respectively. Most viruses that were only detected by panviral PCR were DNA viruses that are shed for extended periods by healthy individuals, e.g. anelloviruses, HHV6, HHV7. In the absence of active replication in the upper respiratory tract, these viruses are not detectable by RNA-seq. Of the remaining viruses, 46% and 21% were detected by both methods, 38% and 64% by RNA-seq only, and 15% and 14% by panviral PCR only in cases and controls, respectively.

4.4 Discussion

Making an etiologic diagnosis in patients with pneumonia is important for understanding the epidemiology, providing appropriate therapy and for limiting unnecessary use of antimicrobials. However, extensive testing using current approaches are unable to identify a pathogen in ~20% of children and ~60% of adults (8,30). We used upper respiratory tract samples from children enrolled in EPIC with well-documented pneumonia but no pathogen identified by culture, molecular, and serologic methods to assess the diagnostic yield of two unbiased pathogen detection tools, RNA-seq and broad-range PCR. In this systematic analysis, we showed that RNA-seq and PVR

had high sensitivity and specificity compared with pathogen-specific, PCR-based tests. We were able to identify potential viral pathogens in ~30% of NP/OP samples from children hospitalized with pneumonia with no identified pathogen despite extensive testing with other methods.

The detected viruses can be broadly categorized into 4 groups: (1) known respiratory pathogens, (2) viruses of unclear pathogenicity, (3) opportunistic viruses that are pathogenic in immunocompromised hosts, (4) viruses not thought to play a pathogenic role in respiratory tract illness. Among known respiratory pathogens, we detected Coxsackievirus, HRV, ADV, HPIV, human parechovirus, and measles virus, which together were detected in 33% of children with CAP. Coxsackievirus, HRV, ADV, HPIV and measles virus more often than among controls (OR >2), but due to the low frequency of detection, these differences did not reach statistical significance. Human parechovirus and echovirus can cause respiratory tract infections but were detected infrequently in both cases and controls (OR 1.3 and 0.4, respectively). The two patients with measles virus detection did not show signs of measles and had recently been vaccinated. Thus, the detected measles virus most likely represents the vaccine strain. Unfortunately, the low number of reads precluded demonstrating this by examining the complete viral genome. Cardioviruses, which are a possible cause of respiratory tract infections (31), were only detected in controls.

HBoV was the most commonly detected virus among children with CAP and no identified pathogen (13 of 70 [19%]) and detection was strongly associated with CAP (OR 10.0, $P < 0.001$). Only 2 of these infections (15%) were co-detections with other putative viral pathogens. HBoV was not targeted as part of EPIC due to uncertainty over

its role as a human pathogen. HBoV is a Parvovirus with a DNA genome. HBoV DNA can be detected for weeks to months following acute infections, which has complicated epidemiologic studies to demonstrate its pathogenicity. Pan viral PCR detected HBoV DNA in 12 of 70 cases (17.1%) and 2 of 90 asymptomatic controls (2.2%). RNA-seq identified HBoV mRNA in 10 of 70 cases (14.3%) and none of 90 asymptomatic controls (odds ratio 31.4, 95% CI 1.8-546, $P < 0.05$). Sequencing reads spanning splice sites of the viral capsid mRNA (32) confirmed that mRNA rather than genomic DNA served as sequencing template (data not shown). This strong association is in contrast with numerous PCR-based studies targeting viral genomic DNA (33,34), suggesting that detection of HBoV mRNA may serve as a marker for acute (i.e. clinically relevant) infections. While these results will need to be confirmed in larger studies, our results suggest that HBoV is associated with CAP and may be a true pathogen.

Human herpesviruses that can cause respiratory tract infections including pneumonia in immunocompromised hosts (e.g. HSV, CMV, parvovirus B19, HHV6) were also more frequently detected in cases than controls. However, children with known immune compromising conditions were excluded from EPIC. Detection of these viruses likely is a result of reactivation of latent infection rather than acute infection. Lastly, we detected a number of viruses not known to cause respiratory tract infections, including EBV, anelloviruses, HHV7, polyomaviruses, papillomavirus. Their detection in NP/OP samples of asymptomatic children as well as CAP patients is consistent with previous reports. Their detection demonstrates both the power of unbiased pathogen detection but also emphasized the importance of using appropriate controls. Interestingly, detection rates for these DNA viruses were much higher by DNA-based PVP than by RNA-seq. It

is possible that RNA-based testing may be more sensitive for DNA viruses during high level replication when mRNA is abundant. Interestingly, we detected *C. trachomatis* by RNA-seq in one infant with pneumonia. *C. trachomatis* is an important but uncommonly diagnosed cause of pneumonia in that age group.

While both RNA-seq and PVP provide broad-range detection of respiratory viruses, each has potential advantages and disadvantages. RNA-seq is highly unbiased, demonstrated by the detection of divergent enteroviruses not identified by PVP, and enables identification of non-viral pathogens, as exemplified by detection of *M. pneumoniae*, *C. trachomatis*. While RNA-seq was more sensitive than PVP for detection of RNA viruses, PVP detected more DNA viruses, many of which were not detected by RNA-seq. This may have been due in part to shedding predominantly of viral particles (containing genomic DNA) with low levels of active replication (i.e. production of mRNA) in the upper respiratory tract. Performing next-generation sequencing with both RNA-seq and DNA-seq might increase the yield for DNA viruses and bacteria, but at increased cost. However, without active replication in the upper respiratory tract where samples were taken from, these viruses may not be detectable by RNA-based approaches.

As hypothesized, broad-range pathogen detection enabled identification of viruses not part of comprehensive test panels (e.g. HBoV, Coxsackievirus, HPIV-4, Echovirus, human parechovirus), genetically divergent strains escaping PCR-based detection (e.g. HRV-A, HRV-C), and unrecognized bacterial infections (e.g. *C. trachomatis*). In addition to the Taxonomer analysis described above, we also performed *de novo* assembly of RNA-seq results and searched resulting contiguous sequences for conserved protein profiles (35) on all data from children with CAP without identifying additional putative

pathogens (data not shown). Despite these extensive efforts, a potential pathogen was not detected in 46 children (65.7%) with CAP of unknown etiology. This could have been due to testing of NP/OP swabs and not lower respiratory tract samples; inadequate timing of sample collection; polymicrobial infections caused by bacterial or fungal pathogens; or non-infectious mimics. It can also not be excluded that highly diverse viruses without homology to known human viral pathogens may have been caused CAP in some of the children. Further advancing the diagnostic yield in children with CAP is likely require additional sampling and host-based markers of infectious processes that may help confirm infectious etiologies even when a pathogen cannot be directly detected.

4.5 References

1. Wardlaw T, Salama P, Johansson EW, Mason E. Pneumonia: the leading killer of children. *Lancet*. **2006**; 368(9541):1048–1050.
2. Margolis P, Gadomski A. The rational clinical examination. Does this infant have pneumonia? *JAMA*. **1998**; 279(4):308–313.
3. Lee GE, Lorch SA, Sheffler-Collins S, Kronman MP, Shah SS. National hospitalization trends for pediatric pneumonia and associated complications. *Pediatrics*. American Academy of Pediatrics; **2010**; 126(2):204–213.
4. Sandora TJ, Harper MB. Pneumonia in hospitalized children. *Pediatr Clin North Am*. **2005**; 52(4):1059–81– viii.
5. British Thoracic Society Standards of Care Committee. British Thoracic Society guidelines for the management of community acquired pneumonia in childhood. *Thorax*. BMJ Group; 2002. pp. i1–24.
6. McIntosh K. Community-acquired pneumonia in children. *N Engl J Med*. **2002**; 346(6):429–437.
7. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. **2015**; 372(9):835–845.

8. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. **2015**; 372(9):835–845.
9. Ruuskanen O, Lahti E, Jennings LC, Murdoch DR. Viral pneumonia. *The Lancet*. **2011**; 377(9773):1264–1275.
10. Huguenin A, Moutte L, Renois F, Leveque N, Talmud D, Abely M, et al. Broad respiratory virus detection in infants hospitalized for bronchiolitis by use of a multiplex RT-PCR DNA microarray system. *J Med Virol*. Wiley Subscription Services, Inc., A Wiley Company; **2012**; 84(6):979–985.
11. Michelow IC, Olsen K, Lozano J, Rollins NK, Duffy LB, Ziegler T, et al. Epidemiology and clinical characteristics of community-acquired pneumonia in hospitalized children. *Pediatrics*. **2004**; 113(4):701–707.
12. Juvén T, Mertsola J, Waris M, Leinonen M, Meurman O, Roivainen M, et al. Etiology of community-acquired pneumonia in 254 hospitalized children. *Pediatr Infect Dis J*. **2000**; 19(4):293–298.
13. Harris M, Clark J, Coote N, Fletcher P, Harnden A, McKean M, et al. British Thoracic Society guidelines for the management of community acquired pneumonia in children: update 2011. *Thorax*. BMJ Publishing Group Ltd and British Thoracic Society; 2011. pp. ii1–23.
14. Wubbel L, Muniz L, Ahmed A, Trujillo M, Carubelli C, McCoig C, et al. Etiology and treatment of community-acquired pneumonia in ambulatory children. *Pediatr Infect Dis J*. **1999**; 18(2):98–104.
15. Pavia AT. What is the role of respiratory viruses in community-acquired pneumonia?: What is the best therapy for influenza and other viral causes of community-acquired pneumonia? *Infect Dis Clin North Am*. **2013**; 27(1):157–175.
16. Lipkin WI. The changing face of pathogen discovery and surveillance. *Nature Publishing Group*. Nature Publishing Group; **2013**; 11(2):133–141.
17. Chiu CY. Viral pathogen discovery. *Curr Opin Microbiol*. **2013**; 16(4):468–478.
18. Morens DM. Novel and re-emerging respiratory viral diseases: Novartis Foundation symposium 290. *Emerging Infect Dis*. [serial on the Internet]. **2009**. Available from: <http://www.cdc.gov/EID/content/15/6/999b.htm>
19. Flygare S, Simmon KE, Miller C, Qiao Y, Kennedy B, Di Sera T, et al. Interactive web-based metagenomics analysis portal for universal pathogen detection and host response-based diagnosis and discovery. *Genome Biology*. In Press **2016**.

20. Rose TM, Henikoff JG, Henikoff S. CODEHOP (CONsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Research*. Oxford University Press; **2003**; 31(13):3763–3766.
21. Rose TM. CODEHOP-mediated PCR - a powerful technique for the identification and characterization of viral genomes. *Virol J*. BioMed Central Ltd; **2005**; 2(1):20.
22. Conrardy C, Tao Y, Kuzmin IV, Niezgoda M, Agwanda B, Breiman RF, et al. Molecular detection of adenoviruses, rhabdoviruses, and paramyxoviruses in bats from Kenya. *Am J Trop Med Hyg*. American Society of Tropical Medicine and Hygiene; **2014**; 91(2):258–266.
23. Finkbeiner SR, Li Y, Ruone S, Conrardy C, Gregoricus N, Toney D, et al. Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *J Virol*. American Society for Microbiology; **2009**; 83(20):10836–10839.
24. Phaneuf CR, Oh K, Pak N, Saunders DC, Conrardy C, Landers JP, et al. Sensitive, microliter PCR with consensus degenerate primers for Epstein Barr virus amplification. *Biomed Microdevices*. **2013**; 15(2):221–231.
25. Schatzberg SJ, Li Q, Porter BF, Barber RM, Claiborne MK, Levine JM, et al. Broadly reactive pan-paramyxovirus reverse transcription polymerase chain reaction and sequence analysis for the detection of Canine distemper virus in a case of canine meningoencephalitis of unknown etiology. *J Vet Diagn Invest*. **2009**; 21(6):844–849.
26. Tao Y, Shi M, Conrardy C, Kuzmin IV, Recuenco S, Agwanda B, et al. Discovery of diverse polyomaviruses in bats and the evolutionary history of the Polyomaviridae. *J Gen Virol*. Microbiology Society; **2013**; 94(Pt 4):738–748.
27. Tong S, Chern S-WW, Li Y, Pallansch MA, Anderson LJ. Sensitive and broadly reactive reverse transcription-PCR assays to detect novel paramyxoviruses. *J Clin Microbiol*. American Society for Microbiology; **2008**; 46(8):2652–2658.
28. Tong S, Conrardy C, Ruone S, Kuzmin IV, Guo X, Tao Y, et al. Detection of novel SARS-like and other coronaviruses in bats from Kenya. *Emerging Infect Dis*. **2009**; 15(3):482–485.
29. Tong S, Singh J, Ruone S, Humphrey C, Yip CCY, Lau SKP, et al. Identification of adenoviruses in fecal specimens from wild chimpanzees (*Pan troglodytes schweinfurthii*) in western Tanzania. *Am J Trop Med Hyg*. American Society of Tropical Medicine and Hygiene; **2010**; 82(5):967–970.
30. Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM, et al. Community-acquired pneumonia requiring hospitalization among U.S. adults. *N Engl J Med*. **2015**; 373(5):415–427.

31. Lin T-L, Lin T-H, Chiu S-C, Huang Y-P, Ho C-M, Lee C-C, et al. Molecular epidemiological analysis of Saffold coronavirus genotype 3 from upper respiratory infection patients in Taiwan. *Journal of Clinical Virology*. Elsevier; **2015**; 70:7–13.
32. Chen AY, Cheng F, Lou S, Luo Y, Liu Z, Delwart E, et al. Characterization of the gene expression profile of human bocavirus. *Virology*. **2010**; 403(2):145–154.
33. Pellett PE. Indictment by Association: once is not enough. *J Infect Dis*. Oxford University Press; **2015**; 212(4):jiv045–512.
34. Williams JV. Déjà vu all over again: Koch's postulates and virology in the 21st century. *J Infect Dis*. Oxford University Press; **2010**; 201(11):1611–1614.
35. Eddy SR. Accelerated profile HMM searches. Pearson WR, editor. *PLOS Comput Biol*. Public Library of Science; **2011**; 7(10):e1002195.

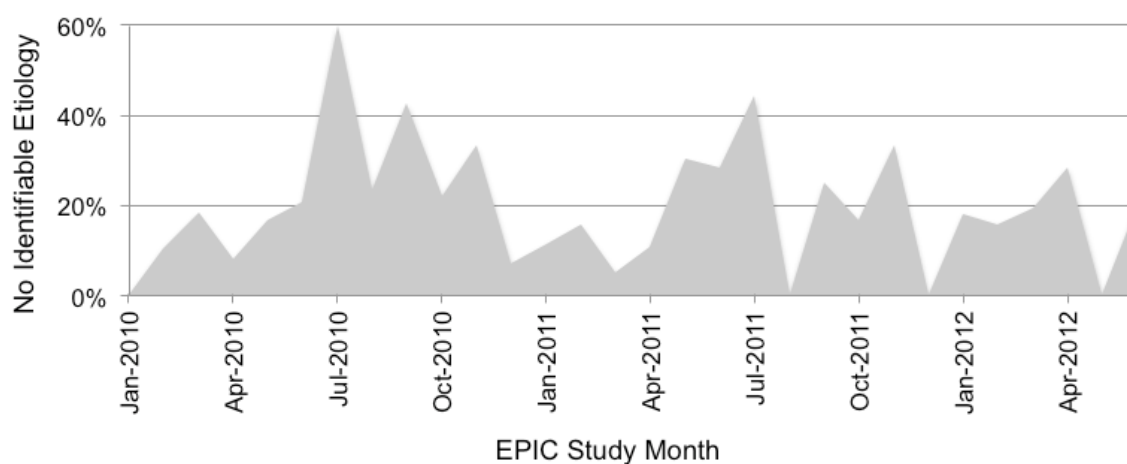


Figure 4.1. Proportion of children with CAP with no identifiable etiology during the 2½-year EPIC study. The proportion of children below age 5 who had no pathogen identified per EPIC testing protocol (n=619) ranged between 0% and 60%. CAP with no identifiable etiology (n=70) was more common during summer/fall than during winter/spring respiratory season.

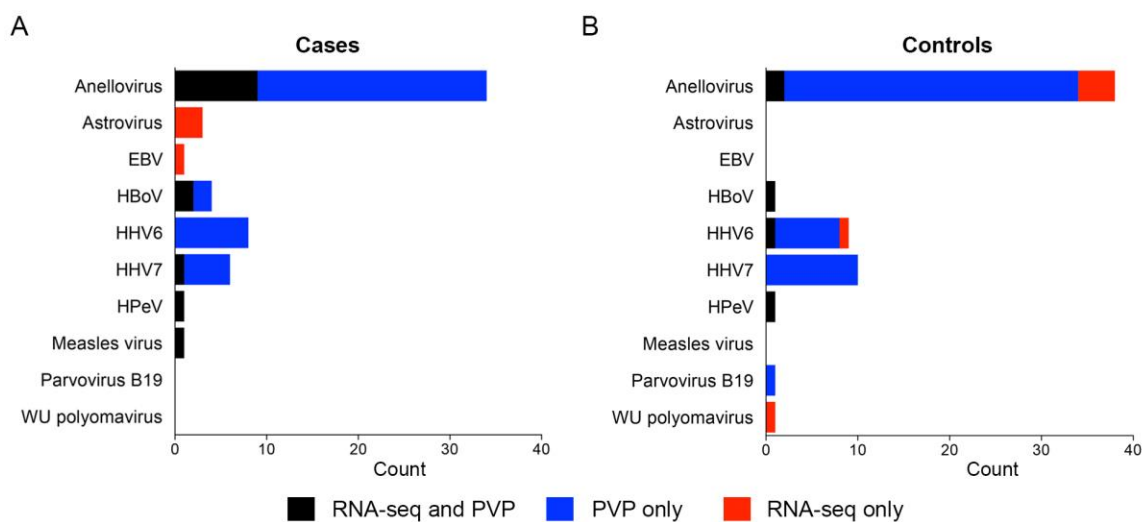


Figure 4.2. Detection of additional human viruses by RNA-seq and panviral PCR in EPIC participants with positive pathogen-specific tests. Human viruses detected by RNA-seq and panviral PCR that were not targeted in EPIC included human parechovirus (HPeV), human bocavirus (HBoV), Epstein Barr Virus (EBV), human herpesvirus 6 (HHV6), and human herpesvirus 7 (HHV7). (PVP – panviral genus/family PCR).

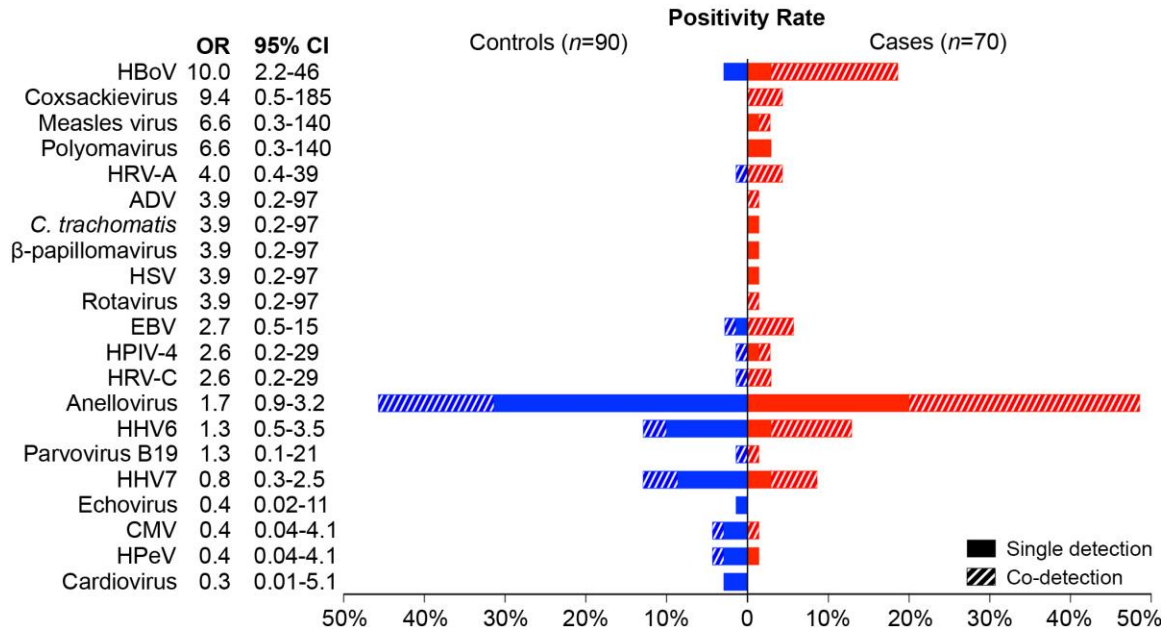


Figure 4.3. Viruses detected by RNA-seq and/or panviral PCR in children with pneumonia with no identifiable etiology ($n=70$, red) and asymptomatic controls ($n=90$, blue). A total of 20 different human viruses were detected in NP/OP samples. In addition, *Chlamydia trachomatis* was detected in one newborn child with pneumonia. Fifteen viruses were more frequently detected in cases than controls (odds ratios > 1), with HBoV ($P<0.001$) having significant associations with CAP (Fisher's exact test). Hashed bars indicate co-detection with one or more additional viruses.

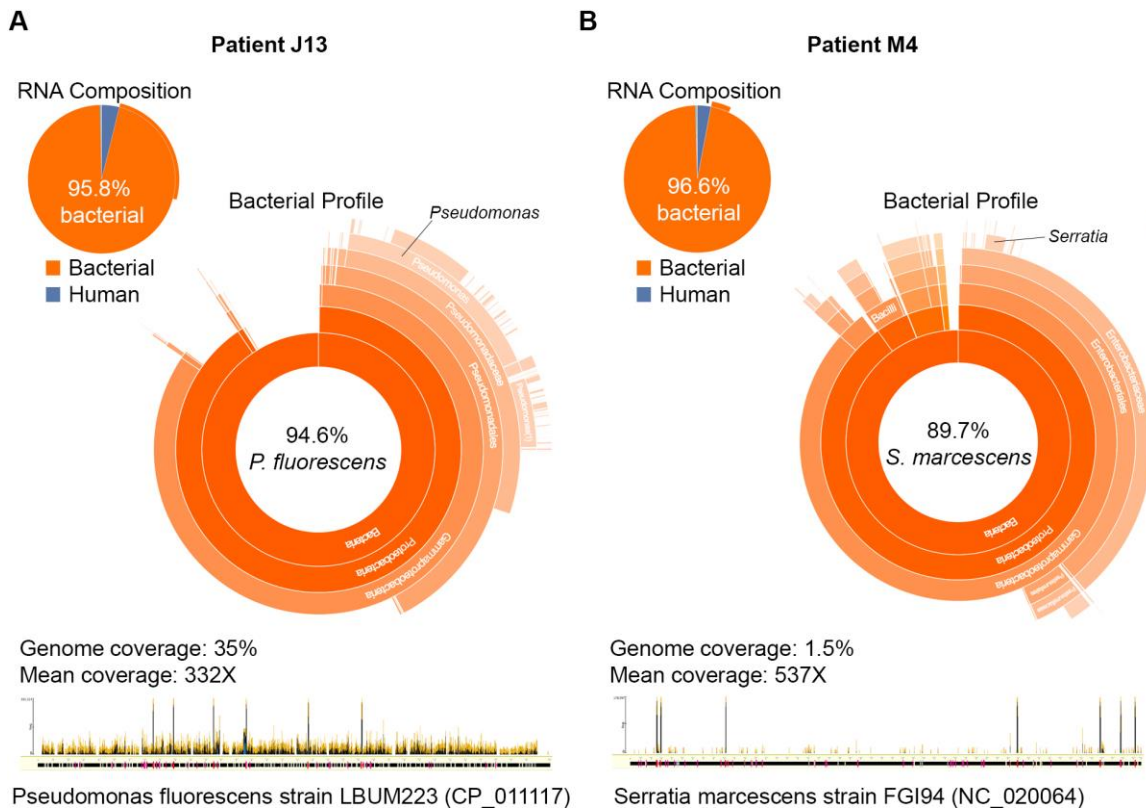


Figure 4.4. An abundant bacterial flora (>95% of sequencing reads) dominated by a single potential pathogen was detected by RNA-seq in NP/OP samples of two children. **A**, 94.6% of sequencing reads generated from the NP/OP sample of patient J13 was identified as *Pseudomonas fluorescens*; 35% of the genome of strain LBUM223 (NCBI accession number CP_011117) was covered at a mean of 332X. **B**, In patient M4, 89.7% of sequencing reads were derived from *Serratia marcescens* covering 1.5% of the genome sequence of strain FGI94 (NCBI accession number NC_020064) at a mean of 537X.

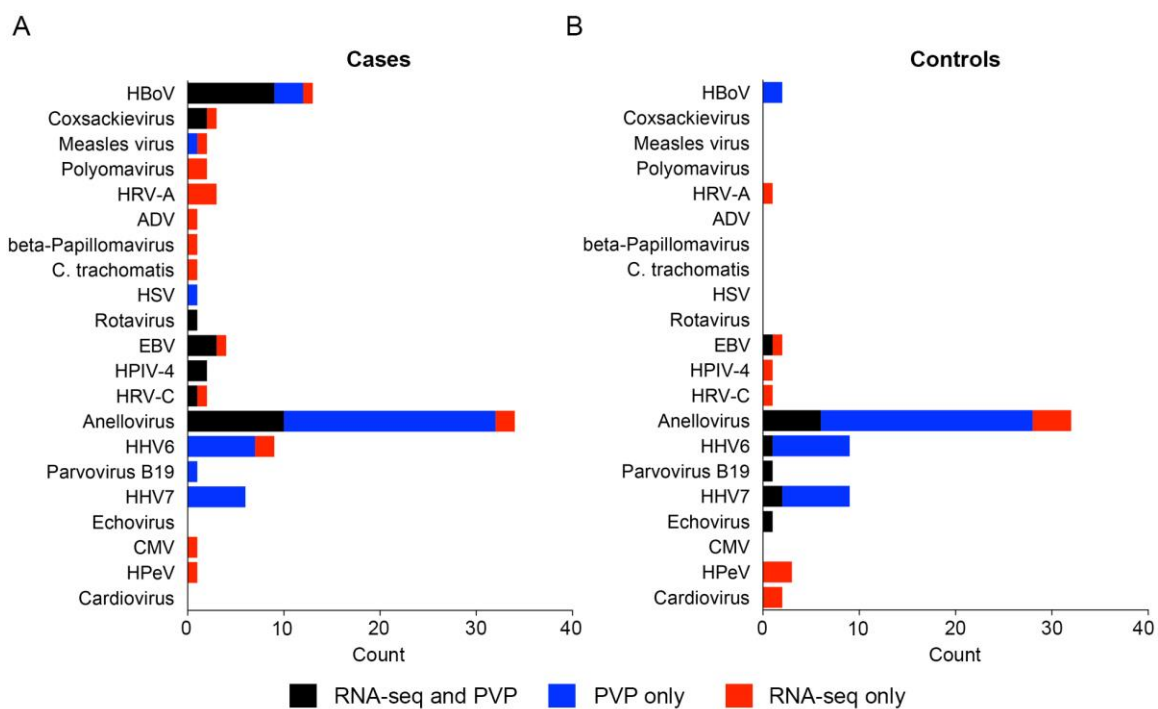


Figure 4.5. Pathogen detection in children with CAP with no identified etiology and asymptomatic controls by method.

Table 4.1 Demographic and clinical information of children selected for study with community acquired pneumonia.

	Community Acquired Pneumonia (n=71)	Control Patients (n=92)	P-value (chi square)
Age group, n (%)			p = 0.05
<1 yr	14 (20%)	23 (25%)	
1-2 yr	27 (38%)	19 (21%)	
2-4 yr	30 (42%)	50 (54%)	
Month of enrollment			p = 0.04
January - March	23 (33%)	13 (14%)	
April - June	25 (36%)	37 (41%)	
July - September	16 (23%)	25 (28%)	
October - December	6 (9%)	15 (17%)	
Symptom, n (%)			n/a
Cough	59 (83%)	0	
Fever	68 (96%)	0	
Anorexia	54 (76%)	0	
Dyspnea	33 (46%)	0	
Symptoms on follow-up			n/a
None	0	57 (62%)	
Cough	0	9 (10%)	
Fever	0	8 (9%)	
Underlying condition, n (%)			n/a
Asthma or reactive airway disease	6 (8%)	3 (3%)	
Preterm birth among children <2 yr	12 (17%)	15 (16%)	
Radiographic findings, n (%)			n/a
Consolidation	32 (45%)	0	
Alveolar or intestinal infiltrate	46 (65%)	0	
Pleural effusion	20 (28%)	0	
Hospitalization			n/a
Length of stay, median (IQR)	3 (2-4)	0	
ICU admission	20 (28%)	0	
Death in the hospital	0 (0%)	0	

Table 4.2 Performance of RNA-seq and panviral PCR compared to pathogen-specific real-time PCR performed per EPIC protocol.

Cases	Positive in this Study												Additional Positives		Combined		RNA-seq		Panviral PCR	
	EPIC				RNA-seq				RNA-				PVP		Sens.		Spec.			
	Positive	Combined	RNA-seq	PVP	Combined	PVP	Combined	seq	Combined	Sens.	Spec.	Combined	Sens.	Spec.	RNA-seq	Sens.	Spec.	Panviral PCR	Sens.	Spec.
IAV	0	0	0	0	0	0	0	0	0	0	n/a	n/a	100%	n/a	100%	n/a	100%	n/a	100%	
IBV	1	1	1	0	0	0	0	0	0	0	100%	100%	100%	100%	100%	100%	0%	0%	100%	
HMPV	8	8	7	7	0	0	0	0	0	0	100%	100%	100%	88%	100%	88%	100%	88%	100%	
HRV	12	12	12	2	0	0	0	0	0	0	100%	100%	100%	100%	100%	100%	17%	17%	100%	
RSV	24	23	23	21	0	0	0	0	0	0	96%	100%	100%	96%	100%	88%	100%	88%	100%	
HPIV	3	2	2	1	1	1	1	1	1	1	67%	98%	98%	67%	98%	33%	98%	33%	98%	
ADV	3	1	0	1	0	0	0	0	0	0	33%	100%	100%	0%	100%	33%	100%	33%	100%	
HCoV	2	2	2	2	0	0	0	0	0	0	100%	100%	100%	100%	100%	100%	100%	100%	100%	
MPNE	7	7	7	n/a	1	1	1	1	n/a	n/a	100%	98%	98%	100%	98%	n/a	n/a	n/a	n/a	
Controls																				
IAV	1	0	0	0	0	0	0	0	0	0	0%	100%	100%	0%	100%	0%	100%	0%	100%	
IBV	0	0	0	0	0	0	0	0	0	0	n/a	100%	100%	n/a	100%	n/a	100%	n/a	100%	
HMPV	4	0	0	0	0	0	0	0	0	0	0%	100%	100%	0%	100%	0%	100%	0%	100%	
HRV	38	34	34	5	1	1	1	1	0	0	89%	96%	96%	89%	96%	13%	100%	13%	100%	
RSV	5	1	0	1	0	0	0	0	0	0	20%	100%	100%	0%	100%	20%	100%	20%	100%	
HPIV	3	3	3	2	1	1	1	1	0	0	100%	98%	98%	100%	98%	67%	100%	67%	100%	
ADV	5	3	0	3	0	0	0	0	0	0	60%	100%	100%	0%	100%	60%	100%	60%	100%	
HCoV	3	2	1	2	0	0	0	0	0	0	67%	100%	100%	33%	100%	67%	100%	67%	100%	
MPNE	0	0	0	n/a	0	0	0	0	n/a	n/a	n/a	100%	100%	n/a	100%	n/a	n/a	n/a	n/a	

PVP – panviral genus/family PCR, Sens. – sensitivity, Spec. – specificity, IAV – influenza A virus, IBV – influenza B virus, HMPV – human metapneumovirus, HRV – human rhinovirus, RSV – respiratory syncytial virus, HPIV – human parainfluenza viruses, ADV – adenovirus, HCoV – human coronavirus, MPNE – *Mycoplasma pneumoniae*

CHAPTER 5

CONCLUSION

This dissertation represents an advance in the field of clinical metagenomics and computational biology. In Chapter 1 and 2, we outlined some of the important considerations for classifying random DNA fragments in metagenomic samples. These considerations included classifying reads from DNA regions with substantial database coverage for known species and using marker genes with clearly understood evolutionary rates. With these concerns in mind, we created Taxonomer, a kmer-based tool that uses a novel analysis workflow with simultaneous read binning and marker gene classification.

In Chapter 2, we benchmarked Taxonomer against other rapid analysis tools that currently represent the best practices for analyzing metagenomic data. These benchmarks established Taxonomer as the most precise rapid metagenomics tool for the detection of virus, bacteria, and fungi to-date. Additionally, we showed Taxonomer has the ability to measure human transcript expression to predict a viral infection, enhancing the tools functionality for clinical usage and interpretation. We combined Taxonomer with the iobio visualization framework (1) to create taxonomer.iobio, which couples the speed and accuracy of Taxonomer with interactive visualizations that allow users to explore data from the highest to lowest taxonomic level. Several use cases were presented to show how Taxonomer and metagenomics could be implemented in the research and clinical

setting.

Chapter 3 compared the metagenomics-based Taxonomer approach to a FDA-approved and commercially available respiratory virus panel (RVP). Sensitivity was greater when detecting viruses by Taxonomer. This improved sensitivity was largely a result of detecting divergent viruses or virus not covered in the FDA cleared RVP panel. While not presented in Chapter 3, the metagenomics approach also provides an avenue to measure bacterial and fungal pathogens, without additional analysis time. This is not possible using the RVP. The benefit of bacterial detection is seen in Chapter 4.

Challenges still exist for the implementation of metagenomics in the clinical laboratory. These challenges include turn-around-time, cost, and interpretation. The RVP panel has a turn-around-time of 6 hours and a simple +/- interpretation of viral presence. In contrast, the metagenomics analysis used in Chapter 3 required ~14 hours of library prep and ~11 days on the sequencer before the data were processed by Taxonomer. When using Taxonomer, the user must explore the resulting data to identify pathogens among other micro flora. The sample cost for metagenomics sequencing can be comparable to that of RVP panel when processing ~24 samples on a single sequencing run. The feasibility of waiting for additional samples prior to testing is not ideal for a time-sensitive diagnostic. One additional attribute of the RVP panel is its ability to be deployed at point-of-care; this is currently not possible for metagenomics-based testing. For small to medium size laboratories, point-of-care testing with commercially available RVP panels may still present the best option for respiratory diagnostics; however, larger laboratories may be able to leverage the additional information that metagenomics-based testing provides in order to report more actionable clinical information like antibiotic

resistance and secondary infections.

In Chapter 4, we utilized metagenomics and Taxonomer to characterize pediatric pneumonia using samples from a large multicenter study. This study was initiated by the CDC to understand the etiology of pneumonia in the community (EPIC) (2,3). Fourteen to 23 percent of pediatric patients with radiological confirmed pneumonia had no pathogen identified using standard diagnostics. In the children that had no identifiable etiology, human viruses were detected in (53) 76% and (55) 61% of the cases and controls, respectively. Several likely respiratory pathogens were identified, including human bocavirus, coxsackieviruses, and human rhinovirus. Additionally, we co-detected viruses in a number of samples of which the implication on pathogenicity is still uncertain. Taxonomer also identified two likely bacterial pathogens, *Pseudomonas fluorescens* and *Serratia marcescens*, in patients with no etiology, showing the added benefit of bacterial screening that metagenomics provides.

Chapter 4 also compared a panviral PCR targeting multiple viral families to the metagenomics approach. While detection limits theoretically should be better for PCR, the metagenomic approach was able to detect more viruses than the PCR panel. This along with the data from Chapter 3 clearly shows that metagenomics can provide suitable depth of coverage to provide a detection limit that rivals PCR. The PCR panel was superior for the detection of DNA viruses (adenovirus), which would need to be actively replicating to be detected via RNA-seq metagenomics.

The majority experiments described in this dissertation were performed on the Illumina HiSeq sequencing platform, which requires ~11 days to finish a sequencing run. This time frame is not ideal and it should be noted that sequencing technologies do

currently exist that enable the sequencing to be done in ~24hrs with similar levels of coverage to those obtain in our studies. If trends continue, new sequencing technologies will enable metagenomics to be performed as rapidly and as inexpensively as commercially available panels or PCR-based tests. In lieu of widespread adoption, the role of metagenomics in clinical laboratories may be to help influence the design of the next-generation of analyte-based tests.

In our studies, metagenomics provided meaningful insight beyond current diagnostics options. In Chapter 3 and 4, we showed that human bocavirus was commonly detected in patients with pneumonia. It was also shown to be a likely cause of pneumonia because it was more often detected in cases than controls. However, bocavirus is not part of the analytes detected in either the RVP panel in Chapter 3 or the EPIC study in Chapter 4. This illustrates how metagenomics can be used to influence the next-generation of analyte-based tests.

Taxonomer's iobio interactive visual front-end simplifies data interpretation, which can be overwhelming for metagenomics data. It allows the user to drill down into the taxonomic lineage and filter results based on coverage and read count. Currently, a large number of reads are simply classified at a high taxonomic level (e.g. Bacteria or Fungi); potentially these reads can offer clinically useful information. For example, if classification of the 16S rRNA reads indicate the presence of a *Staphylococcus* species, then the remaining bacterial reads can be interrogated for markers that may provide species level resolution or for pathogenic features like methicillin resistance. As more of these reads are used to obtain clinical information, the taxonomer.iobio visualization will have to be improved to incorporate these data in a meaningful and intuitive way so it can

be consumed by clinicians.

The speed of Taxonomer is an attribute querying kmer-based datasets, and its accuracy is a result of a novel workflow, which bins sequences and then classifies the reads from genes (16S rRNA; ITS) that are most likely to provide reliable results. We are indebted to the providers of the gene resources we utilized in Taxonomer. These resources delivered highly accurate taxonomic information attached to quality controlled reference sequences (4-7). What is missing are data sources, which could link sample source, diagnosis, co-pathogens, antibiotic resistances, antibiograms, alternative marker genes, pathogenic genes, and epidemiological information to these references or taxa that would allow more clinically oriented information to be queried and then displayed. It is these resources that will allow Taxonomer and other rapid approaches to leverage the entirety of the data generated by metagenomics and concisely display the relevant clinical information. Ultimately, it is tools like Taxonomer that will facilitate the creation of these resources. In the meantime, Taxonomer provides superior ability to profile microbial communities and detect novel microorganisms that standard diagnostics tests cannot achieve, even though challenges remain for widespread implementation of metagenomics.

This dissertation represents an important step forward in biomedical informatics. When the work herein was initiated, metagenomics analysis pipelines could take days to weeks to complete (8). The analysis time was greatly influence by the sample composition. Generally, samples with more fungal or bacterial sequences have longer analysis times. Taxonomer's data flow design reduced the analysis time and alleviated the influence of sample composition by allowing simultaneous screening of relevant taxa

(see Figure 2.1). Along with significant speed increases, Taxonomer leveraged microbiological resources that had been, by a large part, developed outside of medical research where lives are not a stake (4,5,9). This project established algorithms, intelligent data flows, intuitive interactive visualizations, and comprehensive testing to enable ‘*clinical metagenomics*’ from those resources. While many challenges remain in clinical metagenomics the tools and information produced as part of this dissertation have moved the field forward.

5.1 References

1. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods*. **2014**; 11(12):1189–1189.
2. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. **2015**; 372(9):835–845.
3. Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM, et al. Community-acquired pneumonia requiring hospitalization among U.S. adults. *N Engl J Med*. **2015**; 373(5):415–427.
4. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol. American Society for Microbiology*; **2006**; 72(7):5069–5072.
5. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research. Oxford University Press*; **2014**; 42(Database issue):D643–8.
6. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research. Oxford University Press*; **2014**; 42(Database issue):D633–42.
7. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et al. Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol*. **2013**; 22(21):5271–5277.

8. Kostic AD, Ojesina AI, Pedomallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnology*. Nature Publishing Group; **2011**; 29(5):393–396.
9. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*. Oxford University Press; **2013**; 42(D1):D633–D642.