

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Virus-host Cophylogeny Motivated Viral Detection and Discovery Using Viral Microarrays and High-throughput Sequencing

Permalink

<https://escholarship.org/uc/item/90k036tt>

Author

Wootton, Sharon Chao

Publication Date

2011

Peer reviewed|Thesis/dissertation

**Virus-host Cophylogeny Motivated Viral Detection and Discovery Using Viral
Microarrays and High-throughput Sequencing**

by

Sharon Chao Wootton

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Copyright (2011)

Sharon Chao Wootton

for my dear grandparents
and
Cai Xia Ayi,
for her unforgettable wisdom

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Joe DeRisi for his mentorship, support, and for making graduate school one of the best decisions of my life. Everyday in the DeRisi lab is exciting, and Joe has made this possible with his energy and limitless passion for science. He has always given me the freedom to follow my own interests, provided the lab the coolest and most current technology, and I can't be more grateful for his guidance and support.

I would also like to thank Don Ganem for his advice, encouragement, and mentorship. He has always brought indispensable insight to the table, and his presence at Virochip subgroup meetings is always both entertaining and immensely constructive.

I am grateful for all the input and company of my current and past lab-mates. Patrick Tang first introduced me to the bench, and I will always remember his patience and encouragement. I am thankful for the advice of some Virochip postdoc veterans--Graham Ruby, Amy Kistler, and Charles Chiu. A special thanks to the computational quad, Michelle Dimon, Vida Ahyong , Peter Skewes-Cox, and Miguel Betegon, for their entertainment and keeping me sane with our coffee breaks. And of course, I would like to thank Tara Christiansen for holding the lab together and everything she's done for me.

I would like to acknowledge all the collaborations that have made much of this work possible. I would especially like to thank Hal Collard and Dong Soon Kim for working with me on the IPF project. Their profound clinical perspective made the process very rewarding.

I am grateful for all the guidance given to me by my dissertation committee--Patsy Babbitt, Deborah Dean, Ryan Hernandez, Ian Holmes, and Joe. Patsy has been an invaluable source of advice throughout the years, and I can't thank her enough. She always pointed me in the right direction and seemed to have a solution whenever I encountered a computational roadblock. I would like to thank Deborah Dean for her guidance through the quals and thesis process and for always being such a pleasure to talk to. After attending one of her talks at Berkeley in my first year of grad school, I knew I would be very lucky to have her on my committee. A very special thanks to Ryan Hernandez for filling in on my committee on such short notice. And I will eternally be grateful to Ian Holmes, who introduced me to computational biology in undergrad, and I will always remember how excited I was to take his course and work in his lab.


I would like to thank my mother Ping Duan, my father Wei Yang Chao, my sister Angela Chao, and my baby brother Bryan Chao for all their love and support throughout the years. I was born in my parents' Berkeley grad school years, and it's always been a dream of mine to come back and do the same. I would also like to thank my in-laws, the Woottons and the Howards. It is amazing to have the support of all forty of you.

And finally, I would like to thank my husband Jeff Wootton, who has been there from the beginning. He is an inspiration, and I will always be grateful for his love, laughter, and faith in me.

**Virus-host Cophylogeny Motivated Viral Detection and Discovery Using Viral
Microarrays and High-throughput Sequencing**

by

Sharon Chao Wootton


.....
Joseph DeRisi, PhD. Chair

ABSTRACT

The emergence of high-throughput genomic technologies has markedly accelerated virus detection, virus discovery, and viral metagenomics. DNA viral microarrays and high-throughput sequencing platforms allow for the unbiased detection of all pathogens in parallel. In this dissertation, the capacity for a novel or unexpected viral etiology in a number of idiopathic diseases is investigated. More specifically, serum from individuals with acute liver failure, ocular fluids from patients with uveitis, and bronchoalveolar lavage from individuals experiencing an acute exacerbation of idiopathic pulmonary fibrosis are analyzed for all viruses.

An alternative to the syndromic approach to virus discovery is also taken in tandem, where a hypothesized gap in virus phylogeny is targeted specifically. Overlaying viral hosts onto a phylogenetic tree of a conserved herpesvirus gene reveals a clade of herpesviruses found in a number of primates, with a distinct gap suggesting a

homologous herpesvirus in human hosts. Attributes of this gap, such as sequence similarity, viral lifestyle, and tropism, are used to design a set of discovery projects. Saliva samples from patients with full-blown AIDS were collected and banked in an era before antiretroviral therapy, and tonsils, a lymphocyte rich tissue, from adolescents with recurrent tonsillitis was collected and treated with a chemical known to induce the herpesvirus lytic gene cascade. These samples were analyzed with the pan-viral microarray and high-throughput sequencing for the presence of a novel human herpesvirus.

This method of targeted analysis of a hypothesized virus gap is then generalized to all viruses in an effort to precipitate hypotheses of discovery targets, or gaps, in virus trees when they are overlaid onto their corresponding host trees. The propensity for viruses to coevolve with their hosts motivates a framework for translating host homology to homology between known viruses and their suggested orthologs (gap-recognition). These proposed gaps present insight into viral diversity and evolution, and can also be used to motivate targeted microarray-based and high throughput sequencing-based virus discovery efforts. With the rapid advances in massively parallel genomic platforms, there is a clear demand for more informed and targeted molecular and bioinformatic strategies of detection.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vii
TABLE OF CONTENTS	ix
LIST OF TABLES	xi
LIST OF FIGURES	xiii

CHAPTER 1: INTRODUCTION	1
1.1. <i>Viruses and disease</i>	5
1.2. <i>Traditional methods of viral discovery</i>	6
1.3. <i>Pan-viral microarrays</i>	8
1.4. <i>Virus detection with high-throughput sequencing</i>	10
1.5. <i>Viral metagenomics</i>	13
1.6. <i>Viral evolution</i>	14

CHAPTER 2: TARGETED VIRAL DISCOVERY OF HUMAN HERPESVIRUSES: AIDS SALIVA AND TONSILLITIS	18
2.1. <i>Introduction</i>	18
2.2. <i>Background</i>	20
2.3. <i>Methods</i>	22
2.4. <i>Results</i>	27
2.5. <i>Discussion</i>	31
2.6. <i>Acknowledgements</i>	37

CHAPTER 3: VIRAL METAGENOMICS IN DISEASES WITH UNKNOWN ETIOLOGY: UVEITIS AND ACUTE LIVER FAILURE	46
3.1. <i>Introduction</i>	46
3.2. <i>Background</i>	47
3.3. <i>Methods</i>	49
3.4. <i>Results</i>	52
3.5. <i>Discussion</i>	55
3.6. <i>Acknowledgements</i>	61

CHAPTER 4. VIRAL INFECTION IN ACUTE EXACERBATION OF IDIOPATHIC PULMONARY FIBROSIS.....	70
4.1. <i>Introduction</i>	70
4.2. <i>Methods</i>	72
4.3. <i>Results</i>	76
4.4. <i>Discussion</i>	80
4.5. <i>Supplemental</i>	85
4.6. <i>Accession numbers</i>	85
4.5. <i>Acknowledgements</i>	86
CHAPTER 5. VIROGAP: USING VIRUS-HOST COPHYLOGENY TO IDENTIFY TARGETS FOR VIRUS DISCOVERY.....	90
5.1. <i>Introduction</i>	90
5.2. <i>Background</i>	92
5.3. <i>Methods</i>	95
5.4. <i>Results</i>	105
5.5. <i>Discussion and future directions</i>	110
5.6. <i>Acknowledgements</i>	115
CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS.....	131
REFERENCES.....	135
PUBLISHING AGREEMENT.....	146

LIST OF TABLES

CHAPTER 2

Table 2.1. AIDS Saliva samples for deep sequencing.....	43
Table 2.2. Herpesvirus detection in AIDS saliva.....	43
Table 2.3. Array-motivated detection in AIDS saliva.....	44
Table 2.4. Viruses in AS219 deep sequencing run.....	44
Table 2.5. TBLASTx hits against <i>nt</i> for 213-nt read.....	45
Table 2.6. Viruses in D31589.....	45

CHAPTER 3

Table 3.1. Uveitis samples deep sequencing using 454.....	68
Table 3.2. Viruses, bacteria, and eukaryotes identified by deep sequencing	68
Table 3.3. Acute liver failure samples sequenced on Illumina GAIIX.....	69

CHAPTER 4

Table 4.1. Clinical characteristics.....	88
Table 4.2. Respiratory viral detection in acute exacerbation and stable idiopathic pulmonary fibrosis.....	88
Table 4.3. Array-based viral detection in acute exacerbation and stable idiopathic pulmonary fibrosis.....	89
Table 4.4. Acute exacerbation of IPF samples selected for deep sequencing.....	89

CHAPTER 5

Table 5.1. Viral lifestyles.....	130
---	-----

LIST OF FIGURES

CHAPTER 2

Figure 2.1. Iterative pipeline for sequence analysis for 454 read analysis.....	38
Figure 2.2. Clustogram of Virochip oligos for tonsillitis sample.....	39
Figure 2.3. Percentage of 454 reads by taxonomic domain.....	40
Figure 2.4. Illumina reads per barcode at each iteration of read analysis.....	41
Figure 2.5 Taxonomic domains of 454 reads for tonsillitis sample.....	42

CHAPTER 3

Figure 3.1. Diagram of the eye.....	62
Figure 3.2. Taxonomic domain division of 454 reads using BLASTn and TBLASTx from uveitis pool sequencing run.....	63
Figure 3.3. Multiple sequence alignment of assembled <i>Asfarivirus</i> contig.....	64
Figure 3.4. Cluster diagram of microarray results showing B19 parvovirus.....	65
Figure 3.5. HTS reads from FH563 mapped to the hepatitis C virus genome.....	66
Figure 3.6. HTS reads mapped to GB virus C genome.....	67

CHAPTER 4

Figure 4.1. Survival time in torque teno virus (TTV)-positive acute exacerbation of patients with idiopathic pulmonary fibrosis compared with TTV-negative acute exacerbation of patients with idiopathic pulmonary fibrosis.....	62
--	----

CHAPTER 5

Figure 5.1. Virogap schema.....	116
Figure 5.2. Virogap pipeline.....	117
Figure 5.3. Clustering algorithm.....	118
Figure 5.4. Collapsing isolates using iterative clustering.....	119
Figure 5.5. The universality of cytochrome B.....	120
Figure 5.6. dsDNA virus species with host annotations.....	121
Figure 5.7a. All-against-all MCL clustering of all herpesvirus protein sequence ($I=1.0$).....	122
Figure 5.7b. All-against-all MCL clustering of all herpesvirus protein sequence ($I=4.0$).....	123
Figure 5.7c. Host diversity for three largest protein clusters.....	124
Figure 5.8. All-against-all sequence similarity network of herpesvirus DNA polymerase.....	125
Figure 5.9. Sequence similarity networks for rhadinovirus proteins.....	126
Figure 5.10. Sequence similarity network of rhadinovirus DNA polymerase showing partially sequenced proteins.....	127
Figure 5.10. Sequence similarity network of rhadinovirus DNA polymerase showing partially sequenced proteins.....	127

Figure 5.11. Host cytochrome b network with rhadinovirus DNA polymerase network.....	128
Figure 5.12. Top 20 unmatched hosts.....	129

CHAPTER 1. INTRODUCTION

Viruses have been implicated in the pathogenesis of many forms of disease, where some associations are clear cause-and-effect and others are more involuted. In recent years, the discovery of novel viruses have been associated with diseases with unknown etiology or not thought to be caused by an infectious agent. Viruses have been implicated in the pathogenesis of a number of cancers, and vaccines have been developed as a method to prevent disease.

Limitations of traditional methods of detection and characterization have defined the bounds of virus discovery. Most microbes have been difficult to culture, and immunological assays depend on the availability of antigenic and serological cross-reactivity; polymerase chain reaction (PCR) is a culture independent technique that has increased the sensitivity of detection but has its limits in diseases with no known etiology or with multiple viruses present [1]. Emerging technologies, however, have opened the door for viral metagenomics. Chip-based assays, for example, allow culture-independent detection of viruses from environmental or biological specimens without *a priori* knowledge of the viruses present [1]. Moreover, high-throughput sequencing provides an additional degree of freedom by producing sequence without necessity of hybridization between viral nucleic acid and probes derived from known sequence. With the rise of these new technologies, pan-viral analyses can be performed on inherently complex biological or ecological samples to survey for all viruses present; this is important for

answering questions on virus diversity, virus population dynamics, and can even identify viral etiology in a disease not thought to be caused by an infectious agent.

With the absence of universally homologous sequence shared between viral families and the broad sequence diversity within viral families, virus discovery is a nontrivial task.

Due to the anthropomorphic nature of virology, virus discovery in humans has progressed to the point where many of the easily discernible viral infections have been identified.

With microarrays and emerging sequencing efforts, two main approaches to virus discovery have been implemented--surveying viruses involved in a disease with unknown etiology, and surveying the viral metagenomic flora of a high-yield microbiome. This thesis presents viral discovery projects that apply both of these principles. Additionally, this dissertation describes a computational method of generating hypotheses to motivate discovery projects using principles of virus-host cophylogeny.

Chapter 2 describes a project designed to identify the ninth human herpesvirus, a clear gap representing a putative virus in the herpesvirus phylogenetic tree when overlaid with a phylogenetic tree of primate hosts. *A priori* knowledge describing this gap was used to facilitate detection efforts, from the sample selection to the technologies used to probe these samples. More specifically, saliva from individuals with full-blown AIDS in a time before antiretroviral therapies were used to demonstrate the high-yield approach to discovery, as herpesviruses are known to both replicate in human saliva and reactivate when the host is in an immunocompromised state. In the second part of this chapter,

tonsillar cell lines from young adults experiencing recurrent tonsillitis were treated with a chemical known to induce herpesvirus lytic replication and investigated for known and divergent herpesviruses. Tailored technologies were applied to both collections, including a herpesvirus-tiling microarray and application of a gap specific Hidden Markov model to high-throughput sequencing data.

Chapter 3 is an application of the former approach to virus discovery, where two devastating diseases with unknown etiology are investigated using viral microarrays and high-throughput sequencing. Intra-ocular fluid from individuals with idiopathic uveitis was collected to study viral involvement in this disease characterized by inflammation of the eye and temporary to permanent blindness. The rich microbial flora of the eye that was taxonomically binned through high-throughput sequencing revealed the metagenomic aspect of this study. The second disease, acute liver failure is a disease with approximately one third of cases of indeterminate etiology, and serum from individuals with acute liver failure testing negative for the common causes and hepatitis viruses were investigated by microarray and high-throughput sequencing. These studies yielded a small handful of unexpected viruses, including a novel circovirus and a highly divergent asfarivirus in individual cases of uveitis, and a number of known viruses including human parvovirus B19 in the acute liver failure collection.

Chapter 4 describes a study on viral etiologies in acute exacerbation of idiopathic pulmonary fibrosis, where our viral discovery platforms are applied to scrutinize

microbial flora of an interstitial lung disease punctuated by unexplained drastic declines in lung function that often lead to death within days to weeks. A handful of expected respiratory viruses and herpesviruses were detected in the acute exacerbation of idiopathic pulmonary fibrosis samples and zero of the controls, although this number was not significant. A small single-stranded DNA virus, TTV, was discovered in almost one third of the acute exacerbations and none of the cases of stable idiopathic pulmonary fibrosis. Although this finding was a striking dissimilarity between the acute exacerbation and stable idiopathic pulmonary fibrosis groups, TTV was also detected in a large minority of acute lung injury controls, suggesting the most feasible explanation being inflammation-induced active viral shedding in the lungs.

Finally, in Chapter 5, I describe an abstraction of the method used in Chapter 2 to hypothesize the phylogenetic bearings of a novel human herpesvirus, where I overlay the sequence similarity networks of host species on viral sequence similarity networks to identify gaps in regions with the highest degree of cophylogeny. These gaps then represent candidate discovery targets, where adjacent viral nodes to the gap in the viral network can be used to describe various aspects of the putative virus, from the sequence to tissue tropism and viral lifestyle. These indicators can be used to influence all stages of the viral discovery pipeline--sample selection, sample processing, and sequence analysis.

1.1. VIRUSES AND DISEASE

It has been estimated that one-fifth of cancers are caused by a pathogen, of which most cases are viral [2]. Cancers with infectious etiologies have been targets for vaccine-based prevention strategies. Human papillomavirus (HPV) vaccination for prevention of the majority of cervical cancer cases and hepatitis B virus (HBV) vaccination to limit the incidence of primary liver cancer exemplify the best case incentives for identifying the role of viruses in diseases.

The discovery of Epstein-Barr virus (EBV) in Burkitt's lymphoma by electron microscopy in 1964 and subsequent experiments implicating EBV in tumor progression was the first attestation of a virus involved in oncogenesis. Since this discovery, it has also been shown that HPV, Human T-lymphotropic virus, HBV, hepatitis C virus, Kaposi's sarcoma-associated herpesvirus (KSHV), and most recently, Merkel cell polyomavirus also assist in tumorigenesis [2].

Virus and cancer association studies are often obscured by the fact that viruses are not usually solely responsible for the growth of cancer. With the exception of KSHV and HPV, it is even rarely the case that the virus is ubiquitously found in the tumor. For example, 99 percent of adults are infected with EBV, but Burkitt's lymphoma is associated with EBV primarily in areas endemic with malaria and AIDS, where infected individuals are also immunocompromised. The antiquated Koch's postulates, once a hallmark of virology, has been irrelevant in most virus-cancer association studies. It is not

applicable to EBV and Burkitt's lymphoma or other EBV-associated cancers such as nasopharyngeal carcinoma, where cancer generally occurs in conjunction with lifestyle factors. Culturing the virus *in vitro* is difficult, nor has it been used to reinfected into other organisms and reproduce tumors. Even with the discovery that EBV immortalizes primary B cells, it wasn't until recent years that EBV was finally named a carcinogen by the international cancer agency [2].

Two models describe the virus-cancer association. Viruses can assist in carcinogenesis by employing oncogenes that activate cellular growth control pathways or usurp cell regulatory machinery to replicate their own genomes during the lytic cycle. In another model, viruses cause inflammation in the chronically infected, which in turn promotes tumorigenesis through carcinogenic mutations. In both models, it is not necessarily the case that the virus directly leads to tumorigenesis, and this tempered dependence makes connecting viruses and cancer less straightforward than fulfilling a set of universal postulates.

1.2. TRADITIONAL METHODS OF VIRAL DISCOVERY

Early virus discovery techniques have limitations that are now even more pronounced with advances in genomic methodologies. Viral diagnostics and discovery have traditionally been performed by isolation of viruses *in vitro*, electron microscopy, serology, and antigen detection. Previously championed as the gold standard of viral detection, most viruses have not been amenable to *in vitro* culture, and each virus

requires a susceptible host cell, specific host conditions, and a marker for cytopathic effect [3]. Electron microscopy, although not necessitating a set of virus specific reagents, does require a high concentration of virus particles. Hepatitis A virus and rotavirus were detected with immune electron microscopy, although this method would be less sufficient with lower viral titers [3]. Serology is a routine diagnostic tool that is especially useful for identifying chronic infections such as HIV [4], but both acute and convalescent serum are needed. Antigen detection, which mostly involves direct fluorescent antibody staining (DFA) is commonly used for detection of respiratory viruses and viruses that are slow or difficult to grow in culture, but is more relevant diagnostically as it requires *a priori* knowledge of the specific antigen.

PCR has been the pillar of nucleic acid detection and discovery of viruses, as all viruses with known sequence can be detected independently of whether it can be isolated *in vitro*. Even PCR, however, has limitations, as primers must be designed with known viral sequence. Other nucleic acid techniques have been employed for viral discovery. KSHV was detected through a DNA subtraction technique called representational difference analysis (RDA), where DNA from the healthy tissue was subtracted from the Kaposi sarcoma DNA [5].

While most of these methods have merit in diagnostics and have been successful in identifying novel viruses, there still remains a great deal of guesswork about what viruses are present and none of these techniques can test for all known viruses simultaneously.

Microarrays are a nucleic acid based platform that can be used to detect all known, and with careful oligonucleotide design, novel viruses as well. High-throughput sequencing has taken immense strides for delivering massive amounts of data at cost per base orders lower than Sanger sequencing.

1.3. PAN-VIRAL MICROARRAYS

We use a DNA microarray-based platform, Virochip, originally constructed with the most highly conserved 70-mer oligonucleotides (oligos) across all viral families in Genbank. Known viruses are discernible with distinct array signatures that could be predicted by comparing hybridization patterns to theoretical oligo-virus energy profiles [1]; novel viruses, on the other hand, present intensity patterns that are not predetermined, but rather comprise an composite of oligonucleotides from related viruses that share the most sequence similarity. It has been used to rapidly determine the novel agent involved in an outbreak, as in the case of severe acute respiratory syndrome coronavirus (SARS) [1]. The chip was also used to discover a novel human cardiovirus when conserved sequence in the 5'-UTR and 2C gene of Theiler's murine encephalomyelitis was detected [6]. Additionally, the Virochip has successfully detected non-human viruses, and Kistler et al. found a divergent bornavirus in birds with proventricular dilation disease (PDD) [7].

The latest iteration of the Virochip (V5) has been updated with recently added viral genomes, and takes a taxonomically striated approach, where most oligos target ultra-conserved regions in viral families, but some oligos represent lower taxonomic groups

down to the individual species level. This remodel reinvents the chip as an improved diagnostic tool and provides the capability to distinguish between known and novel virus.

Coverage across the viral families is variable, so a herpesvirus tiling microarray, Herpchip, was created to provide oligo depth for herpesvirus specific studies. The DNA polymerase gene is an essential component to viral replication and is one of the most conserved genes of herpesviruses, and, for the most part, other families of large DNA viruses such as phycodnaviruses, ascoviruses, and iridoviruses. It has been target region for pan-herpesvirus PCR for diagnostics [8], and having one of the longest stretches of conserved sequence [9], is a sensible basis for microarray design. This array covers 149 herpesviruses, consisting of 5,532 70-mers.

Both publicly available and in-house analysis tools such as Cluster and E-predict have been applied to ascertain viral signature on the microarray. For Herpchip, an in-house analysis scheme was developed (unpublished) to analyze this array, called HERPredict. The greatest challenge to viral microarray analysis is discerning true viral signal from noise. Enzymatic treatment and virus purification techniques can be applied during the nucleic acid extraction to preferentially select for viruses, and generally, large scale projects have a control set of samples to subtract background.

1.4. VIRUS DETECTION USING HIGH-THROUGHPUT SEQUENCING

Even with the ability to detect all viruses in parallel, obtaining viral sequence is still necessary to confirm a viral microarray signature. This standard has proven difficult with extremely divergent viruses and high ratios of background to viral nucleic acid. Viral array signature can be used to motivate primer design for PCR using oligo sequence. Elution of hybridized cDNA from the array can be used to select for viral sequence for downstream sequencing, although this is dependent on large quantities of hybridized cDNA and an efficient elution protocol.

Earlier attempts at shotgun sequencing of randomly amplified cDNA were limited by throughput, with viral sequence often comprising only a small percentage of total nucleic acid, especially in metagenomic samples or samples with high host background. With the emergence of high-throughput sequencing platforms that can produce up to one billion reads per run and its subsequent decreasing cost per base, all possible viral sequence can be considered in an unbiased manner. Additionally, with the increase in read lengths, the ability to sequence paired-end libraries, and the implementation of paired-end metagenomic assemblers [10], it is now possible to detect drastically divergent viruses at extremely low copy number.

A range of large-scale sequencing technologies are available and becoming increasingly more affordable per base sequenced. The main approaches of high-throughput sequencing include sequencing-by-synthesis and sequencing-by-ligation. All high-throughput

sequencing requires a library prep, usually involving fragmentation of nucleic acid, followed by attachment of platform specific sequencing adaptors onto templates, and subsequent amplification.

This thesis involves the application of two commercially available sequencers that both use sequencing-by-synthesis--the Roche 454 and the Illumina sequencing platforms. 454 was one of the first pyrosequencers commercially available, and while the number of gigabases delivered per full run is now one of the least in its class, it still had the advantage of longer average read lengths of approximately 330-bp [11]. Libraries are prepared through ligation of sequencing adaptors onto DNA templates, and emulsion PCR is performed on DNA molecules that have been isolated to a 1:1 ratio on beads. Template beads are deposited onto a picotiter plate, a microfluidic device with millions of wells designed to capture individual beads. Single dNTPs and sequencing reagents are flowed over the picotiter plate, where the wells are supplied with sulphurylase and luciferase, and light from the wells is imaged indicating incorporation of dNTPs undergoing pyrosequencing.

Illumina has taken the majority of the next-generation sequencing market share, with the ability to produce up to 600 gigabases in one run. Libraries of fragmented DNA or mate-pairs with the Illumina adaptors on each end of the template are bridge amplified into clusters on the Illumina flow cell, and incorporation of a reversible terminator is detected using total internal reflection fluorescence with lasers. The ability to perform paired-end

sequencing with a wide range of insert sizes has advanced *de novo* assembly and variant calling efforts.

The high volume of sequence generated by high-throughput sequencing technologies provides a sensitive method of combing a sample for any sign of viral infection. Only one molecule of DNA is required to detect a virus, and bioinformatic tools can be applied to assemble an entire viral genome *de novo*. Merkel cell carcinoma is a rare form of skin cancer that is more commonly found in immunocompromised human hosts [12]. In 2007, it was discovered by Yuan Chang and Patrick Moore, the husband-wife team who isolated KSHV from Kaposi's sarcoma in 1994, that a polyomavirus was the causative agent of this aggressive form of skin cancer. mRNA from tumors were pyrosequenced on the 454, and only two reads aligned to homologous polyomaviruses. Viral genome walking was used to obtain the rest of the Merkel cell polyomavirus genome.

High-throughput sequencing has also provided a quick solution to divergent viral genome finishing. Conserved, degenerate PCR primers targeting the 5' UTR and VP1 of picornaviruses was used to detect a novel enterovirus in a case of acute respiratory illness in a Nicaraguan child. Extracted RNA from the nose and throat swabs was prepped for high-throughput sequencing on the Illumina GAIIx, and the full-length genome was sequenced [13].

1.5. VIRAL METAGENOMICS

High-throughput sequencing has also been integral to progress in viral metagenomics, which had lagged in the metagenomic surge due to the lack of universally conserved sequence across viral families and disparate mode of evolution when compared to the rest of the tree of life. Early metagenomic studies were performed on microbiomes using targeted PCR and sequencing of 16s and 18s rRNA sequence, and focuses primarily on bacterial genomics [14]. Viral metagenomics first appeared in the form of environmental metagenomics, when viruses of seawater were shotgun sequenced using Sanger sequencing. Here, it is estimated that about 3×10^6 viruses per milliliter in the deep sea exist, and a diverse range of viruses was discovered [15]. In accord with the availability of high-throughput sequencing, there have been numerous studies on the human microbiomes, including the human gut and respiratory viromes [16]. The National Institute of Health's Common Fund established the Human Microbiome project, an initiative to characterize the microbiome of various locales across the human body [16,17]. One of the goals is to describe the core microbiome at each of the 18 sampled sites, and enable studies comparing microbiomes of individuals with a specific disease to controls [18].

The majority of viral families are believed to be known and projected to remain unchanged, given the low frequency of viruses being discovered that do not already fall into general existing clades [14]. This estimate, however, may overestimate headway in the field, as highly divergent viruses are difficult to detect by sequence-dependent

informatic methods. Moreover, within these defined families, there is presumed to be an undetermined number of viruses yet to be discovered, as little has been done to study viruses not hosted only by humans, domestic animals, and plants [14]. There is large disparity between the distribution of viruses and hosts known. Approximately 59 percent of known viruses are vertebrate and insect viruses, 25 percent are plant viruses, 15 percent are bacteriophages, and less than 1 percent are fungal viruses. In eukaryotic species, however, 79 percent are invertebrate, 18 percent plant, and only 3 percent are vertebrate [19]. Because of the anthropocentric nature of virus discovery, few have attempted to study virology in its natural state with viruses of lower plants, wild vertebrates, or non-arthropod invertebrates. Virus discovery in uncharted habitats contributes to better comprehension of virus interactions, diversity, and distribution.

1.6. VIRAL EVOLUTION

The Tree of Life comprises three divergent domains—Bacteria, Archaea, and Eukarya. Viruses, for the most part, have been omitted from the Tree of Life, as they possess few homologous traits linking them to other organisms and even to each other. More importantly, their model for evolution is clearly not monophyletic, and virus categories such as small fast-evolving RNA viruses and the large double stranded DNA viruses are thought to have arisen independently [14,20]. While attempts to infer a common virus origin have been untenable [21], there have been credible estimates of evolutionary relationships at the family levels or even at the superfamily level of viruses [20].

While virus evolution is relatively anomalous, there has been a much needed push to organize viruses into taxonomic categories. In early virus taxonomy, host range, nature of disease, and mode of transmission drove the classification scheme [22]. Following the advancement of molecular and biochemical techniques in the 1970s, genetic material as well as virion properties and serological specificity were included in the evaluation; biologists tried to use all available virus characteristics to classify viruses, assigning arbitrary significance to each feature. In the following decades, there was a shift to the use of sequence information as the primary basis for inferring evolutionary relationships [14,23]. The International Committee on Taxonomy of Viruses (ICTV) currently determines taxonomy and nomenclature of viruses by moderating a community of experts covering the 1915 species of viruses.

With the many modes of evolution between viral families, there are variable ceilings of sequence diversity. Some of the fast-evolving viruses, such as poliovirus and hepatitis C virus, exist more as a cloud of quasispecies. Mutation frequencies within these viruses margin on the line of the maximum tolerated for a viable virus, and their rates of evolution transition between periods of acceleration and periods of equilibrium[24].

Other viruses are able switch hosts by mutations. Canine parvovirus 2, for example, is a small single-stranded DNA virus presumed to have originated from feline parvovirus 2 through a mutation in the capsid gene [3]. Most emerging human diseases are zoonotic transfers, and some RNA viruses vary rapidly enough, from mutations or modular genetic variations, that their outset results in a disease with newfound virulence [24]. Influenza

virus is a classic example of a virus with a high rate of evolution associated with emerging disease. In addition to single mutations that can produce highly virulent strains, the virus is also known for its genetic reassortment of its eight-segmented genome [24]. When reassortment involves the hemagglutinin gene, assumed to occur in mixing vessels such as farmed pigs, it can result in human pandemics.

DNA viruses tend to have lower mutation rates than RNA viruses when normalized across the size of the genome [24]. This is most likely a result of the lack of error correction in RNA synthesis. Coevolution between viruses and their hosts are generally more prevalent in viruses with lower mutation rates and narrow host ranges, and this constrain is one of the keystones of viral evolution. Virus host interactions occur at every stage of the viral life cycle, from the factors involved in tissue tropism to the mechanisms of evasion viruses use to bypass host defenses. More so than the distinction between DNA and RNA viruses, viruses that maintain a persistent lifestyle in their host, such as the case with the large, double-stranded DNA herpesviruses, as opposed to those that have an acute lifestyle, are more likely to exhibit coevolutionary trends. The classic example in virology is the release of a South American myxoma virus in the European rabbit population in Australia and the subsequent strains of attenuated viruses that emerged. When the poxvirus was first introduced, it caused acute, lethal disease that eliminated over 99 percent of infected rabbits. With a dwindling host population, selection for attenuated strains of the virus was applied, and the presiding strains were those that killed less than 90 percent of infected rabbits and allowed for longer survival

times [25]. In turn, the selective pressures applied from virus to host involved the favoring of rabbits that could confer resistance to these virulent strains. Since this classic virology example of coevolution, the ability to create viral phylogenies using viral sequence have prompted the emergence of isolated studies of cophylogeny between viruses and their hosts.

In summary, this dissertation describes a diverse set of studies taking two parallel approaches to virus detection and discovery--targeted and syndromic--using pan-viral microarrays and high-throughput sequencing. A method of recognizing gaps in viral lineages using virus-host coevolution is also presented to precipitate future targeted virus discovery endeavors.

CHAPTER 2. TARGETED VIRAL DISCOVERY OF HUMAN HERPESVIRUSES: AIDS SALIVA AND TONSILLITIS

2.1. INTRODUCTION

The two most rudimentary methods of searching for novel viruses are the syndromic approach, where specific diseases with unknown etiology are targeted, and the practice of screening high-yield populations, such as individuals with severely compromised immune systems. This latter strategy can result in the recovery of a novel viral genome, which can subsequently be used to screen samples of a disease with unknown etiology.

The AIDS saliva study is an application of the latter approach, where we screen saliva collected and preserved from AIDS patients visiting a UCSF dental clinic in the early 1980's. As in the case of the tonsillitis study, this project was initially motivated by the lack of a homologous herpesvirus for humans in one of the two gamma-2 clades of the gammaherpesvirus subfamily. Although all samples were screened for all viruses, known and divergent, the most desirable outcome would be to discover the ninth human herpesvirus.

72 saliva samples were extracted for total nucleic acid, randomly amplified, and cDNA was hybridized to the Virochip. Array calls were followed up with confirmatory PCR and RT-PCR, and a handful of samples that were array positive for a virus but PCR negative were selected for deep sequencing on the Roche 454 and Illumina GAIIx.

The tonsillitis project sits at a mid-point between the two approaches, where a disease is chosen based on its portion of cases with unknown etiology and precedent for being associated with large, double-stranded DNA viruses (herpesviruses, adenoviruses), but additional sample processing was added to induce latent viruses into a replicating state.

The main goal of this project was again the discovery of a novel human herpesvirus, and here, tonsils were harvested from children and adolescents with recurrent tonsillitis, cultured, and treated with a phorbol ester to trigger expression of a lytic cascade of genes in herpesviruses. Cells and cell culture supernatant were split off before induction, 24 hours, and 48 hours after induction, and total nucleic acid was randomly amplified and hybridized to both the Virochip and the Herpchip. 14 different lines of tonsil cells were used, and 5 were selected for deep sequencing on the Roche 454 and Illumina GAIIx based on their increase in microarray signal for gamma-2 herpesvirus oligonucleotides on the induced samples when compared to untreated samples.

A subset of the tonsil samples showed a marked increase in gamma-2 herpesvirus expression by microarray after chemical induction, but in spite of selection of a lymphocyte rich specimen, favorable conditions for viral replication, and the most advanced tools for detection, no novel human herpesvirus sequence was detected.

2.2. BACKGROUND

Herpesviruses--latency and reactivation

Herpesviruses are commonly shed in the saliva, and the individuals who donated these samples were vulnerable to opportunistic pathogens and reactivation of persistent viruses as this was an era before anti-retroviral therapy [26]. Herpesviruses are persistent viruses, many of which are seroprevalent in the majority of the population. Some of the herpesviruses such as human cytomegalovirus (CMV) and Epstein-Barr virus (EBV) do not cause clinically obvious disease in healthy individuals. Through viral latency, herpesviruses maintain their lifelong persistence by residing in cellular sites within the individual, with minimal gene expression to remain undetected by the immune system [27,28]. When the virus reactivates, a cascade of genes are expressed for replication. These events normally occur in a small percentage of cells, and this process is kept in check by the cell-mediated immune system. In immunocompromised hosts such as the donors of these specimens, however, viral replication is uncontrolled and various clinical disease can manifest. Shedding of CMV and EBV is known to be increased in saliva of immunocompromised hosts over immunocompetent hosts [26]. The association between shedding of human herpesvirus 6 (HHV6) and human herpesvirus 7 (HHV7) and HIV positivity has been inconclusive [26].

Induction of lytic cycle

Since the majority of genes would be overlooked by genomic methods of detection in the latent stage, chances of discovering a novel herpesvirus would be greater in the lytic

phase when the virus engages in its cascade of gene expression for viral replication, eventually causing cell death. It has been shown that the lytic phase of select herpesviruses can be induced with the addition of a phorbol esters or sodium butyrate [29]. The processes promoting lytic replication are still unclear, but it is known that overexpression of ORF50 and hypoxia can also induce lytic replication [30].

Tonsillitis

Tonsillitis is an inflammation of the lymph nodes at the back of the throat and is often characterized by symptoms of sore throat, fever, and difficulty swallowing. Tonsils are thought to play a role in filtering foreign particles and activating an immune response, but when overloaded with infection, can become swollen and tender [31]. Tonsillitis can be a resultant of a bacterial (group A streptococcus) or viral (EBV, adenovirus) infection. Gammaherpesviruses have a tropism for lymphocytes, and the lymphoid tissue in tonsils are a natural environment for herpesviruses to reside.

Tonsillectomies

Tonsillectomies, although declining in rate in the United States, are one of the most common surgical procedure on children [32]. In the past 4 to 5 decades, there has been much debate about the role of surgery for tonsillitis, and guidelines of the American Academy of Otolaryngology-Head and Neck Surgery list now state that a recurrence of three or more infections of the tonsils and/or adenoids per year indicate a need for

tonsillectomy or adenotonsillectomy [32]. The tonsils collected in this study are all cases of recurrent tonsillitis.

2.3. METHODS

Sample collection of AIDS saliva

Samples were collected at the UCSF Dental Clinic in the early 1980's. Patients who were HIV+ with full-blown AIDS were asked to rinse their mouths prior to collection and to drool into a 2 milliliter screw cap tube. These saliva samples were flash frozen and stored in a -80C freezer until further processing.

Sample collection and culture of tonsils

14 unique sets of tonsils were obtained via the Cooperative Human Tissue Network during routine tonsillectomies at Vanderbilt University Medical Center. Tonsils were processed for human lymphoid aggregate cultures as previously detailed [33]. The primary tonsil cell lines were split, and a portion was treated with phorbol-12-myristate-13-acetate (PMA) and the rest was cultured without PMA. Cells and supernatant from the PMA-treated cultures were collected after 24 hours and 48 hours.

Virochip analysis

Half the amount of saliva was processed, or at least 200 microliters if the original specimen was less than 400 microliters. Samples were freeze-thawed twice to lyse cells, treated with DNase and RNase, and extracted using the Qiagen QIAamp Ultrasens Virus

kit (Qiagen, Inc., Valencia, CA) according to the manufacturer's instructions. Both cells and supernatant of the tonsil cell cultures were also treated with DNase and RNase and extracted using the Ultrasens Virus kit.

Total nucleic acid was randomly amplified, labelled, and hybridized to the Virochip as previously described [1]. Hybridization patterns were analyzed for viral signal using E-predict and Cluster [34,35]. The tonsillitis arrays were further examined with a script to analyze the amount of increase in gammaherpesvirus hybridization signatures between mock treated and PMA-treated cells and supernatant after normalization. Tonsils that demonstrated the highest increase between untreated and treated gammaherpesvirus oligo signal were selected as samples of interest for further study.

Viral signature was followed up with confirmatory PCR using published primers targeting the microarray suggested virus or primers designed based on the sequence of the array oligonucleotides. PCR positives were sequence confirmed through shotgun sequencing on the ABI 3130.

Herpesvirus tiling array

In addition to the pan-viral approach, a herpesvirus-specific microarray was designed with 70-mer oligonucleotides tiled across the ultra-conserved DNA polymerase gene of every fully or partially sequence herpesvirus. When arrays suggested a new herpesvirus, the sample was analyzed through PCR, shotgun sequencing, and deep sequencing.

Deep sequencing

Three saliva samples were selected for deep sequencing based on their microarray signature for a gamma-2 herpesvirus and absence of PCR confirmation of the only known human gamma-2 herpesvirus--KSHV. One of these samples was negative for all human herpesviruses, and the other two were PCR positive for EBV, but PCR negative for KSHV (**Table 2.1**).

cDNA from the random amplification was prepared for sequencing on the Roche 454 according manufacturers instructions (Roche Diagnostics Corporation, Indianapolis, IN). Samples AS215 and AS219 were each loaded on one of the two regions of the microfluidic sequencing device and sequenced on the GS FLX instrument. Sample AS221 was one of 12 barcoded samples on a separate run of the GS FLX.

Five of the tonsils, selected based on their microarray signatures for a gammaherpesvirus, were prepared for deep sequencing using an in-house protocol as previously described [36]. The samples that had the clearest increase of gammaherpesvirus oligo signal between either the mock treated and 24 hour PMA-treated samples or mock treated and 48 hour PMA-treated samples were selected. Nucleic acid from the cell culture supernatant of both 24-hour and 48-hour PMA-treated tonsils was used for the library prep as preliminary sequencing of the cells revealed that greater than 99 percent of reads would have been host nucleic acid. The five samples were barcoded with a 4-bp barcode

and paired-end sequence was produced on one lane of the Illumina GAIIx with 87-bp read from each end. Additionally, one of these five samples, D31589, was sequenced on the Roche 454 as one of the 12 barcoded samples alongside saliva sample AS221.

Deep sequence analysis of AIDS Saliva

A pipeline for binning 454 reads using iterative BLAT and BLAST was implemented (**Figure 2.1**). A first pass image quality filtering was done on the GS FLX instrument, and reads that passed were filtered for low complexity sequence using a Lempel-Ziv-Welch (LZW) compression scheme [37]. Additionally, reads that had six or more consecutive *N*s were removed. Next, host sequence was removed by a BLAT to the human genome [38], followed by a more lenient BLASTn [39]. The non-human reads were then successively aligned to the NCBI *nt* database using first a MEGABLAST with a word size of 28-bp and an e-value of 10^{-9} , then a BLASTn with a word size of 16 bases and e-value of 10^{-7} , followed by a more permissive BLASTn with a word size of 7 bases and e-value of 10^{-5} . At the final step, a translated BLASTx against *nt* was performed using a word size of 3 amino acids and an e-value of 10^{-5} . At each iteration, reads that hit entries in the database were extracted from the FastA file, the top e-value hit was kept and used to sort the read into a bin based on the taxonomic identifier of the subject organism. Reads that did not hit anything in *nr/nt* were used as seeds for *de novo* assembly from the original dataset using the PRICE assembler [10], and assembled contigs were aligned again to *nt* using BLAST.

High-throughput sequencing analysis of tonsillitis samples

Paired-end Illumina reads were basecalled using Gerald from Illumina's basecalling pipeline. The 4-bp barcode was removed, and reads were separated into separate FastA files based on their original sample. Reads with no matching barcode were filtered into another file. Reads with six or more *Ns* were removed, and low complexity sequence was then filtered out using a LZW compression cutoff of 38 [37]. Only read-pairs with both reads passing the complexity filter were fed to the remainder of the pipeline. Reads were filtered for both the human genome and transcriptome using BLAT with default parameters. Reads that did not hit either database were used as queries for MEGABLAST against the human genome and transcriptome using a word size of 24 and e-value of 10^{-16} . The reads passing the human filters were then used as queries for MEGABLAST against the BLAST *nt* database with the default word size and e-value of 10^{-7} . MEGABLAST output was then condensed into a file, with only the most significant hit saved. These hits were trimmed from the original set of reads and remaining reads were again aligned using BLASTn to *nt* with a word size of 16 and e-value of 10^{-7} , and then once more with an e-value of 10^{-3} . The final set of reads were aligned to *nt* using TBLASTx with a word size of 3 and e-value of 10^{-2} . Hits at every iteration of BLASTn and TBLASTx were reduced to a non-redundant set of hits based on lowest e-value, and binned taxonomically. Reads that fell all the way through the TBLASTx were aligned to a database of all herpesvirus sequence. These non-aligned reads were also used as seeds for assembly off of the original set of reads using PRICE [10].

454 reads from the tonsillitis sample, D31589, sequenced along with 11 other barcoded samples were analyzed using the same pipeline used in the saliva sequence analysis (**Figure 2.1**).

2.4. RESULTS

Sample collection

Saliva samples ranged from 250 microliters to over 1 milliliter, with an average of approximately 750 microliters. Tonsillitis supernatant was frozen at -80C in a lysis buffer, with several milliliters saved for future processing.

PCR detection of viruses in AIDS Saliva

After hybridizing a total of 72 samples to the cDNA microarrays, viral signature for herpesviruses, common respiratory viruses, and other viruses were detected. Of these array positives, a subset was confirmed during followup PCR.

A panel of herpesvirus PCRs revealed a majority of salivas had evidence of herpesviruses (**Table 2.2**). More specifically, 43 samples were sequence confirmed for EBV. Of these samples, 18 had double infections with both EBV types 1 and 2. CMV was present in seven of the samples. Herpes simplex virus 1 and HHV-6 were detected in 5 samples each.

A wide variety of other viruses were also detected by array-motivated confirmatory PCR. Rhinovirus was detected in five samples, Molluscum contagiosum virus was detected in three samples, and TTV, although detected by array in 15 samples, was PCR positive in seven samples (**Table 2.3**). At least one of these TTVs that were detectable by array but not PCR was found during shotgun sequencing. The primers used were published primers (NG059, NG061, NG063) for the conserved N22 region [40].

Viral detection in tonsillitis

Two tonsillitis specimens exhibited clear herpesvirus signal on the Virochip--samples D31589 and D32452. After normalizing intensities, these herpesvirus oligos also showed a pronounced increase in intensity after treatment with PMA (**Figure 2.2**). The clustrogram of D31589 shows an 'off-on' pattern for many gammaherpesvirus and other herpesvirus oligos in both the cells and supernatant. A panel of herpesvirus PCRs were run on all treatment combinations of the 14 tonsils, and no virus was isolated from these samples.

High-throughput sequencing of AIDS saliva

The three saliva samples selected for their gamma-2 herpesvirus array signature were deep sequenced on the GS FLX. The average read length was between 200 and 230-bp. After low pass initial quality filtering on the instrument, 121,147 reads were generated for AS215, 167,578 reads for AS219, and 79,930 reads for the barcoded AS221. The latter two were EBV positive by both array and PCR, and the sequence analysis of these

454 reads confirmed this finding. The viruses found in AS219 are shown in **Table 2.4**, and the breakdown of reads in **Figure 2.3**. The majority of the reads were host DNA, and a minority of reads hit bacteria, fungi, and virus. Three percent of reads could not be binned or binned after assembled. This sample was selected for its array signature suggesting a gamma-2 herpesvirus, however, and after all steps of BLAST and assembly, no detectable gamma-2 herpesvirus was found. AS215 and AS221 also contained no detectable herpesvirus.

Divergent viruses in AIDS Saliva

While no novel human herpesviruses were found in this sequencing run, new viruses from families of greater within-host diversity were discovered. In AS221, seven reads hit various types of human papillomavirus (HPV) averaging around 40% amino acid identity to homologous papillomaviruses; in AS219, 64 reads were from a divergent torque teno virus (TTV); and one read was from another type of single-stranded DNA virus in the family Circoviridae (**Table 2.4**). All reads were used as seeds for assembly from the remaining unmapped reads and off of the human subtracted dataset, but given the small number of reads in the starting dataset, the coverage on the viruses was not enough to grow lengthier contigs.

The HPV reads mapped to the E1, L2, and L1 gene of various different genera of *Papillomaviridae*, hitting alphapapillomaviruses, betapapillomaviruses, and gammapapillomaviruses. The TTV and circovirus reads had a mere 54% and 47% amino

acid identity to their respective closest viruses. With such divergence, the e-values for the BLAST hits were high, but almost all hits beyond the top hit were also consistent with the top hit. In the case of the circovirus, the top hit at the time of discovery was the Finch circovirus with a e-value of 0.016, followed by other circoviruses that aligned at higher e-values, including several strains of Porcine circovirus (**Table 2.5**). Upon re-BLAST against a more recent *nt* database, the top hit is now cyclovirus NG12, a virus in the new genus Cyclovirus that has been isolated from human and chimpanzee stool [41].

Deep sequencing of tonsillitis

Nearly 25 million read pairs were generated from the lane of Illumina sequencing, with approximately the same number of reads between the five barcoded tonsillitis samples. The majority of the reads were filtered out when the human BLAT and MEGABLAST were applied, and after MEGABLAST, BLASTn, and TBLASTx to *nt*, only approximately 0.1 percent of the original dataset remained unattributed (**Figure 2.4**).

An example of the breakdown of reads is shown for the 454 run (**Figure 2.5**). 95 percent of reads were of human origin. Three percent and one percent of reads were filtered out because they were shorter than 30 bases or low complexity, respectively. Almost no reads were of bacterial or viral origin. After all steps of BLAST, only 2 percent of the original dataset remained unattributed.

In the 454 reads, the bacterial reads were from organisms known to occupy the oropharynx, including *Haemophilus influenzae*, *Treponema denticola*, and *Fusobacterium*, a common genus of bacteria that has been implicated in periodontal diseases. Only plant viruses were sequenced in this run.

The lane of Illumina sequence, however, yielded a small number of viral hits. All viruses only have five or less reads, and all were nearly 100% identical to their subject viruses (**Table 2.6**). Five reads were 100 percent identical to human parvovirus B19 and one read hit the 5' UTR of human echovirus 11. Five reads from this sample and a similar number of reads from other tonsillitis samples were between 96-100% identical to bovine viral diarrhea virus 1, a common contaminant in the nutrient serum used in cell culture [42].

2.5. DISCUSSION

Two sets of samples were used in pursuit of a novel human herpesvirus. Virus discovery, especially when targeting a specific large double-stranded DNA virus, involves chance and speculation, but the odds of finding a new virus can be augmented by knowing the sequence of neighboring viruses around the gap of interest and careful sample selection. In this study, the gap of interest is in the gamma-2 genus of *Herpesviridae*, and the saliva and tonsillitis specimens are known to both contain herpesviruses. The former set of samples takes the 'high-yield' approach, as herpesviruses and other viruses are known to be shed in saliva [43], with higher degrees of herpesvirus shedding in

immunocompromised hosts [26]. The latter takes advantage of the fact that the gamma-2 herpesviruses are lymphotropic viruses that commence a cascade of lytic gene expression when reactivated.

Saliva

Targeted viral detection of known viruses has been studied extensively in the saliva of HIV+ patients [26], but only with recent advances in technology has it been possible to characterize all known and somewhat divergent viruses in an unbiased manner. Prior to viral microarrays and deep sequencing, the scope of past studies was defined by the limitations of traditional methods of detection. The conserved 16s and 18s rRNA sequence in non-viral life forms have made the bacterial flora of saliva a point of interest [44]. Virochip provides a means to defining the viral flora of saliva without an assumption of which viruses for which to probe. With the emergence of high-throughput sequencing platforms, this task becomes even more comprehensive, as the power of the study is not as severely affected by the a low signal to noise ratio, a complication arising when hybridizing a complex metagenomic sample that includes mostly bacterial, fungal, and host nucleic acid to an array that probes only for viruses. This technology has led to inclusion of oral cavity metagenomics in the push to characterizing the human microbiome [17,45]. This set of saliva samples differs from current oral microbiome studies because it constitutes sampling at a time before the common use of anti-retroviral therapies, so patients who donated in this study were severely immunocompromised with full-blown AIDS. Thus, while the best possible outcome would be to discover a novel

human herpesvirus, this study also uses the latest genomic technologies and bioinformatic tools to examine all known viruses and all potentially novel viruses in saliva of immunocompromised hosts.

Although only sequences from the eight known human herpesviruses were detected, an assortment of other viruses, including unexpected and divergent viruses, was found. First, common respiratory viruses such as rhinovirus were detected, perhaps evidence of opportunistic infections on a weakened immune system. Second, there was an overrepresentation of some of the herpesviruses in saliva. Shedding of EBV, for example, was detected in the majority of the samples, compared to 18% detected in normal saliva [46]. Finally, the ability of microarrays and deep sequencing to detect divergent viruses was evident upon the recovery of novel HPV, TTV, and circovirus sequence.

HPV

Papillomaviruses (PVs) are circular double-stranded DNA viruses that are about 6.8kb to 8.4kb in length [47]. There are over 100 types of HPV and over 100 types of PVs found in a broad range of vertebrate hosts. The divergent HPV reads were disparate enough from the closest genotype to be considered a new type. Greater than a 10 percent nucleotide sequence difference is necessary in the E6, E7, and L1 gene to formalize a new genotype [48], and HPV reads in AS221 were at most 80 percent nucleotide identity to the nearest genotype. The nearest genotype was HPV type 116, which was identified in 2009 as a novel gammapapillomavirus with enough genomic dissimilarity to be

considered a new PV species [47]. There was not enough read coverage of the divergent HPV to assemble the genome, and because these reads were on separate contigs, it is unclear whether they originated from a single HPV infection or more. It is not uncommon to find HPV in the saliva, and rates of detection range from one to 60 percent of oral samples from healthy individuals [48]. Various studies have commented on the higher rates of HPV infection in HIV+ individuals, a product of differing socioeconomic status and behavior or a result of the immune system's inability to clear the infection ultimately leading to HPV persistence [49].

TTV

TTV is a small single-stranded DNA virus of unknown pathogenicity. It has been isolated from a number of diseased samples, including hepatic and pulmonary disease, but the role of the virus in the diseases' etiologies is unclear [50]. TTVs are part of a diverse group, possessing about 40 percent genome-wide heterogeneity. In addition to detecting TTV in serum, feces, and nasal secretions, TTV has been previously isolated from saliva [51]. The sequence variability of this TTV isolate to other genotypes was consistent with the current degree of divergence in the family.

Circoviruses

Circoviruses are also small, circular single-stranded DNA viruses that have only recently been detected in humans with advances in viral metagenomic technologies. Viruses in the *Circoviridae* family had only been found in pigs and birds, and have been

associated with a variety of diseases [41]. Porcine circovirus 2 have been linked to respiratory, enteric, and systemic diseases [52]. The reads found in the 454 run were only detectable during the final round of sensitive undirected translated BLAST, since the reads were too divergent to pick up with a reasonable e-value threshold from nucleotide BLAST. Since the virus was found in saliva, it is unclear whether sequence was a result of direct viral shedding from individual or a transient passenger of the oral flora. In this particular sample and in other samples of this set, tobamoviruses such as tomato mosaic virus was recovered appearing to originate from the leafy components of meals ingested prior to donation. The most similar circovirus by sequence was cyclovirus NG12 (80 percent nucleotide identity) of the genus *Cyclovirus*, and was one of several circoviruses isolated from human and chimpanzee feces [41]. An assortment of cycloviruses, although all but one of different species of those found in stool, was also isolated from the muscle of farm animals, calling for further studies to determine the natural host of the virus.

Rationale behind absence of a novel herpesvirus

The result of no detectable novel herpesvirus is not unexpected, but this study not only increased chances of finding a new herpesvirus by examining high-yield samples, but also used the most sensitive and comprehensive tools currently available. There are many explanations to why a novel herpesvirus was not found, the most obvious reason being that a ninth human herpesvirus does not exist. There is a possibility that there was speciation between the *RV1* clade and *RV2* clade of all other primate herpesviruses but not for human herpesvirus 8. This would be an instance of a virus ‘missing the boat,’

where speciation occurred with the hosts, but a virus failed to continue with one of the two bifurcated lineages. This seems unlikely given the high degree of coevolution between herpesviruses and their hosts, the persistent nature of the virus, and the lack of a required vector that the virus would rely on for person-to-person transmission. Another explanation is that the virus is not shed in saliva. While there was a high detection rate of EBV and CMV, there was no evidence of KSHV in these saliva samples. It is unclear if clinically occult open oral sores was used as an exclusion criterion during the collection of saliva. KSHV is the closest human herpesvirus to the gamma-2 herpesvirus gap of interest, and it is present in a smaller percentage of the general population. Another point is that the discovery of a new herpesvirus is heavily dependent upon the diversity and number of currently known herpesvirus sequence. The clade of interest has a limited number of primate herpesvirus sequence, and most of these sequences are from only one or two genes. All of these sequences were used in both Virochip and Herpchip design as well as the database construction against which deep sequencing reads were aligned. With more sequence being added with every new non-human herpesvirus discovered and with the feasibility of complete genome assembly using the increasingly available deep sequencing technologies, the gamma-2 herpesvirus profile will only get more robust over time.

Future directions

The discovery of a novel human herpesvirus may come coupled with other metagenomic samples or diseases with unknown etiology. AIDS saliva represent a rich source of viral

diversity, and the deep sequencing reads that did not align to anything in the *nt* database should be re-aligned again as more viruses are added to Genbank. Although no novel human herpesvirus was discovered in these samples, a few novel viruses were found and represent promise for pursuing targeted virus discovery when ‘gaps’ can be identified in viral-host cophylogenies.

2.6. ACKNOWLEDGEMENTS

I would like to thank Jinjong Myoung for his work on the tonsil cultures, Patrick Tang for his guidance and protocols on saliva extraction, James Graham Ruby for his Illumina library prep protocol and PRICE, and Joe DeRisi and Don Ganem for their guidance.

FIGURES

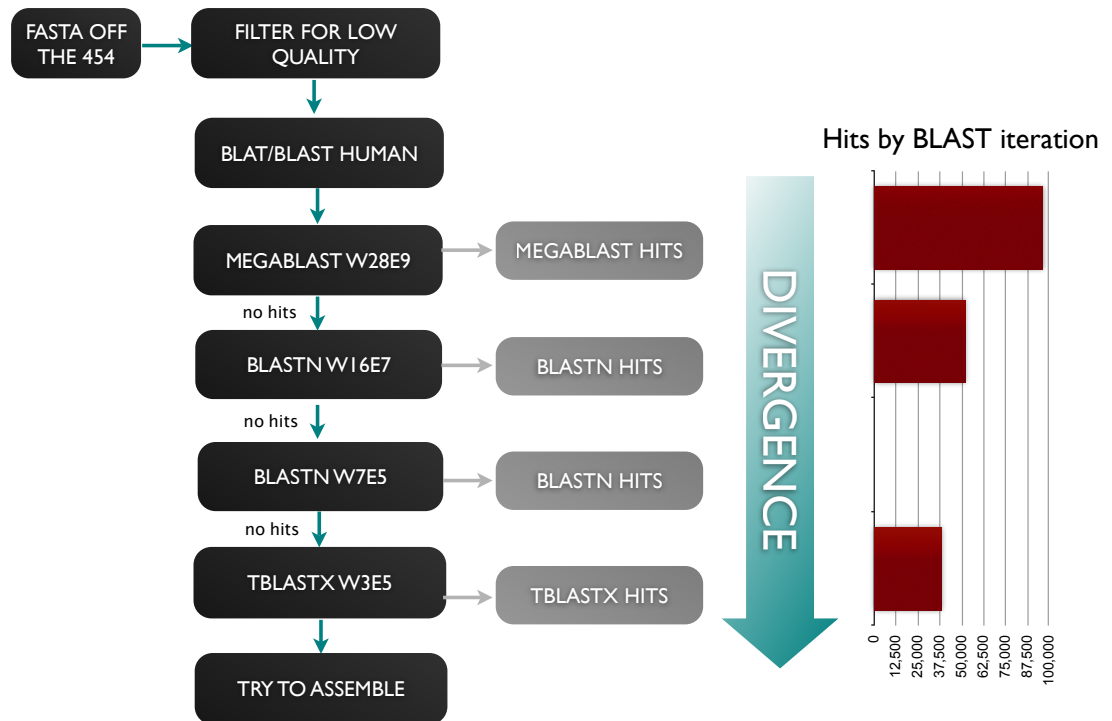


Figure 2.1. Iterative pipeline for sequence analysis for 454 read analysis. BLAST hits (grey) are peeled off at every iteration and binned taxonomically. The graph (right) shows the number of hits attributed at each step.



Figure 2.2. Clustrogram of Virochip oligos for sample D31589, showing an increase in herpesvirus oligo intensity between untreated cells and supernatant (column one and three) and PMA treated cells and supernatant (column two and four)

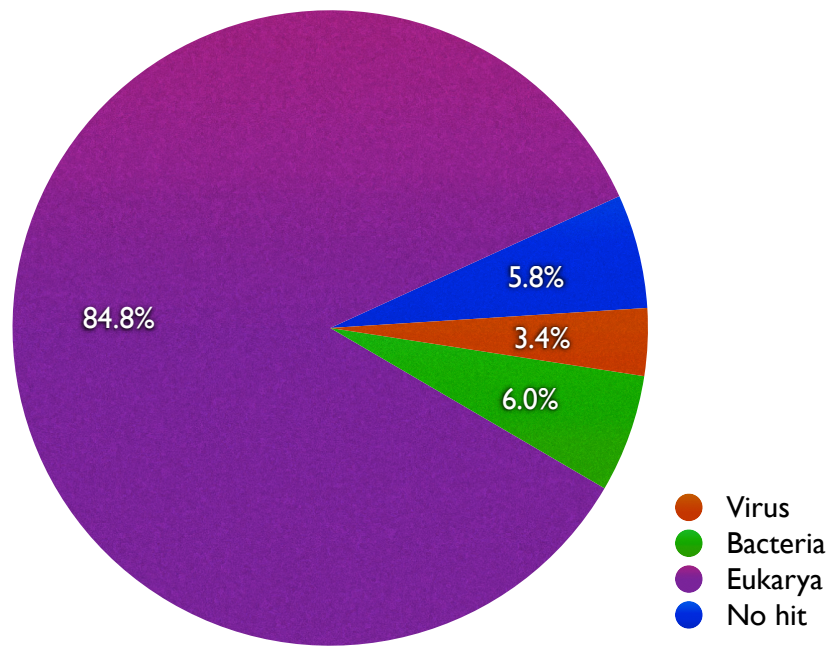


Figure 2.3. AS219 percentage of 454 reads by taxonomic domain. Reads that fell through the pipeline without aligning to any Genbank sequence with a significance below the threshold are in ‘no hits’ (blue).

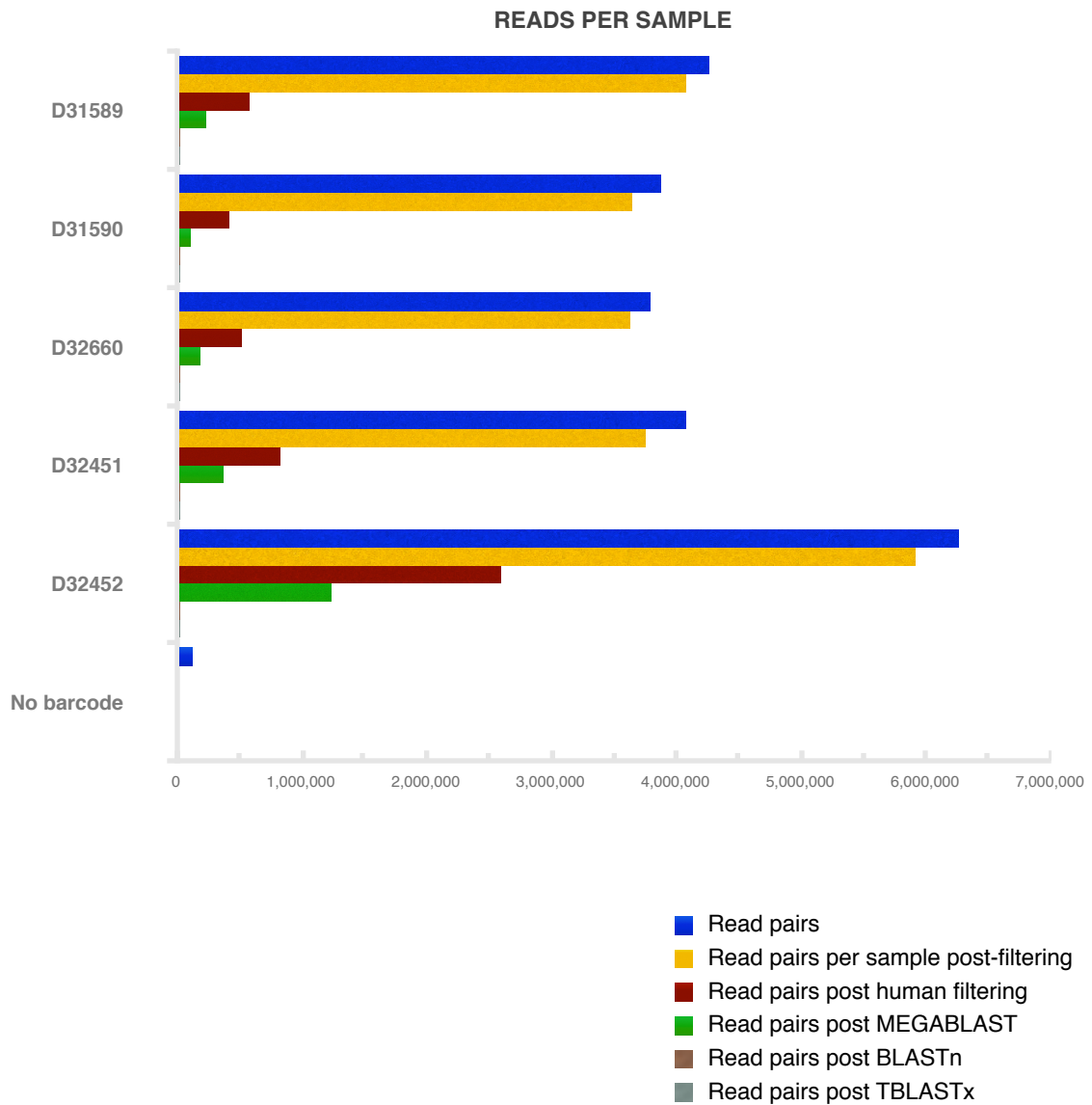


Figure 2.4. Illumina reads per barcode at each iteration of read analysis. Initial number of reads, and reads lefts per sample after each step of filtering and BLAST iteration.

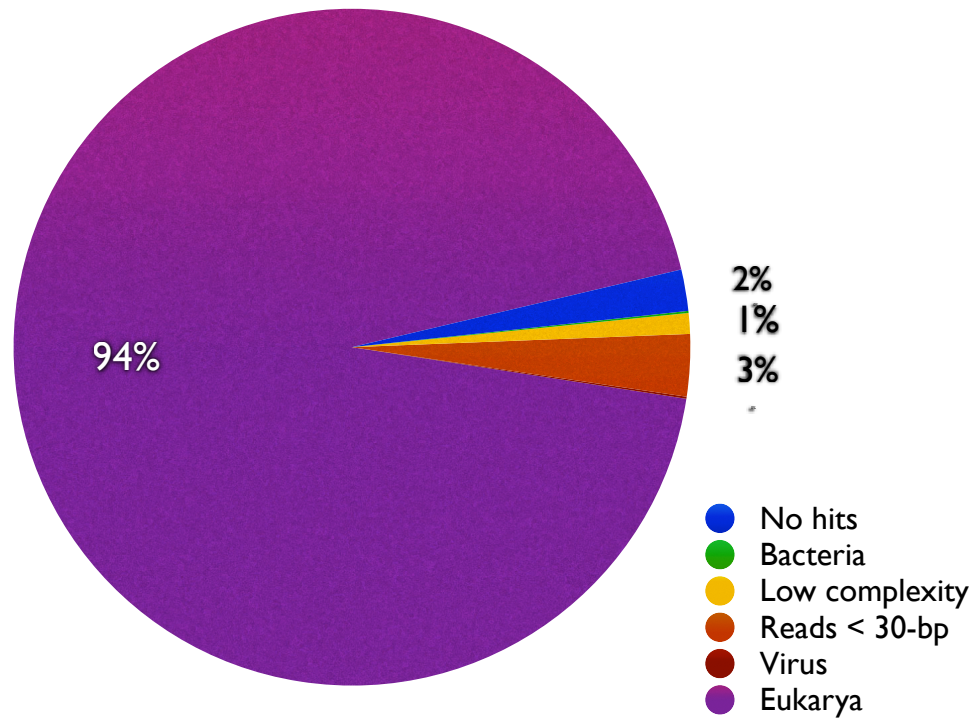


Figure 2.5. Taxonomic domains of 454 reads for D31589. Short reads, or those shorter than 30 nucleotides long were removed (red).

TABLES

TABLE 2.1: AIDS SALIVA SAMPLES FOR DEEP SEQUENCING

Accession	Sample ID	PCR findings	Array findings
AS87-3429	AS215	All herpesviruses (-)	CMV; Rhadinovirus
AS89-0240	AS219	EBV(+) TTV(+) KSHV(-)	Rhadinovirus, TTV
AS88-7790	AS221	EBV(+) KSHV(-)	Rhadinovirus

TABLE 2.2: HERPESVIRUS DETECTION IN AIDS SALIVA

Virus	PCR confirmed (n=72)
HHV-1: Herpes simplex virus 1	5 (7)
HHV-2: Herpes simplex virus 2	0 (0)
HHV-3: Varicella zoster virus	0 (0)
HHV-4: Epstein-Barr virus	43 (60)
HHV-5: Cytomegalovirus	7 (10)
HHV-6: Human herpesvirus 6	5 (7)
HHV-7: Human herpesvirus 7	0 (0)
HHV-8: Kaposi's sarcoma-associated herpesvirus	0 (0)

TABLE 2.3. ARRAY-MOTIVATED DETECTION IN AIDS SALIVA

Virus	PCR confirmed (n=72)
Rhinovirus	5 (7)
Adenovirus	0(0)
Molluscum contagiosum virus	3 (4)
TTV	7 (10)

TABLE 2.4. VIRUSES IN AS219 DEEP SEQUENCING RUN

Virus	Num. reads	Perc. identity
EBV	3214	95-100% n.t.
TTV isolate 3h	64	54-95% a.a.
HPV 17	7	100% n.t.
Human RV	1	100% n.t.
Circovirus	1	47% a.a.

TABLE 2.5. TBLASTX HITS AGAINST NT FOR 213 NT READ

Sequences producing significant alignments:	Score (Bits)	E Value	N
gb GQ404854.1 Cyclovirus NG12, complete genome	48.8	7.00E-04	1
gb DQ845075.1 Finch circovirus, complete genome	44.3	0.016	1
gb HQ839721.1 Circoviridae SFpork/USA/2010 isolate SFpork8 R...	43.4	0.031	1
gb HQ738638.1 Circoviridae PorkNW2/USA/2009 Rep protein gene...	43.4	0.031	1
gb GQ404851.1 Chimpanzee stool avian-like circovirus Chimp17...	43.4	0.031	1
gb JN377559.1 Bat circovirus ZS/Yunnan-China/2009 isolate 31...	42.9	0.042	1
ref XM_002612110.1 Branchiostoma floridae hypothetical prote...	42.9	0.042	1
gb JF938079.1 Bat circovirus ZS/China/2011 isolate YN-BtCV-2...	42.5	0.058	1
gb HQ839722.1 Circoviridae SFpork/USA/2010 isolate SFpork9 R...	42.5	0.058	1
ref XM_002598906.1 Branchiostoma floridae hypothetical prote...	41.6	0.11	1
gb EF524534.1 Porcine circovirus 2 strain HN0602, complete g...	41.6	0.11	1
gb JN377562.1 Bat circovirus ZS/Yunnan-China/2009 isolate 00...	41.1	0.15	1
gb EU148505.1 Porcine circovirus 2 isolate DK1990PMWSfree, c...	41.1	0.15	1
gb HM142896.1 Porcine circovirus 2 isolate BX, complete genome	40.6	0.2	1
gb FJ644561.1 Porcine circovirus 2 isolate HB05, complete ge...	40.6	0.2	1
gb EU503037.1 Porcine circovirus 2 strain Taizhou0512, compl...	40.6	0.2	1
gb EF524528.1 Porcine circovirus 2 strain HB05, complete genome	40.6	0.2	1
gb DQ923524.1 Porcine circovirus 2 isolate 15/23R from Brazi...	40.6	0.2	1
gb DQ923523.1 Porcine circovirus 2 isolate 15/5P from Brazil...	40.6	0.2	1
gb AY484415.1 Porcine circovirus 2 isolate NL_PMWS_3, comple...	40.6	0.2	1
gb AY321991.1 Porcine circovirus 2 strain Fd10, complete genome	40.6	0.2	1
gb AY321990.1 Porcine circovirus 2 strain Fd7, complete genome	40.6	0.2	1

TABLE 2.6: VIRUSES IN D31589

Virus	Number of reads	Average % identity
Human parvovirus B19	5	100
Human echovirus 11	1	98
Bovine viral diarrhea virus 1	5	97

CHAPTER 3. VIRAL METAGENOMICS IN DISEASES WITH UNKNOWN ETIOLOGY: UVEITIS AND ACUTE LIVER FAILURE

3.1 INTRODUCTION

An alternate approach to virus discovery is to take an unbiased survey of all viruses involved in a disease with unknown etiology. In these studies, instances of a novel or known virus in the set of diseased samples and controls is compared. While this does not define the type of relationship between the virus and disease, it does suggest an association that can be further studied for its epidemiology, pathology, and pathogenesis. The primary goal of these studies is the discovery of a divergent virus that could play a role in the etiology of the disease. A secondary goal of these studies is to enumerate all infectious agents in these samples, a goal that becomes particularly interesting in specimens of metagenomic origins.

Two very different diseases are used to exemplify this approach--uveitis, an inflammation of intra-ocular cavities of the eye, and acute liver failure (ALF), an uncommon but devastating disease with high morbidity. In uveitis, three types of bodily fluids are examined--anterior chamber fluid, vitreous fluid, and matched serum. This study represents the first time eye microbial flora has been described using deep sequencing. These fluids, originally thought to be sterile, revealed complex microbial diversity that made isolating a single agent in uveitis etiology difficult. Novel viruses, including a divergent circovirus and virus from the family *Asfarviridae* (family of African Swine

Fever viruses) were found, although variability of clinical definitions within the collected uveitis specimens made followup association studies plausible only with a more widespread and homogenous collection.

In the ALF study, 81 serum specimens from patients with fulminant hepatitis, a severe form of ALF were examined using microarray, PCR, and high-throughput sequencing (HTS). These samples were from a subset of ALF patients testing negative for hepatitis A-E virus. Human parvovirus B19 was detected in one sample by both Virochip and PCR. A subset of samples that showed signs of viral signature by microarray were prepared for HTS. From this lane of sequencing, GB virus C and hepatitis C virus were each detected in one of the 14 samples.

3.2 BACKGROUND

Uveitis

Uveitis is an inflammation of the middle region of the eye, which includes the choroid, iris, and pars plana (**Figure 3.1**). It is usually accompanied by either temporary or permanent blindness, and accounts for an estimated 2.8 to 10 percent of cases of blindness in the U.S [53]. As a umbrella disorder for an abundance of many diseases involving inflammation of the eye, uveitis presents in various clinical forms and has been associated with viral, bacterial and fungal infection, trauma, as well as a manifestation of autoimmune diseases such as sarcoidosis and rheumatoid arthritis. Between 14 and 48 percent of cases are idiopathic and have an undetermined cause [53].

Infectious causes of uveitis

Past studies on the role of infection in uveitis rely on ability to culture the organisms *in vitro*, and even in recent years with the inclusion of PCR as a diagnostic, a hypothesis of which infectious agents to test is required. Toxoplasmosis and tuberculosis represent about 6.5 and 10.5 percent (n=200) of cases of uveitis [54]. Herpes simplex virus (HSV), varicella zoster virus (VZV), and cytomegalovirus (CMV) are routinely tested for, with one study finding about 25 percent (n=119) cases of anterior uveitis being herpetic uveitis. Herpesviruses establish lifelong latency, and their reactivation and involvement in uveitis has been shown to vary with the individual's immunocompetence. Rubella has also been named as a potential agent involved in anterior uveitis, and no other viruses have been implicated in uveitis etiology. The high number of idiopathic uveitis and narrow range of viruses associated with uveitis has partially been limited by the lack of a pan-viral detection platform.

Acute liver failure

ALF appears in many clinical forms, but it normally presents with necrosis of hepatocytes. Most cases of ALF are either drug-induced or viral hepatitis, although it is estimated that less than 1 percent of viral hepatitis leads to acute liver failure [55]. It is estimated that 15 percent of adult cases and 50 percent of pediatric cases of ALF have no specified etiology [56]. Currently, the only viruses to be associated with acute liver failure are hepatitis A, B, D, and E and HSV.

3.3 METHODS

Collection of fluids from individuals with uveitis

Individuals with uveitis were recruited for the study upon detection of ocular inflammation when visiting Stanford University Ophthalmology, the Palo Alto VA, and Valley Medical Center of Ophthalmology. Aqueous humour was collected during anterior chamber paracentesis, and vitreous humour was collected if a vitrectomy was necessary as a component of treatment. Sterilization of all instruments, administration of antibiotic prophylaxis, and treatment of the surface of the eye with povidone-iodine occur prior to collection of ocular fluids. A total of 53 individuals contributed aqueous humour, and four individuals donated vitreous humour. Matched serum was also collected when possible, with nine serum samples matched to aqueous humour extraction, and two serum samples from patients who declined the anterior chamber paracentesis consent forms so no ocular fluids were collected. Of the 53 aqueous humour collections, 10 were control group patients with cataracts and no signs of inflammation. None of the individuals with vitrectomies were in the control group. All collections were stored at -80C until further processing.

Collection of acute liver failure samples

Serum from individuals experiencing acute liver failure was collected by Dr. Will Lee at the University of Texas Southwestern Medical School. 81 serum samples were processed for this study.

Microarray and PCR analysis

All specimens were treated with DNase and RNase prior to total nucleic acid extraction with the Qiagen QIAamp Ultrasens kit (Valencia, CA). Nucleic acid was randomly amplified using the standard Round A/B protocol described earlier. All uveitis samples were hybridized to the Virochip V4, and all ALF samples were analyzed on the Virochip V3. Arrays were analyzed using Cluster [35], and any viral signature was followed up with PCR.

Deep sequencing of uveitis

Five cases of uveitis were selected for deep sequencing on the Roche 454 GS FLX and another five were sequenced on the 454 FLX Titanium (**Table 3.1**). Samples were selected as candidates for HTS based on either array signature suggestive of a viral infection or based on the clinical diagnosis of the patient. Samples UV0019, UV0037, UV0039 were all selected for the viruses suggested by array, and UVSC07 was selected based on the donor's immunocompromised state as an AIDS patient with low CD4 counts. This individual was diagnosed with bilateral retinitis with diffuse necrosis.

Within uveitis being a spectrum of disorders, a subset of samples were selected for a deep sequencing run on the 454 Titanium based on their homologous clinical phenotype.

White dot syndromes are a group of diseases, usually in young adults, characterized by 100 to 200 micron white-yellow dots visible on the retina and fovea upon ocular examination [57]. In some cases, individuals diagnosed with white dot syndromes report

cough, fever, and malaise and then temporary or prolonged blurred or loss of vision. One common form of white dot syndrome was sequenced on the 454 GS FLX, a sample of aqueous humour from an individual with acute zonal occult outer retinopathy (AZOOR). In the pooled 454 Titanium run, UV0021 was included for its diagnosis as a class of white spot syndrome called acute posterior multifocal placoid pigment epitheliopathy (APMPEE), a disease with acute vision loss where approximately one third of individuals experience flu-like symptoms prior to vision loss [57]. UV0002 is a sample from an individual with serpiginous choroiditis, and UV0005 is a sample from patient diagnosed with birdshot choroiditis, both recurrent white dot syndromes that appear to affect middle aged adults. The final two samples, UV0033a and UV0041, are cases of chorioretinitis, which include vitreous inflammation and anterior uveitis. Samples were pooled at equal concentrations after amplification and size selection, and the pool was prepped for HTS using the standard Roche 454 DNA prep.

The FastA files were processed using the pipeline described in Chapter 2. In brief, short reads and reads of low complexity sequence were filtered out, and the remaining set of reads was sequentially passed through increasingly lenient iterations of BLAST to attribute the reads to records in NCBI. Of the reads that were assigned, a rank ordering of NCBI GIs was generated based on the number of reads mapping to them, with information of average percent identity, the taxonomic identifier of the source organism, and the record description included.

Deep sequencing of acute liver failure

14 ALF samples were chosen for HTS on the Illumina GAIIx based on interesting array signature. Barcoded, randomly amplified cDNA from each sample was prepared using an in-house developed library prep described in the Chapter 4. The pooled samples were run on one lane of a paired-end Illumina GAIIx flow cell with 65-bp read from each end. Reads were analyzed using a modified pipeline of the one detailed in Chapter 2.

3.4 RESULTS

Uveitis

The 454 GS FLX runs generated two sets of data--one where the pool was run on one of the two regions of the fluidic sequencing platform and another where the pool was run on both regions of the device. The former generated 109,785 reads with an average read length of 178, and the second run generated 187,812 reads with an average read length of 224 bases.

The majority of the reads were attributed to a specific NCBI record using a nucleotide BLAST. Most of these reads were of human origin (**Figure 3.2**), and there was a large minority of reads that were of fungal (18.1 percent) or bacterial origin (0.3 percent).

TBLASTX revealed mostly divergent bacteria and fungi, with a small percentage being of viral origin (4.7 percent).

A sampled list of the diverse spectrum of viruses, bacteria, and eukaryotes are enumerated in **Table 3.2**. Divergent strains of pathogenic strains of bacteria, including *Bartonella henselae*, *Rickettsia typhi*, *Bordetella pertussis*, and *Borrelia burgdorferi* were detected. 11 reads mapped with perfect identity to *T. gondii*, the parasite responsible for toxoplasmosis [54]. Sequence from a host of divergent viruses was identified, including a divergent human papillomavirus, a novel circovirus, and a virus most similar to African swine fever virus (ASFV). These divergent viral reads were found in the final round of attribution, during the translated BLAST to *nt*.

Reads were assembled into contigs for each virus. This contigs were used as seeds into the original dataset, but no further assembly could be furnished. Profile HMMs were created using sequence from the circovirus family and the ASFV family, *Asfariviridae*, but no additional reads were indicated by the model. The circovirus contig aligned most closely to beak and feather disease virus in the *rep* gene at 47 percent amino acid identity. The ASFV-like contig had 35 percent amino acid identity to strains of ASFV. The nucleotide alignment to nearest strains of ASFV is shown in **Figure 3.3**.

Acute liver failure

Microarray analysis using Cluster and E-Predict suggested viral signature for 15 samples. Through a clustering on both arrays and all viral oligos, human parvovirus B19 signal in sample FH539 drove the clustogram (**Figure 3.4**). The suspected virus was confirmed by PCR using primers and cycling conditions as previously published [58]. PCR product

was run on a 1.5 percent 1xTAE agarose gel, and bands of the expected size were cut and sequenced on the ABI 3730 sequencer (ElimBio, Hayward, CA). The remaining 14 were prepared for downstream sequencing analysis.

The 14 barcoded ALI samples were run on one lane of the Illumina GAIIx, and 27.4 million read pairs were generated for this lane. After filtering out low complexity sequence, reads were reconciled such that a pair was removed when one of the two was low complexity. 25.1 million read pairs remained after filtering. The breakdown of reads per sample is shown in **Table 3.3**, post-filter. 333,434 read pairs were then removed for lacking a perfect match to an expected barcode.

After binning all reads according to taxonomy, two of the samples had viral reads: one with reads to hepatitis C virus (HCV) and human herpesvirus 6, and the other with reads to GB virus C. All reads were near perfect matches to existing genomes in Genbank. Reads mapping to both viruses extend across the genome, but were not dense enough to assemble the entire genome. 162 reads from FH563 mapped across the entire HCV genome, which consists of one open reading frame (ORF) that codes for structural proteins on the N-terminal portion and non-structural proteins on the other end (**Figure 3.5**). Most N-terminal reads map to the start of the part of the ORF coding for the E2 protein, the envelope protein. Regions of the ORF coding for non-structural proteins are generally evenly covered by remaining reads.

Reads mapping to the GB virus C genome are relatively sparse but map across the length of the genome, with a paired-end read in the 5' UTR and coverage across the putative E1, p7-NS2, NS3, NS5a and NS5b regions (**Figure 3.6**).

3.5 DISCUSSION

In these studies, two diseases with idiopathic forms of the disease were probed for evidence of viral infection through the examination of samples collected at the time of illness using unbiased and high-throughput genomic technologies. Uveitis and acute liver failure are manifestations of different etiologies, including systemic disease, infection, trauma (uveitis), and drug overdose (ALF). This study analyzed the subset of individuals with idiopathic disease for novel viruses or viruses not previously realized to be associated with the disease.

Uveitis

Uveitis is a form of intra-ocular inflammation that occurs as an indication of a variety of diseases. Sarcoidosis, HLA-B27-associated spondylarthropathies, and juvenile rheumatoid arthritis are some of the common systemic diseases associated with uveitis [59]. Additionally, a number of pathogens have been affiliated with subtypes of uveitis. Tuberculosis, toxoplasmosis, and herpetic uveitis primarily from HSV, VZV and CMV have been known to account for over one-third of uveitis cases [54]. Alphaherpesviruses HSV and VZV have been known to play a role in uveitis as early as the 1950s, with clinical presentation of keratin precipitate, but increasing evidence has also suggested

causative capacity for CMV and non-herpesviruses such as rubella virus [60]. It is more common to see herpesviruses playing a part in uveitis in immunocompromised individuals, and this is usually the case in CMV-associated anterior uveitis. This was part of the rationale for the selection of sample UVSC07 for pyrosequencing. The patient had an extremely low CD4 count at the time of sampling, leaving her prone to opportunistic infection and the possibility of uveitis caused by viral infection or reactivation. Idiopathic uveitis is the dominant class of uveitis, representing approximately 48 percent of cases [53], but as molecular techniques of diagnosis evolve, more pathogens are surfacing as etiological agents. Other viruses such as EBV, HHV6, and chikungunya virus, a vector-borne *Alphavirus*, have been detected in patients with uveitis, but it is unclear whether these viruses are involved in pathogenesis or simply detected incidentally [60].

Ascertaining an association between detected pathogens in uveitis samples and disease is curtailed by the difficulty in obtaining adequate quantities of the same subtype of uveitis. Many of the ocular disease, including the types of white dot syndrome--the group of diseases studied at depth in this study--are uncommon. Subsets of the 47 uveitis aqueous humour and vitreous humour specimens were isolated based on their comparable ocular presentations and sequenced on the Roche 454. 454 pyrosequencing revealed complex metagenomic heterogeneity, with evidence of novel bacteria and viruses. Although three divergent viruses--HPV, circovirus, and ASFV--were found in these samples, there were not enough other samples with similar clinical phenotypes to consider, and these viruses were not detected in other samples regardless of matched phenotype.

The pyrosequencing performed on aqueous humour is the first attempt to study intraocular fluid as a metagenomic environment. It is presumed the aqueous humor is a sterile environment, and it appears through high-throughput sequencing that these eye chambers harbor more microbes than anticipated. An obvious explanation of the microbial diversity detected in the eye is contamination from the conjunctiva during anterior chamber paracentesis in spite of the surface of the eye and equipment being sterilized using standard procedures. The number of bacteria genera, however, are consistent with previously published findings on bacterial diversity on the conjunctiva [61].

A number of reads mapping to divergent pathogenic strains of bacteria were also identified, but the read depth on these samples was not enough to assemble larger contigs. The strain of *Bartonella* spp. detected appeared to map to various species within the genus of *Bartonella*, with one 190 base contig aligning with 73 percent amino acid identity to *B. bacilliformis*, one 170 base read aligning with 57 percent amino acid identity to *B. quintana*. and one 150 base read mapping with 65 percent amino acid identity to *B. henselae*. These were top alignments in BLAST, although *Bartonella* spp. are phylogenetically similar to other detected bacteria--*Rickettsia* and *Brucella*. *Bartonella* spp are gram-negative bacteria that are transmitted through vectors such as ticks, lice, and sandflies. Over 22 *Bartonella* species exist, with at least eight being known to infect humans [62]. *B. henselae* is the bacteria at the root of cat scratch fever, a

disease transmitted through bites or scratches from cats. A recent report documents the evidence for *Bartonella* spp. associated uveitis [63]. *B. henselae* and *B. grahamii* have been detected in past uveitis studies. In 1,417 uveitis patients included in this study, 2.6 percent had both ocular fluid that was PCR positive for *Bartonella* spp. and serum that had specific bands with Western blot [63]. Given the amount of microbial heterogeneity found in the aqueous humour during our high-throughput sequencing run, it is unknown whether this particular divergent strain of bacteria is involved in pathogenesis, but should be further investigated.

Besides bacteria, two of the viruses found were particularly interesting because they represented the first of their clade to be found in humans at the time of sequencing. The novel circovirus was discussed previously in the AIDS saliva project. The divergent *Asfarivirus* mapped only to members of the family *Asfarviridae*, a family of large double-stranded DNA viruses that only consists of one virus called African swine fever virus (ASFV). Comparative sequence analysis of the DNA polymerase of *Asfarviridae* to other viruses shows distant but clear phylogenetic similarity to other large double-stranded DNA virus families *Poxviridae*, *Herpesviridae*, and *Phycodnaviridae* [64]. ASFV causes acute hemorrhagic fever in pigs, but there has been no known zoonotic infection in humans [64]. Novel ASFV-like sequence was recently also detected in human serum and feces, also through the use of high-throughput sequence [64]. The *Asfarivirus* contig found in the uveitis study did not align to any of the 36 published ASFV-like sequence contigs. As evident in **Figure 3.3**, our contig was so divergent that

nucleotide BLAST to *nt* using the most lenient of parameters was not sensitive enough to identify it as *Asfarivirus* sequence. *Asfarivirus* reads were called in the final round of TBLASTx, where the sensitivity of a translated BLAST was able to detect an alignment across the entirety of the reads with only 35 percent amino acid identity. This finding could suggest that a human homolog of ASFV could exist, although further studies are required to determine the entirety of the viral genome and whether it causes human disease.

Acute liver failure

ALF is a severe disease that involves an acute onset of hepatic encephalopathy that is followed by jaundice and multi-organ failure. Its many etiologies include, most commonly, drug overdose, malignancies, and viral hepatitis [65]. About one third of all cases are of unknown etiology [66]. In the United States, the most common viral causes of ALF are hepatitis A (3 percent), hepatitis B (7 percent), and hepatitis C (<1 percent), and outside the United States, there is a larger number of cases attributed to hepatitis E. This cohort consists of ALF cases of unknown cause, and have all been tested negative for hepatitis A-E. 81 serum samples from individuals experiencing ALF were extracted and investigated by viral microarray. Viral signature on the microarray was confirmed by PCR. 14 cases with viral signature but negative by PCR for the suggested virus were barcoded and sequenced on one lane of the Illumina GAIIx. Although all samples received were screened for hepatitis A-E, one case of hepatitis C virus (HCV) was

detected in the multiplexed sequencing run. In a previous study of viral hepatitis induced ALF, hepatitis C virus accounted for only one case of ALF out of more than 1000 [56].

One of the barcoded ALF samples deep sequenced, FH569, yielded sequence for GB virus C, or hepatitis G virus, a virus not known to cause any disease. Despite originally being called hepatitis G, it has not been associated with either acute or chronic hepatitis, and does not have a tropism for liver cells [67]. GB virus C is a flavivirus genetically similar to HCV that can persist for years in serum. It has been detected in as many as 1.3 to 1.7 percent of healthy blood donors and nearly 13 percent of commercial blood donors [68]. While there is no pathological evidence substantiating it as a driver for acute or chronic hepatitis, there have been several reports suggesting a role in hepatitis while most cases of GB virus C infection are asymptomatic [68,69]. In one particular study of acute liver failure, 11 of the 22 (50 percent) patients had GB virus C RNA detected by semi-nested reverse transcriptase PCR, compared to the five of 106 (5 percent) control serum samples [68]. Five of the 11 infected with GB virus C were also tested positive for HBV, most likely due to shared risk factors. Earlier reports have made the case for ALF due to concurrent HBV and either HCV or hepatitis D virus infection, but neither of the latter were detected in the GB virus C positive ALF samples. As ALF patients often receive transfusions, several other studies have contradicted the suggestion of GB virus C and ALF, citing a need to control for such factors [67].

Finally, the most interesting virus detected was parvovirus B19, which conferred the most obvious viral signature by microarray (**Figure 3.4**). Human parvovirus B19 is a small single-stranded DNA virus with a tropism for human erythroid progenitors [70]. Most cases of human parvovirus B19 are asymptomatic, but can cause fifth disease, or erythema infectiosum [70] usually in children. A rash typically occurs, and is generally accompanied by fever and influenza-like illness [70]. Aplastic anemia has also been known to occur from persistent parvovirus B19 infection. Case reports of parvovirus B19 in ALF have been published [71,72], but large scale studies have not substantiated this suggestion [73].

Uveitis and ALF are both diseases that have documented viral etiologies, yet still consist of a large subset of idiopathic cases. Headway in molecular tool and sequencing technology development have diminished these share of cases, but no prior studies have taken an unbiased survey of all viruses, including the possibility of detecting divergent viruses. A number of unexpected viruses were found in each set, and further studies are need to follow up on the pathogenic role, if any, these viruses have on disease.

3.6. ACKNOWLEDGEMENTS

The uveitis study was done in collaboration with Dr. Alex Nugent and Dr. Hiroyuki Nomoto with advice from Dr. Mark Blumenkranz from Stanford Medical Center, Ophthalmology. Acute liver failure samples were collected and banked by Dr. Will Lee from UT Southwestern Medical Center.

FIGURES

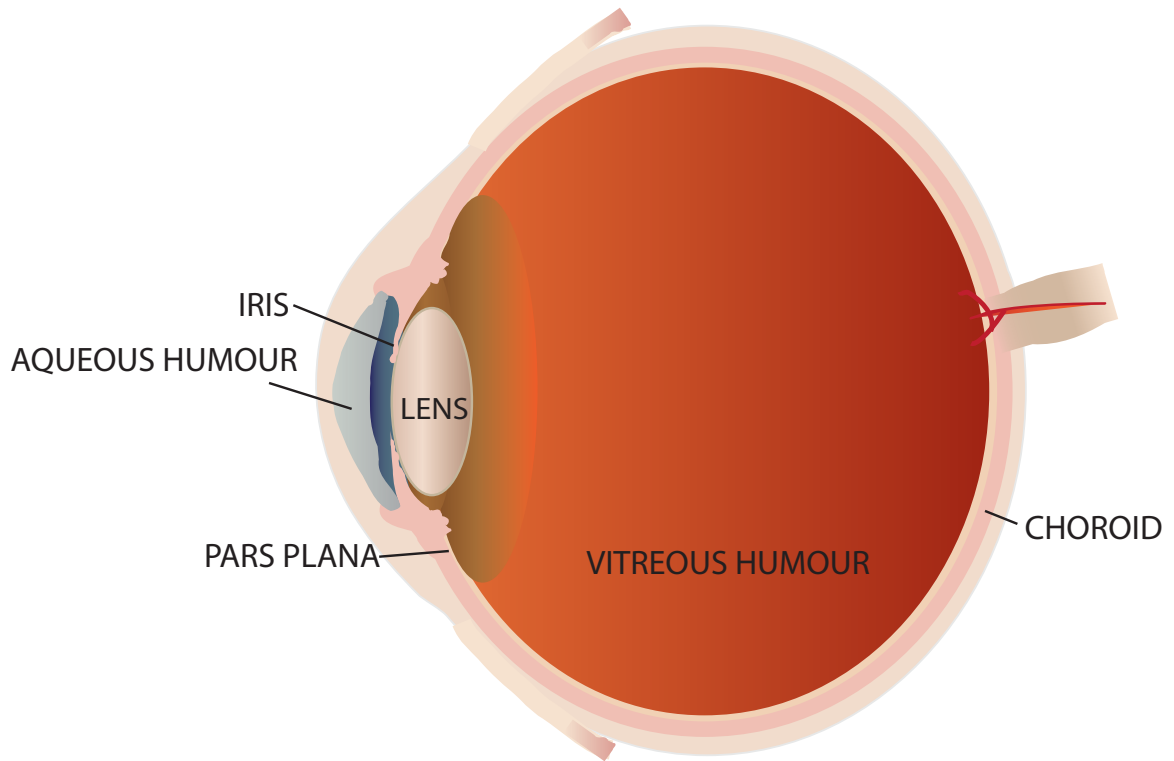


Figure 3.1. Diagram of the eye. Uveitis commonly involves inflammation of the pars plana, a subcomponent of the ciliary body, the choroid, and the iris. The sampling sites are the vitreous and the aqueous.

BLASTN HITS ATTRIBUTED TRANSLATED BLAST HITS ATTRIBUTED

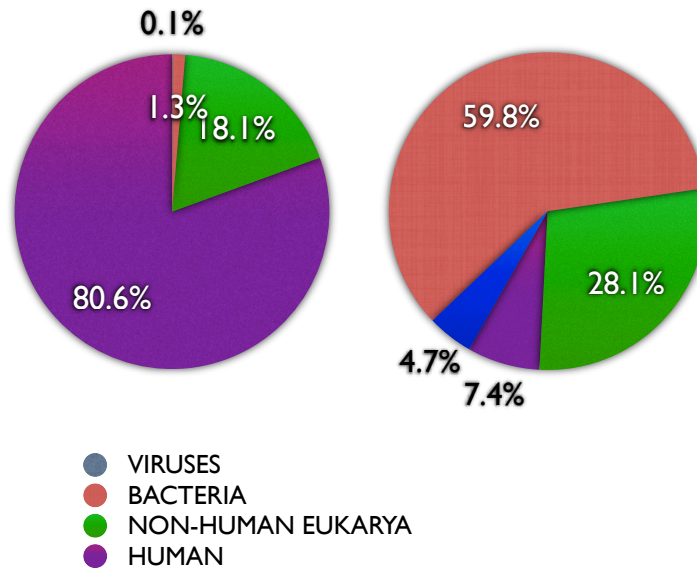


Figure 3.2. Taxonomic domain division of 454 reads using BLASTn and TBLASTx from uveitis pool sequencing run.

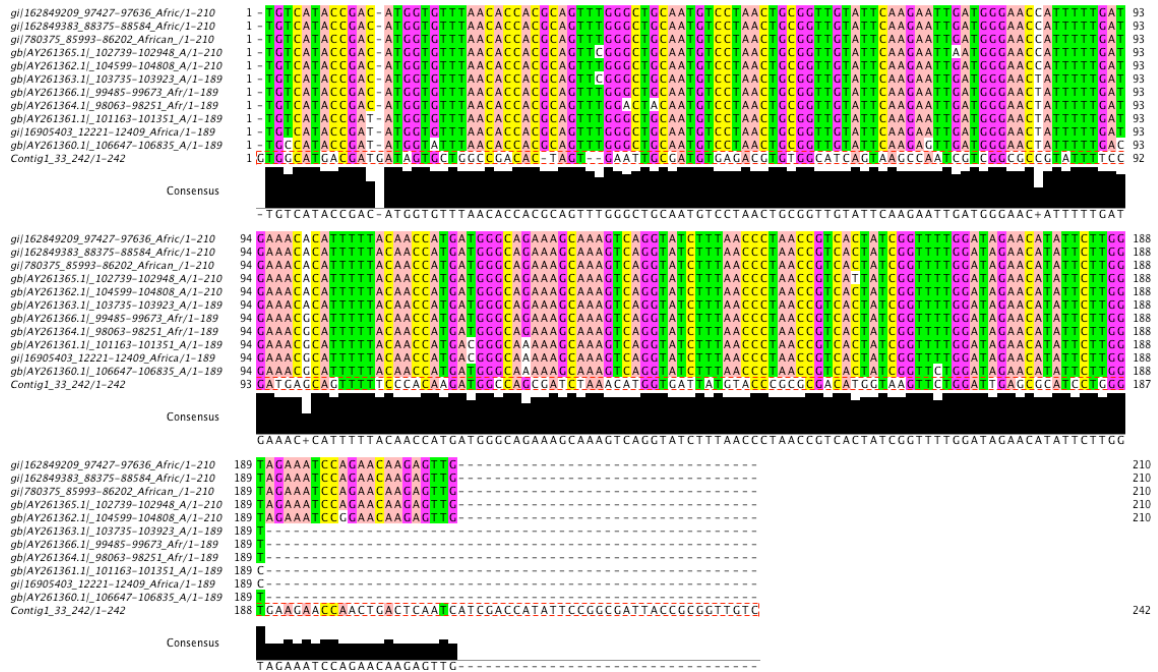
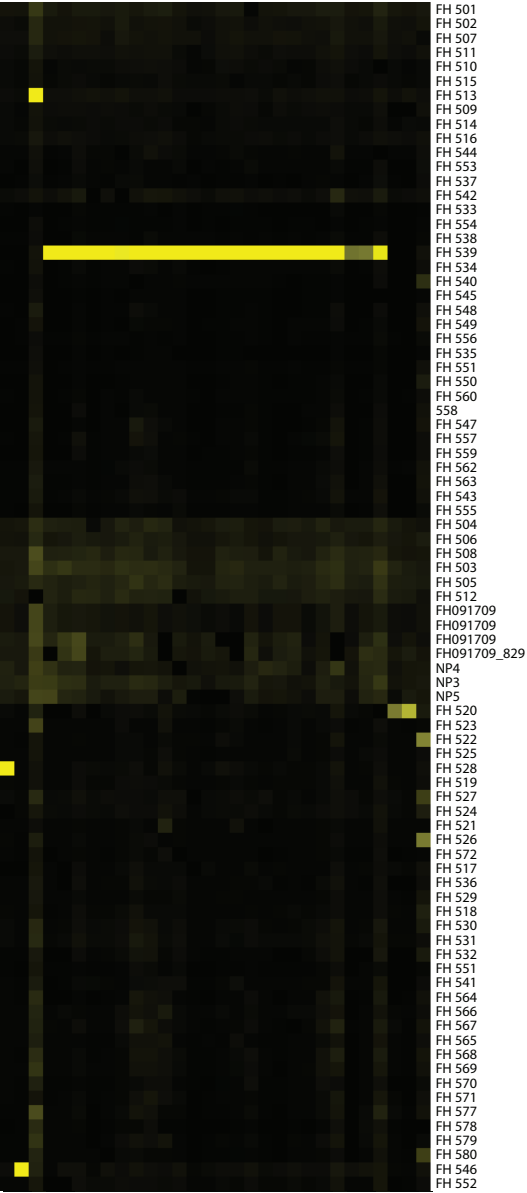


Figure 3.3. Multiple sequence alignment of assembled *Asfarivirus* contig, in the last row of each block, to strains of ASFV, depicted in Jalview.



ALF SAMPLES BY ARRAY

FH 501
 FH 502
 FH 507
 FH 511
 FH 510
 FH 515
 FH 513
 FH 509
 FH 514
 FH 516
 FH 544
 FH 553
 FH 537
 FH 542
 FH 533
 FH 554
 FH 538
 FH 539
 FH 534
 FH 540
 FH 545
 FH 548
 FH 549
 FH 556
 FH 535
 FH 551
 FH 550
 FH 560
 558
 FH 547
 FH 557
 FH 559
 FH 562
 FH 563
 FH 543
 FH 555
 FH 504
 FH 506
 FH 508
 FH 503
 FH 505
 FH 512
 FH091709
 FH091709
 FH091709
 FH091709_829
 NP4
 NP3
 NP5
 FH 520
 FH 523
 FH 522
 FH 525
 FH 528
 FH 519
 FH 527
 FH 524
 FH 521
 FH 526
 FH 572
 FH 517
 FH 536
 FH 529
 FH 518
 FH 530
 FH 531
 FH 532
 FH 551
 FH 541
 FH 564
 FH 566
 FH 567
 FH 565
 FH 568
 FH 569
 FH 570
 FH 571
 FH 577
 FH 578
 FH 579
 FH 580
 FH 546
 FH 552

9651821_30_c [Grass carp hemorrhagic virus | Reoviridae - Aquarionvirus
 1199791_64_c [Kendrick Italian haemovirus | Comoviridae - Neovirus
 1739046_69_c [Bulgarian fledging disease virus - 4 | Polyomavirus
 21389702_180_c [Front-foot-month disease virus - type 6/17.2 | Picornaviridae - Arthropovirus
 96227127_80 [Human papillomavirus type 35 | Papillomaviridae - Alphapapillomavirus
 9632996_4667_P70_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4883_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_99_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_106_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4527_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4602_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4667_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4754_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4522_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4602_P70 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_104 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_4883_P70_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_99 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_106 [Human parvovirus B19 | Parvoviridae - Erythrovirus
 3749708_104_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9632996_50_c [Human parvovirus B19 | Parvoviridae - Erythrovirus
 9630311_7_c [Gibson ape leukemia virus | Retroviridae - Gammaretrovirus
 20070095_96_c [Biome streak mosaic virus | Potyviridae - Tritonvirus
 4662323_330_c [Venezuelan equine encephalitis virus | Togaviridae - Alphavirus

PARVOVIRUS B19 OLIGOS



Figure 3.4. Cluster diagram of microarray results showing B19 parvovirus signature in FH539. Both arrays and all Virochip oligos are clustered, with arrays listed below the clustogram, and human parvovirus B19 oligos driving the cluster to the right of the clustogram. Cluster values are sum-normalized array intensities.

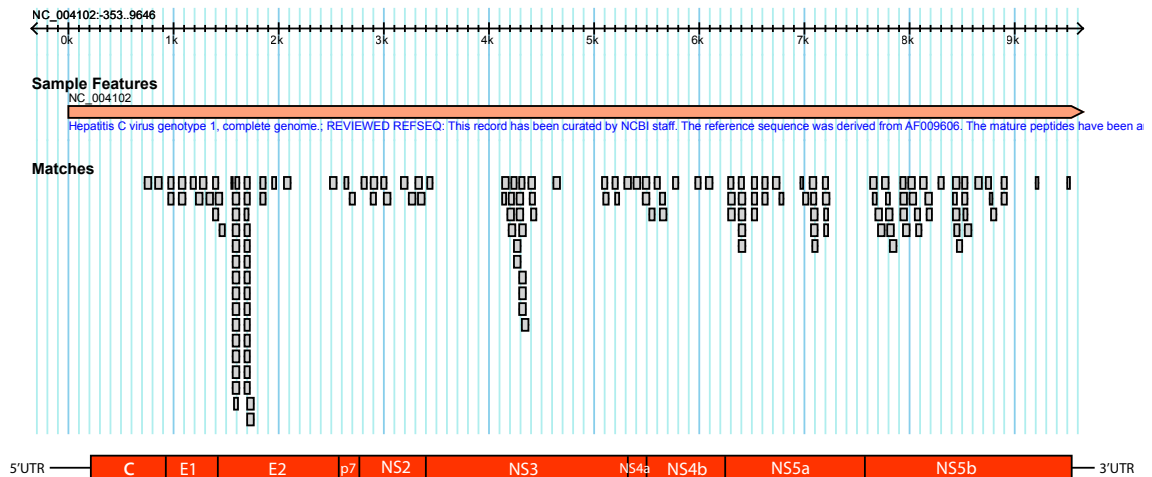


Figure 3.5. HTS reads (grey tick marks in Matches) from FH563 mapped to the hepatitis C virus complete genome (orange). The genome organization is shown below in red.

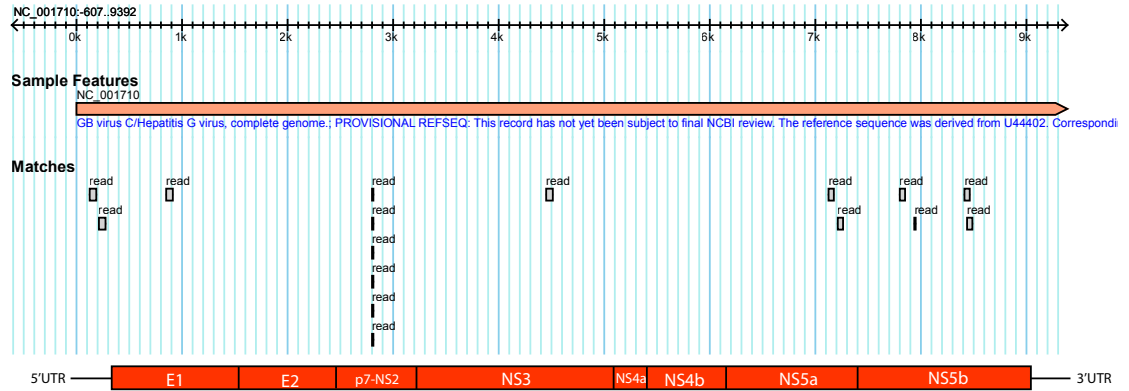


Figure 3.6. HTS reads (grey tick marks in matches) mapped to GB virus C complete genome (orange). The GB virus C genome organization is shown below in red.

TABLES

TABLE 3.1. UVEITIS SAMPLES DEEP SEQUENCED USING 454

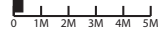
Sample	Clinical	Array call	Reason for inclusion	Platform	
UV 0019	Iridocyclitis	Alphavirus	Array	454 FLX	
UV 0037	posterior uveitis, diabetes retinopathy	CMV	Array	454 FLX	
UV SC07	AIDS, retinitis	--	Clinical	454 FLX	
UV 0039	pan-uveitis, aplastic anemia	Seadornavirus	Array	454 FLX	
WHITE DOT SYNDROME	UV 0047	AZOOB	Parvovirus	Array/Clinical	454 FLX
	UV 0021	APMPPE	--	Clinical	454 Titanium
	UV 0002	Serpiginous choroiditis	--	Clinical	454 Titanium
	UV 0005	Bilateral birdshot choroiditis	CMV	Array/Clinical	454 Titanium
	UV 0033a	Chorioretinitis	--	Clinical	454 Titanium
	UV 0041	Multifocal Chorioretinitis	--	Clinical	454 Titanium

VIRUSES	# READS	AVG % ID
PARAMECIUM BURSARIA CHLORELLA VIRUS	86/72	75% AA
ACANTHOCYSTIS TURFACEA CHLORELLA VIRUS 1	42	72% AA
HUMAN PAPILLOMAVIRUS TYPE 60/55/48	14	56% AA
CIRCOVIRUS (BEAK & FEATHER/COLUMBID)	36	47% AA
AFRICAN SWINE FEVER VIRUS	17	35%
PROKARYOTE		
BARTONELLA HENSELAE, BACILLIFORMIS, QUINT.	31	70% AA
RICKETTSIA TYPHI	11	61% AA
BORTADELLA PERTUSSIS	43	82% AA
BORRELLIA BURGDORFERI	16	73% AA
EUKARYOTE		
TOXOPLASMA GONDII	11	99%
CANDIDA ALBICANS	284	99%
PROTOHECA WICKERHAMII	2	95%
T. BRUCEI/CRUZI AND LEISHMANIA INFANTUM	14	43% AA
SARCOCYSTIS MIESCHERIANA	2	61% AA

Table 3.2. Viruses, bacteria, and eukaryotes identified by deep sequencing.

TABLE 3.3. ACUTE LIVER FAILURE SAMPLES SEQUENCED ON ILLUMINA GAIIX

Sample	Barcode	No. Read Pairs	Virus found (avg. % i.d.)
FH503	ATA	~4.2M	-
FH517	AGC	~0.5M	-
FH519	ACG	~1.0M	-
FH520	TAA	~3.0M	-
FH521	TTT	~2.5M	-
FH526	TGG	~3.0M	-
FH528	TCC	~1.0M	-
FH546	GAC	~1.5M	-
FH547	GTG	~1.0M	-
FH560	GGT	~3.0M	-
FH563	GCA	~2.0M	Hepatitis C virus (98% n.t.); HHV6B (100% n.t.)
FH566	CAG	~0.5M	-
FH567	CTC	~1.0M	-
FH569	CGA	~0.5M	GB virus C 96% n.t.



CHAPTER 4. VIRAL INFECTION IN ACUTE EXACERBATION OF IDIOPATHIC PULMONARY FIBROSIS

Citation: Wootton SC, Kim DS, Kondoh Y, Chen E, Lee JS, Song JW, Huh JW, Taniguchi H, Chiu C, Boushey H, Lancaster LH, Wolters PJ, DeRisi J, Ganem D, Collard HR. Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med.* 2011. 183(12):1698-702.

4.1. INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a progressive fibrotic lung disease with no known cause or cure [74][75]. Patients normally present with decreased capacity for lung exchange and periods of hypoxemia [76]. IPF is one of the most common forms of interstitial lung disease, and median time of survival after the time of diagnosis is around 3 years [77]. The frequency of the disease increases with age, and is marginally more prevalent in men than women [76].

While some patients experience a gradual progression of disease over time, many have periods of relative stability punctuated by episodes of acute respiratory worsening that are often fatal [78][79]. These marked episodes of severe respiratory decline are acute exacerbations of IPF [77].

Acute exacerbation of IPF is defined as an acute worsening of dyspnea characterized radiologically by the presence of bilateral ground glass abnormality on high-resolution

computed tomography (HRCT) and clinically by the absence of any identifiable cause [77]. It is estimated that between 5 and 10 percent of patients with IPF will experience an acute exacerbation annually, with more physiologically advanced disease at higher risk [77][80]. Upon hospital admission, mortality is over 60 percent, with many experiencing severe hypoxemia and requiring mechanical ventilation. Even after discharge, the mortality of survivors within 6 months after the event is over 90 percent [76]. It remains unclear whether acute exacerbation of IPF represents a primary acceleration of the underlying fibroproliferative process in IPF or is a clinically occult secondary complication (e.g. infection) [77][81].

Acute exacerbation of IPF is often accompanied by fever, increased cough, and myalgia, suggesting an infectious etiology. Respiratory viruses have been considered a particularly likely cause of acute exacerbation of IPF, based on the similarities in clinical and radiological presentation between acute exacerbation of IPF and viral pneumonitis, and the poor sensitivity of standard methods of detection [77]. Preliminary evaluations of the role of infection in acute exacerbation have yielded mixed results [82][83].

In this study, we tested the hypothesis that acute exacerbation of IPF is caused by occult viral infection by prospectively collecting BAL from patients experiencing acute exacerbation of IPF and controls (stable IPF and acute lung injury (ALI)). We used multiplex PCR and an array-based discovery platform (Virochip) to test for the presence of known and novel viruses, and applied next-generation parallel sequencing (high-

throughput sequencing) to a subset of acute exacerbation samples to improve the sensitivity for detecting the presence of virus and to fully describe their microbial flora.

4.2. METHODS

Study Population

Patients with acute exacerbation of IPF were identified prospectively from two centers (University of Ulsan, Korea; Tosei General Hospital, Japan). Diagnostic criteria for acute exacerbation of IPF were prespecified and required the following: 1) previous or concurrent diagnosis of IPF; 2) unexplained worsening within 30 days; 3) HRCT evidence of bilateral ground glass abnormality and/or consolidation; 4) no evidence of pulmonary infection by respiratory culture; 5) exclusion of alternative causes of respiratory worsening [77]. All acute exacerbation patients had negative clinical evaluation for infectious causes including routine bacterial and viral BAL cultures.

Control patients with stable IPF and ALI were identified from a single center (University of Ulsan, Korea) and underwent bronchoscopy at the time of diagnosis. BAL from a case of IPF that was found positive for rhinovirus and cytomegalovirus (CMV) was included with the samples as a blinded positive control. IPF and ALI was defined by consensus criteria [74,84,85]. All centers received approval from their institutional review board or equivalent, and all patients provided informed consent.

Sample collection and processing

In all cases, bronchoscopy was performed as part of patients' clinical evaluations. In most cases, BAL was collected within the first 48 hours of admission to the hospital. In general, BAL was performed in a single sub-segment of the right middle lobe or lingula, with at least 100 milliliters of sterile saline instilled. A subset of acute exacerbation samples and stable IPF controls underwent phlebotomy at the time of bronchoscopy. All samples were stored at -80 degrees Celsius until ready for processing. Total RNA was extracted from 200 microliters of each sample using the RNeasy mini kit (Qiagen, Inc., Valencia, CA) according to the manufacturer's instructions.

PCR analysis

A blinded, nested respiratory multiplex was run on the BAL samples from acute exacerbation of IPF samples and stable IPF controls for pre-specified respiratory viruses (influenza virus, human parainfluenza virus, respiratory syncytial virus (RSV), human rhinovirus, human enterovirus, human coronavirus, human metapneumovirus, and human adenovirus) [86]. All PCR positives were run on a 1.5 percent 1xTAE agarose gel stained with Ethidium Bromide, cut at the expected bandwidth, and extracted for DNA using the QiaQuick gel extraction kit (Qiagen, Inc., Valencia, CA). This product was then directly sequencing using standard di-deoxy sequencing on the ABI3730 sequencer (ElimBio, Hayward CA). PCR was also carried out to confirm viral signatures indicated from Virochip analysis (herpes simplex virus (HSV), Epstein-Barr virus (EBV), and torque teno virus (TTV)) using previously published nested primer sets [87][88][40]. Additional

PCR for TTV was performed on BAL samples from ALI controls and serum samples from a subset of acute exacerbation of IPF and stable IPF controls in an effort to better define the epidemiology of TTV.

Pan-viral microarray analysis

Acute exacerbation of IPF samples and stable IPF controls were randomly amplified to generate cDNA which was hybridized blindly to the Virochip, a pan-viral microarray, as previously described [1]. Arrays were scanned using the Axon 4000B scanner and intensities were calculated using GenePix 6.0 (Axon Instruments, Union City, California). The presence of a viral signature was determined using Cluster 3.0 [35] and E-predict [34].

Deep sequencing and read analysis

BAL from twelve of the study patients with acute exacerbations of IPF were selected for high throughput sequencing on the Illumina Genome Analyzer Iix platform (Illumina, San Diego, CA) based on the presence of symptoms of a viral-like illness including fever, cough, and myalgia. A subset of these selected BAL samples were from individuals who were also mechanically ventilated.

Libraries were prepared for deep sequencing using an in-house protocol of obtaining the correct topology of sequencing primer on cDNA inserts. Nucleic acid extracted from BAL was first primed using abbreviated Illumina A and B adaptors attached to a unique

3-bp sequence tag (barcode) followed by a random hexamer. The 3-bp barcode was incorporated into each of the twelve samples to allow for 12-plex barcoded sequencing within a single lane on the Illumina flow cell. cDNA was amplified for 25 cycles of PCR using the barcoded adaptors, and the PCR product was run on a 4 percent native polyacrylamide gel at 4C to select for a narrow size distribution centered around 250-bp. The amplicons were then precipitated with 100 percent ethanol at 4C and resuspended in 16 microliters of water. Two microliters were carried into a second round of PCR amplification using the abbreviated A adaptor and a full length B adaptor for 15 cycles using 22-bp of the 3' end of the Illumina A adaptor and 61-bp of the Illumina B adaptor as primers. The product was size selected once more for products around 304-bp, which would carry the correct A/B topology. Ethanol precipitated DNA was then PCR amplified for ten cycles using the full length Illumina A adaptor and the 5' end of the B adaptor. This final library was sequenced on one lane of a paired end deep sequencing run with 65-bp read from each end of the insert.

Low complexity reads with inadequate Lempel-Ziv-Welch (LZW) compression ratios were removed, and barcodes indicated by the sequence of the first 3-bp of each read were used to bin reads according to their original sample [89]. All remaining reads were filtered for high identity human sequence using a stringent BLAT to the human genome [38]. Passed filter reads were then pushed through a pipeline of sequential BLAST alignments to the NCBI NT database [39]. First, high identity hits were isolated using a MEGABLAST with a word size of 28 against *nt*. Reads that did not align significantly to

nt with a high word size were then aligned to *nt* again using MEGABLAST with word size of 12 and e-value of 10^{-7} , followed by a sensitive BLASTn alignment with a word size of 7 and an e-value of 10^{-3} . All hits at every step were sorted by their NCBI taxonomy identifiers (NCBI taxIDs).

Statistical methods

Clinical data are expressed as means or percentages, unless otherwise stated. The primary comparison was between BAL samples from acute exacerbation of IPF and stable IPF controls. Additional comparisons were made between BAL samples from acute exacerbation of IPF and ALI controls, and serum samples from acute exacerbation of IPF and stable IPF controls. In all cases, intergroup comparisons were performed conservatively using non-parametric methods (Wilcoxon signed-rank test) and Chi-squared/Fishers exact analyses as appropriate. Regression analysis was performed to determine the relationship between clinical factors, PCR positivity, and survival. Clinical data analysis was performed using SAS 9.1 (SAS Institute, Cary, NC). Statistical significance was defined as a p value < 0.05 .

4.3. RESULTS

Patient characteristics

Forty-three patients with acute exacerbation of IPF (35 Korean, 8 Japanese), forty patients with stable IPF, and twenty-nine patients with ALI with BAL were enrolled between 2006 and 2009. Their clinical characteristics are summarized in **Table 4.1**.

Patients with ALI controls were primarily respiratory in etiology as these were the cases that underwent BAL for clinical purposes. Serum was obtained from a subgroup of the acute exacerbation of IPF patients (n=22) and stable IPF (n=31) controls.

Viral detection by PCR and pan-viral microarray

Four acute exacerbation of IPF BAL samples (9%) were positive for common respiratory viruses by initial multiplex PCR (two for rhinovirus, one for coronavirus (human coronavirus-OC43), and one for parainfluenza virus-1). All stable IPF samples were negative for common respiratory viruses ($p = 0.12$, **Table 4.2**).

Array analysis of acute exacerbation of IPF BAL samples revealed the presence of torque teno virus (TTV) and several human herpesviruses. To pursue this finding further, we carried out sensitive genome-specific PCR reactions for HSV, EBV and TTV on all samples. This yielded 15 additional BAL positives in acute exacerbation of IPF samples (**Table 4.3**). Of these, only TTV was significantly more common in acute exacerbation of IPF compared to stable controls (28 percent vs 0 percent, $p=0.0003$). Four BAL samples revealed double infections: two with TTV and rhinovirus, one with TTV and parainfluenza virus-1, one with TTV and HSV. One BAL sample revealed a triple infection of TTV, EBV, and coronavirus. Overall, 14 (33 percent) of acute exacerbation of IPF samples were positive for virus compared to no positives in the stable IPF samples ($p < 0.0001$). There was no difference in the frequency of fever and myalgia between

virus positive and virus negative cases, and there was no significant difference in the use of corticosteroid treatment.

Viral detection by high-throughput sequencing

BALs from eleven of the study patients with acute exacerbation of IPF were selected for high-throughput sequencing based on the presence of symptoms of a viral-like illness including fever, cough, and myalgia (**Table 4.4**). The twelfth sample, IPF264, was included to keep the heterogeneity of the barcode nucleotides for cluster identification during the Illumina run, and this sample was selected because of its human parainfluenza virus (HPIV) signature on the Virochip but was PCR negative for all human parainfluenza viruses. We elected to sequence this subset of 12 samples to investigate the possibility of viruses being overlooked by PCR and microarray. Of these samples, two were PCR-positive for tested viruses—one for TTV and one for TTV and HSV. After initial quality filtering, approximately 26 million pairs, or 52 million total reads, comprised the primary dataset. Each of the 12 barcoded acute exacerbation of IPF samples was represented with at least 3 million high quality reads. Over 98 percent of the reads were derived from human origin, and of the remaining reads, approximately 0.1% were recognizably bacterial in origin. Only a few hundred were potentially attributable to known non-human eukarya and viruses. Aside from bacteriophages, only three viruses (two TTVs and one HSV) were found, consistent with the PCR results. After all stages of mapping to a sequence database were finished, approximately 0.6 percent of the original dataset remained without attribution. When these reads were passed to PRICE as seeds

for assembly upon the earlier set of human filtered reads, no additional virus was detected.

Comparison of TTV positive and negative acute exacerbations

There were no significant differences in age, gender or baseline pulmonary function in TTV positive acute exacerbation of IPF patients compared to TTV negative acute exacerbation of IPF patients. TTV positive patients appeared sicker, with 58 percent requiring mechanical ventilation (vs. 29 percent in TTV negative patients, $p = 0.09$) and 75 percent dying at 60 days (vs. 42 percent in TTV negative patients, $p = 0.06$). Overall mean survival time in the TTV positive patients was 29 days (vs. 88 days in TTV negative patients, $p = 0.19$, **Figure 4.1**). Bivariate regression analysis of potential independent predictors of survival time (presence or absence of prednisone treatment at the time of BAL, mechanical ventilation and the time of BAL, and TTV positivity on BAL) revealed only mechanical ventilation as statistically significant (hazard ratio 2.30, $p = 0.03$). TTV positivity was not a significant predictor of survival time (HR 1.65, $p = 0.20$).

TTV positivity in acute exacerbation of IPF and stable IPF serum

Six of 22 patients (27 percent) with acute exacerbation of IPF were PCR positive for TTV in serum, compared to five of 31 patients (16 percent) with stable IPF ($p = 0.34$). Three of the six acute exacerbation patients (50 percent) that were PCR positive in serum

were also PCR positive in BAL. Four of the 16 acute exacerbation patients (25 percent) that were PCR negative in serum were PCR positive in BAL.

TTV positivity in ALI BAL

The statistically significant link between BAL-associated TTV and acute exacerbation of IPF prompted us to also examine BAL samples from 29 patients with ALI. These BAL samples were collected using the same protocol as the acute exacerbation of IPF samples, and the majority of the ALI samples were from patients with pneumonia.

TTV was detected in seven of 29 BAL samples (24 percent) from ALI patients; this was not significantly different from the prevalence of TTV in BAL samples from acute exacerbation of IPF (28 percent, $p = 0.73$).

4.4. DISCUSSION

Using highly sensitive PCR, pan-viral microarrays, and high-throughput sequencing technologies in a large, well-described cohort of patients with acute exacerbation of IPF and controls, we found that most cases of acute exacerbation of IPF had no evidence of an underlying viral infection. This suggests that viral infection is not a common cause of acute exacerbation of IPF.

Overall, we found viral nucleic acid in the BAL of 33 percent of patients with acute exacerbation of IPF; no viruses were found in samples from stable IPF controls. There were two rhinovirus-positive samples, one coronavirus-positive sample, and one

parainfluenza virus–positive sample, suggesting that a small minority (9 percent) of acute exacerbations of IPF may be caused by occult infection with common respiratory viruses. Surprisingly, the most common virus detected in the BAL of acute exacerbation of IPF patients was TTV, which was present in 28 percent of acute exacerbation BAL samples. This finding was not unique to acute exacerbation of IPF because 24 percent of BAL samples from ALI controls were also TTV positive.

Two recent studies have commented indirectly on the possible role of occult viral infection in acute exacerbation of IPF. The first study performed gene expression microarrays on whole lung tissue from 8 patients who died of acute exacerbation of IPF, 23 patients with stable IPF, and 15 healthy controls [83]. The authors concluded that acute exacerbation of IPF was characterized by a pattern of enhanced epithelial injury and proliferation, but found no gene expression profiles indicative of a response to viral or bacterial infection. In a second study of 27 patients presenting with acute decline in fibrotic lung disease (13 of whom had confirmed acute exacerbation of IPF), 5 had antigenic or PCR evidence of viral infection (one parainfluenza virus, two HSV, and two CMV infections), three of which were missed on standard viral culture [82].

Our study expands significantly on previously published reports. First, we take an unbiased approach to viral discovery using cutting-edge genomic methodology. It is the first study to do this in acute exacerbation of IPF. Our use of sequencing to confirm all suspected viruses rules out the possibility of spurious PCR results, a common pitfall of

the technique. Second, our large cohort of well-defined patients with acute exacerbation with adequate controls allows for greater certainty regarding our conclusions. Third, we have identified an unexpected virus (TTV) that was associated with 33 percent of acute exacerbations, and that was absent in stable IPF.

The pathogenetic significance of TTV in acute exacerbation of IPF BAL is unclear. TTV is a non-enveloped single-stranded circular DNA virus that exists in a genetically diverse clade [50,90]. The virus seems to have broad tissue tropism because it has been detected in peripheral blood mononuclear cells (PBMCs) and bone marrow, spleen, liver, and lung [50]. Infection with TTV in the human population is worldwide, with prevalences of viremia ranging from 8 to 80 percent depending on the population studied and detection methodology used. When only considering the hemi-nested PCR of the N22 region used in this study, rates of TTV DNA found in healthy blood donors range from 8.4 to 12 percent [91,92] and do not seem to correlate with the geographic location of the patients. Most infected subjects are asymptomatic, and to date efforts to link TTV viremia with any acute or chronic pathologic state have been unsuccessful [50]. Although there have been reports of TTV in the upper respiratory tract (nasopharynx and oral cavity) [93], TTV has not been identified in BAL fluids. TTV has previously been detected in the serum of 12 (36 percent) of 33 Japanese patients with IPF. In this study, TTV appeared more frequently in cases that progressed to acute exacerbations, and TTV positivity was suggested to correlate with worse survival [94]. Our findings do not show a correlation between the presence of TTV in the serum and the presence of TTV in the

BAL, or any correlation between serum TTV positivity and a diagnosis of acute exacerbation.

It is possible that *de novo* TTV infection in the lung causes direct alveolar epithelial cell injury and acute respiratory worsening. If so, this process does not seem to be unique to acute exacerbation of IPF because we detected TTV at a similar frequency in BAL from patients with ALI. Although this does not exclude a potential role for TTV in the pathogenesis of acute exacerbation of IPF, it is also compatible with the idea that inflammation or injury in the lung may nonspecifically trigger local TTV replication, or may result in increased vascular permeability in the lung allowing circulating virus to enter the alveolar compartment. In the latter two cases, the presence of TTV would represent a consequence of lung inflammation rather than its cause. The idea that local TTV replication might be enhanced by underlying inflammatory signaling is supported by *in vitro* studies of PBMCs from donors who are TTV negative [95]. These PBMCs were infected *in vitro* with TTV, cultured with and without the presence of phytohemagglutinin, lipopolysaccharide, and interleukin-2, and then examined for evidence of TTV replication. In this experiment, TTV mRNA and replicative intermediates were only found in the stimulated PBMCs, consistent with an infection-amplifying role for inflammatory signaling.

The methodologies used in this study have unparalleled sensitivity for viral detection. The multiplex nested PCR is several fold more sensitive than virus culture and direct

immunofluorescent tests, with the ability to amplify less than 10 copies of target nucleic acid [86]. For viral discovery, however, PCR is of limited use because it identifies only *a priori* viral targets. Pan-viral microarray precludes the need for a preconceived list of targets, although even with its proved sensitivity, its benefit is dependent on the signal-to-noise ratio of the nucleic acid [1]. The use of deep sequencing to look further for evidence of viral infection in a high-risk subpopulation of patients with acute exacerbation therefore adds confidence to our results, because it produces an unbiased, high-resolution description of the microbial landscape of the sample tested. This technology has been used to identify novel viruses in human diarrhea, and to describe the microbiome of the distal gut [96][97], but never in BAL [98][99]. In the current study, two samples subjected to deep sequencing were positive for known viruses by PCR and pan-viral array screening. Using an efficient and sensitive pipeline for sorting reads, these positive PCR findings were confirmed, and no additional viruses were detected in these or the other samples tested. Interpretation of these findings must be tempered by the fact that existing computational methods for recognizing potential viral genomes are imperfect, and may fail to identify novel agents with only limited homology to known viral genera. The same is true of array-based viral detection methods.

One important limitation of this study is the potential for false-negative results because of the timing of sample collection. BAL was performed early in the course of hospitalization, most commonly in the first 48 hours after admission, and the median time from symptom onset to sampling was 7 days. Importantly, no difference in the time from

symptom onset to sample collection was found between virus-positive and virus-negative cases. The duration of replicating virus in BAL is largely unknown, and it is possible that a virus could have stopped shedding during this time. In this study, we have maximized our likelihood of detecting virus by obtaining BAL samples early after admission and using highly sensitive viral detection techniques.

In summary, this study used unbiased, highly sensitive genomics-based discovery methods to investigate the role of viral infection in a large, well-characterized cohort of patients with acute exacerbation of IPF. The results of this study suggest that most cases of acute exacerbation of IPF are not caused by viral infection. Future research into the etiology of acute exacerbation of IPF should confirm these findings, further investigate the role of TTV, and consider other possible occult complications (e.g., aspiration) that may cause acute respiratory worsening in these patients.

4.5. SUPPLEMENTAL MATERIAL

The online data supplement to this article can be found on the AJRCCM website (<http://ajrccm.atsjournals.org/>).

4.6. ACCESSION NUMBERS

Illumina reads and information on the run were uploaded to the NCBI Short Read Archive (SRA) under the accession SRX042016. All microarray data were uploaded to the Gene Expression Omnibus (GEO) under the accession GSE27578.

4.7. ACKNOWLEDGEMENTS

This project was done in collaboration with Dr. Harold Collard. I would also like to thank Amy Kistler for her effort with the initial sample processing and for offering her review. I also thank Drs. Naftali Kaminski and Talmadge E. King Jr. for their thoughtful input and guidance in the early stages of this project, and James Graham Ruby for his contribution to the sequencing library preparation.

FIGURES

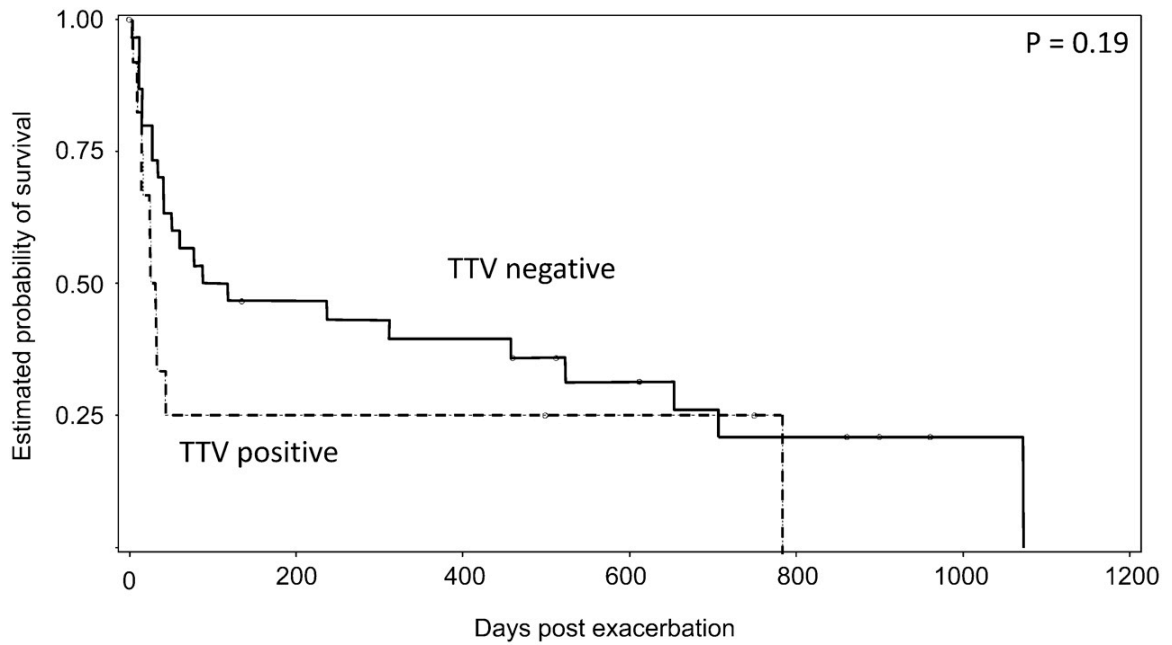


Figure 4.1: Survival time in torque teno virus (TTV)–positive acute exacerbation of patients with idiopathic pulmonary fibrosis compared with TTV-negative acute exacerbation of patients with idiopathic pulmonary fibrosis. Mean survival time was 29 days versus 88 days, respectively. P=0.19.

TABLES

TABLE 1. CLINICAL CHARACTERISTICS

Variable	Acute Exacerbation (n = 43)	Stable (n = 40)	ALI (n = 29)
Age, yr	65	66	60
Male sex, %	88	75	66
Surgical lung biopsy, %	21	28	NA
Smoking, %	84	75	48
Baseline FVC, %	73	79	NA
Baseline DL _{CO} , %	60	70	NA
Mechanical ventilation, %	37	NA	100
Immunosuppressive therapy, %	60	NA	NA

Table 4.1. Clinical characteristics

TABLE 2. RESPIRATORY VIRAL DETECTION IN ACUTE EXACERBATION AND STABLE IDIOPATHIC PULMONARY FIBROSIS

Virus	Acute Exacerbation (n = 43)	Stable (n = 40)	P Value
Any respiratory virus (%)	4 (9)	0 (0)	0.12
Rhinovirus (%)	2 (5)	0 (0)	0.49
Coronavirus (%)	1 (2)	0 (0)	1
Parainfluenza (%)	1 (2)	0 (0)	1
Adenovirus (%)	0 (0)	0 (0)	–
Enterovirus (%)	0 (0)	0 (0)	–
Influenza (%)	0 (0)	0 (0)	–
Metapneumovirus (%)	0 (0)	0 (0)	–
Respiratory syncytial virus (%)	0 (0)	0 (0)	–

Table 4.2. Respiratory viral detection in acute exacerbation and stable idiopathic pulmonary fibrosis

TABLE 3. ARRAY-BASED VIRAL DETECTION IN ACUTE EXACERBATION AND STABLE IDIOPATHIC PULMONARY FIBROSIS

Virus	Acute Exacerbation (n = 43)	Stable (n = 40)	Acute lung injury (n = 29)	P Value*
Torque teno virus (%)	12 (28)	0 (0)	7 (24)	0.0003
Epstein-Barr virus (%)	2 (5)	0 (0)	NA	0.49
Herpes simplex virus (%)	1 (2)	0 (0)	NA	1
Cytomegalovirus (%)	0 (0)	0 (0)	NA	–

* P value is for comparison of acute exacerbation with stable control.

Table 4.3. Array-based viral detection in acute exacerbation and stable idiopathic pulmonary fibrosis

TABLE 4.4. ACUTE EXACERBATION OF IPF SAMPLES SELECTED FOR DEEP SEQUENCING

Sample	Reason for inclusion	PCR	Array call	Barcode
IPF 204	Fever and myalgia	(+): TTV	TTV	GAG
IPF 205	Fever and myalgia	(+): HSV, TTV	HRV14	AGG
IPF 216	Fever and myalgia			GGA
IPF 219	Fever		HPIV	CCA
IPF 220	Fever			TAT
IPF 227	Fever and myalgia			CAC
IPF 247	Fever and myalgia	(+): TTV	HMPV	GTC
IPF 264	Array		HPIV	ACC
IPF 268	Fever and myalgia			ATT
IPF 278	Fever			TCG
IPF 296	Fever and myalgia			CGT
IPF 306	Fever			TTA

Table 4.4. Acute exacerbation of IPF samples selected for deep sequencing

CHAPTER 5. VIROGAP: USING VIRUS-HOST COPHYLOGENY TO IDENTIFY TARGETS FOR VIRUS DISCOVERY

5.1. INTRODUCTION

Viruses, for the most part, were left out of the early metagenomics movement, as traditional techniques of virus detection have their limitations. Additionally, the extensive divergence between and within virus families and the lack of a universal gene, such as the 16s and 18s rRNA sequence found in other life forms, render sequence-dependent informatics methods insufficient. In response to the microbial metagenomics revolution, methods of “gap-filling” have been employed to provide a systematic approach of selecting underrepresented bacterial and archaeal lineages in the Tree of Life to target with high-throughput sequencing [100]. Virogap applies the underlying goal of gap-filling to virus discovery, where the assumption of coevolution between virus and host allows translation of host organism homology to homology between known viruses and their proposed orthologs. These gaps not only give insight into the approximate representation across all viral lineages but also generate hypotheses of potential targets for virus discovery. Using viral and host sequences, viral sequence networks for all viral families are mapped to their corresponding host networks to expose clades of orthologs lacking virus counterparts in particular hosts (gap-recognition), ultimately to motivate targeted chip-based and high throughput sequencing (HTS)-based methods of gap-filling.

Virogap is an alternative approach to virus discovery, a process that has traditionally

involved studying either disease cohorts with no known etiology or surveying "high-yield" groups such as immunocompromised patients. This idea of gap-targeted viral discovery has been used informally, as in the case of the the subfamily gammaherpesvirinae of herpesviridae. This subfamily of herpesviruses comprises two genera—lymphocryptoviruses, which had been isolated from both Asian and African apes, and rhadinoviruses, which had been only found in African apes [101].

A previous study used this observation to design a set of degenerate primers based on neighboring rhadinovirus sequence, and discovered a novel virus in Asian apes [101].

The rhadinovirus genus is even more interesting if you consider its phylogenetic tree based on the ultra-conserved DNA polymerase. It reveals two distinct sub-clades which mirror each other with almost the same set of primate hosts. The human host, however, is an exception, and Kaposi's sarcoma-associated virus (KSHV) is a rhadinovirus that lacks a counterpart virus in the opposing lineage. It is thus believed by many of those studying herpesviruses that a ninth human herpesvirus exists, and it would likely be in the second rhadinovirus lineage in the position of the gap [101].

While this dissertation features virus to host gap recognition in double-stranded DNA viruses, this dissertation discusses a method that can be easily generalized. This method may be used at a genus level or even at the family level, and can be broadened to apply to all viruses. It recognizes targets for viral discovery, and can be used to describe the selection of specimens to investigate, the design of the discovery platforms (microarrays),

and the ensuing bioinformatics to handle the vast amount of metagenomic sequencing data.

5.2. BACKGROUND

Coevolution between virus and host

Earlier studies have shown that viruses, usually microparasitic in nature, and their hosts often coevolve [23,102]. The necessity for viruses to coexist with their hosts is a major driving force for coevolution. In particular, viruses must be able to replicate enough to be communicable while still tempering the severity of disease such that there is enough host survival for the propagation of the virus. Viruses must also overcome mechanisms of host resistance such as RNA silencing in plants and fungi, restriction endonucleases in prokaryotes, and immune systems in vertebrates [103]. Virus evolution can therefore be understood by studying the evolution of its hosts and host-microbe interactions.

However, the degree of cospeciation varies, and is oftentimes most obvious in viruses that can maintain lower nucleotide substitution rates, such as those with DNA polymerases. It has been suggested that the amount of cospeciation depends on whether the virus can maintain persistent infection in specific host individuals (**Table 5.1**). On the other end of the spectrum, viruses that lead an acute lifestyle are usually associated with high virulence and high transmissibility, and therefore find new hosts during their short window for replication. These viruses, in turn, are highly dependent on the host population structure and rarely cospeciate with host species.

Coevolution between DNA viruses and their hosts has been shown in several families, including *Poxviridae*, *Polyomaviridae*, and *Herpesviridae* [102,104,105]. In the family *Herpesviridae*, parallel evolution between hosts and microbe has been phylogenetically demonstrated for avian, mammal, and reptile hosts; within subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae*, and *Gammaherpesvirinae*, it is evident there are congruent branchings of lineages and evolutionary distances between the phylogenetic trees of the viruses and their hosts [102].

Cospeciation in RNA viruses is a topic of contention—it has been both demonstrated and refuted for hantaviruses, spumaviruses, and arenaviruses [106][107][108]. With the high rate of RNA polymerase errors due to its lack of proofreading mechanisms it seems unlikely that RNA virus evolution, compared to the protracted duration of host evolution, that viruses diverged from each other at the same time their hosts did. However, RNA viruses that can maintain persistent infection in effect replicate less often and are believed to cospeciate with their hosts [106]. Additionally, certain retroviruses that are under less severe levels of immune selective pressures have shown signs of cospeciation. For example, simian foamy viruses (SFVs) are ubiquitous, persistent, non-pathogenic retroviruses that infect almost all primates. The phylogenetic tree for the highly conserved *pol* gene in 44 SFVs and mitochondrial cytochrome oxidase subunit II (COII) of 55 primates reveals 22 cospeciation events [109].

Cophylogenetic reconstruction

Cophylogeny mapping is the process of using the observed relationships between the leaves of associated phylogenetic trees to infer the most likely co-evolutionary events [110]. This is a computationally expensive process that has been shown to be NP-complete [111], and an approximation of the solution is not even tractable when the trees are not highly congruent.

Sequence similarity networks

As an alternative to phylogenetic reconstruction, relationships can be represented in a network of all-against-all pairwise distances between sequences. This is a powerful tool that has been shown to produce results comparable to phylogenetic analysis, but allows all interactions to be considered instead of just the most optimal model [112].

Assumptions

To clarify, this method of gap-recognition in the lower branches of the viral tree of life does not assume gaps in the tree necessarily equate to unidentified viruses, but rather creates a system of hypothesis generation to motivate viral discovery. Virus discovery indeed involves a certain measure of uncertainty, but this approach, as any other targeted metagenomics project, creates a bias that presents an advantage in virus discovery. In particular, the gaps would encapsulate information about the host species as well as sequence from neighboring viruses that could influence experimental design.

Additionally, if the viruses cluster by disease or another phenotype, it would be possible to use these characteristics as a guide in syndrome-based viral discovery.

5.3. METHODS

A simple schematic of Virogap is shown in **Figure 5.1**. With sequence from all viral proteins (**Figure 5.1.a**), clustering of an all-against-all network is performed (**Figure 5.1.b**). Using the largest protein cluster (**Figure 5.1.c**), known hosts are mapped to each virus. The basal example would be a one-to-one mapping of virus to host, as depicted, although a one-to-many mapping is also permitted in this implementation. The host sequence from a widely conserved protein is then used to create a host network (**Figure 5.1.e**). The lowest common ancestor (LCA) of these hosts is determined from taxonomy, and a network of every organism under the LCA is created (**Figure 5.1.f**). Only some of these ‘unmatched hosts’ (yellow) are highly connected to ‘matched hosts,’ and only these are retained in the list of putative gaps (**Figure 5.1.g**). Finally, using the distances of unmatched host nodes to the rest of the matched host nodes, new nodes of viral gaps are created in a new viral network (**Figure 5.1.h**).

Virogap can be reduced to four main components (**Figure 5.2**). First, host information is mapped onto all viral species. Second, viral protein clusters are created, optimized to contain homologous proteins across as many viruses as possible only once. In parallel, conserved host sequence is acquired and host proteins are clustered. Finally, gaps in the host cluster are scored, and high scoring gaps are mapped back to their relative position

in the viral cluster.

Networks and visualization

All relationships are described as all-against-all pairwise alignments using BLAST. The sequence similarity networks consist of nodes that represent protein sequences and edges that indicate sequence similarity [112].

All-against-all BLAST networks are visualized in Cytoscape, a powerful open-source platform for analyzing and visualizing large, complex networks [113]. Tabular BLAST output is converted to a collection of .SIF, .EDA, and .NOA files, where edge weights are determined by BLAST bit scores and node files describe the sequence record name and other relevant properties. In all cases, the edge-weighted force-directed layout was used to visualize the network.

Mapping virus to hosts

Public sequence repositories containing virus data vary greatly in their coverage and accuracy of host annotations. They are often under-annotated, incomplete, incorrect, or map a virus to broad host range. The first step to alleviating these inadequacies involves aggregating different sources, where more reliable sources take precedence over less reliable sources and more specific annotations would extend off of a general host frame. Genbank records, for example, note the host organism only at the discretion of the submitter; SWISS-PROT is well curated with nearly all of their virus records annotated

with standard nomenclature, but lacks entries for all known proteins; and the International Committee on the Taxonomy of Viruses (ICTV) also lists host range only for well characterized viruses, but this range is too broad to be useful.

SWISS-PROT

SWISS-PROT is a curated database that is a subset of UniProt Knowledgebase. The computer-annotated records derived from DDJI/Genbank/EMBL-Bank are stored in UniProt Knowledgebase and TrEMBLE. In SWISS-PROT, however, annotations are critically reviewed, albeit not entirely representative of all known viruses. These records are cross-referenced to Genbank records, which aids in the effort to consolidate annotations across multiple sources.

Genbank

This is the primary source of host annotations for viruses not included in SWISS-PROT. The ideal Genbank record includes a NCBI taxID for the virus host. In many cases, however, submitters have entered either the scientific name of the host species or a common name. In each of these instances, the NCBI released *names* file is used in an attempt to retrieve a NCBI taxonomy ID [114]. This file is an accumulation of all scientific names, common names, and common misspellings for all organisms mapped to unique taxonomy IDs.

Reference set of viral species

NCBI contains an up-to-date list of viral species, as even partial sequences of novel viruses are uploaded to Genbank. The NCBI taxonomy has identifiers for all viral families, genera, species, subspecies, etc. While the goal is to map all viral species to their hosts, there are instances in NCBI where the sequence of the virus is kept only with the subspecies or isolates, and in some cases, these subspecies are misclassified as taxa of unknown rank when they should be species. An example of this is the clade under human herpesvirus 8, correctly listed as a species on NCBI taxonomy. Its children nodes, however, are other gammaherpesvirus-2 viruses from other primates, including chlorocebus rhadinovirus-1 and gorilla rhadinovirus-1. NCBI taxonomy is based on a compilation of sources, and while fairly accurate, does require a bit of oversight to correct these exceptions. A script designed to handle these insufficiencies was created and run to adjust the reference list in a semi-automated process.

Filling in, propagating, and correcting incomplete host information

All dsDNA viral protein sequence from Genbank and Swiss-Prot was downloaded. Genbank sequence was filtered to remove patent sequence and third party annotated sequence. Virus sequence was added to the NCBI taxonomy-based scaffold using the NCBI taxIDs of their source virus, and each record was parsed for host annotations. When records are missing information, the *names* file from the NCBI taxonomy FTP site is used to obtain a scientific host name and a NCBI taxID. In cases where the virus species has no associated SWISS-PROT and Genbank records because all the sequences

are associated with subspecies and isolates, it is necessary to propagate host information from the subspecies level up to the species level. This is done in a supervised automated manner to moderate incorrect taxonomies.

The remaining virus species still lacking host annotations are manually annotated by literature surveying beginning with the Pubmed record linked to the sequence entry. A final consistency check was performed to make sure no incorrect host annotations were carried through the automated parts of assembling host annotations.

Viral clusters

All possible viral sequence is used in this analysis for the purpose of including all known viruses. Ideally, only Refseq sequence would be used, as these records have been curated, well-annotated, and are non-redundant [115]. The curation of Refseq sequences, however, are biased towards non-viral sequences, particularly vertebrates. The percentage of curated viral species is almost the lowest of all taxonomic groups at 17 percent (as of 2009), while 86 to 92 percent of vertebrate species are curated [115]. The power of Virogap would be greatly reduced with the inclusion of only 17 percent of viral species, so all Genbank sequence was used.

Obtaining all Genbank sequence under a taxon

All sequence under a specified taxonomy identifier is collected into a FastA file. This sequence includes all entries, including partial sequence, entire proteins, and complete genomes. Additionally, the FastA files have sequence from isolates and subspecies of viruses.

Markov Clustering

The Markov Clustering Algorithm, or MCL, is an unsupervised clustering algorithm developed by Stijn van Dongen that simulates flow between nodes of the network to determine cluster boundaries [116]. The granularity of clustering is determined by the inflation constant that represents the contraction of flow in the network and is used to convert the stochastic matrices used during the clustering process [117].

We optimized the clustering process by choosing the inflation constant that produces the largest clusters of singly-represented hosts. This promotes selection of clusters of single lineages that are most likely each comprising sequence from the same protein (e.g. DNA polymerase). In some cases, such as the rhadinovirus lineage, there are multiple lineages within the genus. Maximizing the cluster for the most unique host species allows differentiating between sublineages within

clades. It would almost be an ideal metric to evaluate the clusters by adding up the number of hosts that were represented in the cluster only once, but in some cases, this would give greater weight to clusters with fewer hosts represented but all hosts represented only once than clusters that covered a broad range of hosts, but had most of virus hosts represented more than once. We choose a metric more lenient than the former approach but also favors singly represented hosts by scoring the cluster as follows:

$$S = \sum_{i=1}^n \frac{1}{x_i}$$

where x_i is the number of times host i is represented by viruses in the cluster.

Choosing an inflation constant based on homogeneity of protein annotation would be another approach, but annotations are rarely included in the Genbank record, and when included, lack consistency of nomenclature between homologous proteins. The consensus can be used, however, to validate clustering with some certainty.

The initial networks were generated using an all-against-all BLAST with a loose e-value cutoff of 10^{-5} . This lenient BLAST cutoff was to allow distant homologs to cluster together, while spurious associations or clusters with multiple lineages would be separated when the MCL clustering algorithm is optimized to have maximal single instances of unique species.

Collapsing sequence to remove isolates and redundant sequence entries

To reduce redundancy and to find the most representative sequence for a group of isolates, we have developed an algorithm using clustering based on pairwise alignments (**Figure 5.3**). The *nrd* program developed by Holm and Sandler has a similar goal of obtaining representative sequence of clusters using pairwise alignments, but sets an identity threshold[118]. Similarly, this method creates clusters, but optimizes to maintain homogeneity of the source virus amount the clusters. Viruses have dissimilar sequence identity thresholds that define new species.

In the above algorithm described in **Figure 5.3**, the clusters are generated to maintain the same source virus across all its proteins. It is an iterative process, which creates clusters as each new sequence is added. Each cluster is associated with an e-value cutoff to maintain stringent consistency among all clusters. After the first loop through the pairwise distances, clusters are created that represent isolate sequence from the same virus, where one cluster may be the all ovine herpesvirus 2 DNA polymerase sequence (**Figure 5.4, maroon**) and another cluster may be ovine herpesvirus 2 glycoprotein B sequence (**Figure 5.4, blue**). In the case of no isolate sequence, a cluster can be just one single unique sequence (**Figure 5.4, brown**).

In the second iteration, a representative sequence is selected from each cluster. This sequence has the highest cumulative bit score with other nodes in the cluster (**Figure 5.4, yellow border**).

Finally, there are cases in which only partial sequence of a protein is included. These short sequences, often uploaded to Genbank as PCR amplicons, overshadow sequence similarity and drive the clustering process by length alone. Thus, all sequence that is more than a threshold below the median sequence length of the cluster is removed, and the virus host is noted to eliminate it as a predicted gap downstream.

Host clusters

Including all ‘unmatched hosts,’ (gH_m) or hosts without homologous viruses in the viral cluster requires finding all hosts that are within the phylogenetic or taxonomic limits described by the ‘matched hosts’ (H_m); in other words, the lowest common ancestor (LCA) taxonomic identifier that all n hosts H_n share should be the root node of all other hosts included in the host cluster. Obtaining all host sequence under the LCA taxonomic identifier puts only a loose inclusion criterion on putative hits, from which highest ranked gaps are selected based on sequence.

Conserved host sequence

Highly conserved sequence across a diversity of species has been described through the use of multiple sequence alignments of species groups [119]. There has also been an effort to classify entire clusters of orthologous proteins or domains. Some instances of conservation are only present in subsets of the species spectrum. For example, the 3'-UTR is a widely conserved region of vertebrate genomes due to a possible role in post-transcriptional regulation [119]. On the other hand, the ATP-binding subunit of ABC transporters is a domain that is only found in prokaryotic life.

Only using conserved sequence reduced the burden of needing to gather or store all possible host sequence, which ranges from vertebrates to yeast to bacteria and even archaea. For example, only about three to eight percent of the human genome shares conservation with other vertebrate genomes [119].

PFam

PFam comprises families of domains, or functional units of a protein, that can span species across the tree of life. [120] Cytochrome b is a mitochondrial protein coding gene that is used extensively for phylogenetic studies with its widespread presence across the Tree of Life (**Figure 5.5**) as well as its sequence variability between organisms [121]. The annotation framework allows unlimited appending of PFam host sequences to the mapping, so the choice of which host protein family can be changed to a protein other than cytochrome B if it is more appropriate to represent the desired host range.

Host clustering

Host clustering is performed after the LCA of matched hosts are determined, and again after only the top unmatched hosts are retained. The clustering is done using MCL from an all-against-all BLAST as described for viral clustering.

Gap evaluation: Incident matched edges

Putative gap scores are calculated by cumulating all incident bit scores from matched hosts $H_{1...N}$. A baseline value is determined based on the median cumulative incident bit scores for matched host H_m for all hosts $H_{1...N}$ when the network of all hosts under the LCA is considered. Only gH_m for all unmatched hosts above the determined threshold are considered, and rank ordered based on their gap scores.

Gap evaluation: Ratio of matched to unmatched hosts- density of subtrees

This is a similar concept to the above scoring, although it takes a more direct approach to comparing the number of matched hosts to the number of unknown hosts. This, as in the above scoring metric, are affected by both how well virus and host cospeciate, but also, to some degree, how many viruses have been discovered in this clade.

5.4. RESULTS

Index of viruses

The entire NCBI taxonomy structure for double-stranded DNA viruses was downloaded [122], and these taxonomy records were used as the scaffold to which all apposite

information was integrated. After correcting for incorrectly ranked virus species and subspecies, there were 1,894 species records to curate. Records of other taxonomic ranks were also included in all analysis to accommodate instances where sequence and annotation data were associated with records taxonomically below the virus species.

Curating virus host annotations

After filtering out third party annotated and patent sequence, 291,078 Genbank dsDNA virus protein records were obtained. At the time of procurement, there were 30,154 Swiss-Prot virus records, with 16,171 of the records belonging to dsDNA viruses. Upon incorporation of virus sequence to the NCBI taxonomy-based scaffold using the NCBI taxIDs of their source virus and host annotation distillation from these records, only 363 of the 1,894 virus species had host NCBI taxIDs, and an additional 842 species had some form of a host name (**Figure 5.6**). The NCBI *names* file was used to translate most of these host names to a NCBI taxID, resulting in an increase to 1,148 virus records with taxIDs, and only 57 that remained untranslated. Using the NCBI taxonomy hierarchy, host annotations and sequence were propagated upwards from isolates and subspecies to species records, with supervision. These increased the total number of records with host annotations to 1,276, with only 58 without NCBI taxIDs. In the final round of host mining, the names of the viruses were parsed for indications of a host, and these names were translated to NCBI taxIDs using the NCBI *names* file. This final parse resulted in 1,781 records with host annotations, and the rest were left for manual annotation from the literature.

Reducing sequence redundancy

Virus clustering was performed at the family, subfamily, and genus levels. Primary analysis was performed on *Herpesviridae*. All herpesvirus protein sequence under the family NCBI taxID, 10292, was obtained using a sequence lookup of Genbank records associated with this taxID and all taxIDs below it in the virus scaffold, thereby bypassing the slow process of querying NCBI. The high volume of information stored in records of often redundant viral isolates required reducing the sequence set into a nonredundant set using the clustering method described above (**Figure 5.3**). After collapsing and choosing the most representative sequence from each cluster, the herpesvirus dataset was reduced from 33,338 to 5,726 protein sequences.

Virus clusters

A range of inflation ($I=1.0..4.6$) constants for MCL clustering was considered, with steps of 0.6 (**Figure 5.7**). The three largest clusters generated at each step of inflation constant course were evaluated for the amount of host diversity. The highest host diversity score is cluster one, the DNA polymerase, at inflation constant 4.0 (**Figure 5.7b**), which at $s=76.77$ is slightly higher than the second highest diversity scores (77.75 for cluster one at $I=1.6$ to 3.4) for the largest cluster at other inflation constants and much higher than scores for the second and third largest clusters at almost any other inflation constant (**Figure 5.7c**). The diversity score is also 76.77 at inflation constant 4.6, but this conflict is resolved by staying with the more inclusive clustering scheme. The order of the

clusters vary minimally, except at inflation constant 1.0, where at such a lenient clustering threshold, the largest cluster is a mix of several different types of proteins, and the second largest cluster is DNA polymerase (**Figure 5.7a**).

After selection of the DNA polymerase cluster at inflation 4.0, the e-value threshold in the all-against-all BLAST must be calibrated in order to visualize the substructure of the network (**Figure 5.8**). The loose e-value used in MCL clustering at 10^{-15} shows the same knot of sequences seen in the protein-wide herpesvirus network (**Figure 5.8a**), but at a much more stringent e-value threshold of 10^{-120} , the subfamilies of the network begin to appear (**Figure 5.8b**).

Even with previous efforts to isolate single lineages using MCL, it appears there are at least three subfamilies and distinct lineages within these subfamilies. This is an indication that starting with a lower virus taxID would be more informative. Since a range of taxIDs is selected through the taxonomy hierarchy, the *rhadinovirus* genus of the subfamily *gammaherpesvirinae* is also clustered (**Figure 5.9**). The host diversity score remains unchanged across all inflation constants from 1.0 to 4.6 (step=0.6) (**Figure 5.9c**), so the most inclusive clustering is chosen at $I=1.0$ (**Figure 5.9a**).

Partial sequence

The *rhadinovirus* DNA polymerase cluster appeared to have a small number of partially sequenced records, where only a portion of the DNA polymerase was included (**Figure**

5.10). These sequences were mostly rat herpesviruses and appeared to be uploaded to Genbank from a common source. The short sequence was eliminated and their hosts were noted in the collection of ‘matched hosts,’ so they would not be considered as gaps in downstream analysis.

Host PFAM sequence

PFAM sequence from cytochrome b-N terminal was downloaded, comprising 70,453 total sequences. The file was redundant, with more than one cytochrome b sequence for each organism, so a unique sequence FastA file was created, containing 26,703 sequences. All sequence was added to individual hosts of the virus to host map.

Virus to host mapping and gap recognition

Hosts of all included *rhadinoviruses* were looked up in the virus to host map, and an all-against-all BLAST of this FastA file was performed to generate a sequence similarity network (**Figure 5.11**). This network was linked to the *rhadinovirus* DNA polymerase sequence similarity network using loose edge weights. The host organisms ranged from the expected primate species, including mandrillus, macaques, chimpanzees, and gorillas, but also included hosts as far reaching as the bank vole, feral cattle, and the wild boar. The lowest common ancestor of the *rhadinovirus* hosts was the very general taxon Eutheria, and all PFAM sequence under this taxon was culled. 7,442 species lie under the taxon Eutheria, of which 7,417 are unmatched hosts. A distance matrix was created to describe all-against-all BLAST alignment scores between the 7,442 matched and

unmatched species, and the cumulative bit scores of incident edges between each node and matched species nodes was calculated. This loose inclusion criterion allows all possible unmatched hosts, including those that are minimally connected to matched hosts in the graph. The distribution of matched hosts to other matched hosts was used as a reference of connectivity. The median value, upper quartile, top n percent, or highest value of cumulative bit score distances between matched host nodes and other matched host nodes can be further applied as a threshold to reduce the set of unmatched candidate host gaps. The median was 2749.5, and upper quartile at 3569.6. Use of the upper quartile threshold permitted 167 unmatched hosts to be included as high ranking gaps. A rank ordering of these gaps was produced, and only the top 20 unmatched hosts are shown in the *rhadinovirus* host network for the sake of visualization (**Figure 5.12**). The median score in this set was 5,211. Most of these hits are primates, reflecting the predominance of matched primate hosts, and the highest ranking gap is *Macaca leonina*, a species in the *Macaca* genus that includes the host of *Macaca nemestrina* rhadinovirus. *Macaca leonina* was traditionally cast as a subspecies of the species *Macaca nemestrina*. Aside from primates, some interesting gaps include the southern bog lemming, which is more genetically similar to voles, the host of the vole rhadinoviruses, *Microtus agrestis* rhadinovirus and *Myodes glareolus* rhadinovirus.

5.5. DISCUSSION AND FUTURE DIRECTIONS

Virogap is a hypothesis-generating program for identifying targets for virus discovery using the basic principle that viruses coevolve with their hosts. Viruses are first mapped

to their respective hosts using a course of annotation mining. Sequence similarity networks using all-against-all pairwise distances in viruses are drawn in tandem with the sequence similarity networks of their corresponding hosts. And finally, hosts without corresponding viruses are rank ordered as gap candidates, and it is proposed that viral orthologs in these hosts exist with sequence similarity to the original known virus set.

This method of gap-filling in viral trees is also revealing to viral evolution and metagenomics in two respects. First, not all viruses coevolve with their hosts, and viruses that cospeciate with their hosts are evident when sequence similarity networks are compared. Second, virus discovery has been greatly skewed towards diseases in humans and organisms humans rely on for food, and this approach of gap-recognition would unearth gaps in the virus tree of life that are likely present only because they have been overlooked.

The rationale behind Virogap has been alluded to in specific, small-scale discovery efforts (Chapter 2), but no effort to date has been made to scale this methodology to be applicable to all viruses, nor has there been a formalized scheme for gap prediction. While the example used for demonstrative purposes in this dissertation pertains to herpesviruses, host annotations have been gathered for all dsDNA viruses, and the abstract implementation can be generalized for all viruses once host annotations are curated for ssDNA, RNA viruses, and retroviruses. Double-stranded DNA viruses were chosen as an initial starting point because of genetic stability and tendency to maintain a

persistent lifestyle in the host. Herpesviruses, for example, are known for their ability to establish lifelong latency in hosts. Certain ssDNA and RNA viruses are also known to maintain persistence in their hosts, and thus tend to coevolve highly with their hosts. Multiple hosts are allowed in this framework, so viruses with a propensity for host switching would not require modification of host annotation extraction.

Acquiring all host annotations for viruses was the most cumbersome, requiring extraction of NCBI taxIDs for the majority of virus entries. Most of the annotation gleaning was largely automated, except for 113 dsDNA viruses that had to be manually annotated. Updating the host annotations should build off of current annotations rather than rerunning the pipeline to avoid the final step of manual annotation. Instead, new viruses should be appended to the dataset, and recently added records for existing viruses should be mined for host data to cross check current annotations.

Next, the inclusion of all protein sequence rather than just Refseq curated sequence introduces a number of issues concerning inadequate sequence. As of 2009, Refseq only contains sequence for 17 percent of viral species, and restricting the dataset would generally increase sequence annotation quality but severely limit the power and utility of this study. Non-Refseq records not only do not require a standardized host field, but also include partially sequenced genes. In the case of *rhadinoviruses*, there was a network of partial DNA polymerase sequence that was visible through Cytoscape visualization of the sequence similarity network (**Figure 5.10**). Currently, all sequence that is greater or less

than 30 amino acids from the median length of the cluster is removed from the network, but this metric should further investigated to gauge the amount of variation in sequence length of full-length proteins.

Because gaps are rank ordered, the threshold for gap calling is scalable according to a user defined ceiling. Characterizing the distribution of cumulative distances between matched hosts and other matched hosts in the network of all hosts under the LCA can be used as a reference for a lower bound gap cutoff. Using the upper quartile of matched to matched hosts in *rhadinovirus* allowed only 167 of the 7,417 unmatched species, and the top 20 gaps that were visualized in Cytoscape were well above the matched to matched upper quartile threshold. Most of these high ranking gaps were in primate hosts, reflecting the preponderance of known *rhadinovirus* primate hosts. Sequence proximity to matched hosts was the underlying force for these gap predictions, and while sequence similarity and taxonomy are not always correlative, the taxonomically adjacent primate species were called as high scoring gaps.

This method of gap recognition in viral sequence similarity networks was demonstrated in *rhadinovirus*, but a formalized estimate of the ability to predict virus gaps should be performed for each network created. The value of a leave-one-out analysis is variable across viral families, and even within viral families, the ratio of matched hosts to unmatched hosts varies drastically. The most recent viruses discovered from each network could each be left out and the score of that virus could be calculated. This score

depends heavily, however, not only to how well it fits in the family host range, but also how closely related it is to the majority of matched virus hosts; in other words, if the left out virus host happens to be one of the unexpected outlying hosts, it may underestimate its gap score. A more comprehensive approach would be to leave-one-out for every virus-host in the network, but as linked networks are created for each case, this would be computationally expensive.

An valuable feature that would assist in estimating gap significance would be an implementation of cophylogeny estimation, which can be approximated by comparing virus and host networks. One way of correlating network similarity would be to calculate the sequence similarity distances between pairs of nodes between the two groups, where a correlation coefficient between sum normalized distances of viruses and their hosts is computed. This gets a bit more complicated with viruses with multiple hosts. The hosts of the virus can be collapsed into a single node, with distances to the metanode being an average of the sub-nodes.

Another critical feature would be to approximate the location of the gap in the viral network. This would allow for the extraction of homologous sequence from adjoining viral nodes of the virus gap. First, all adjacent matched host nodes to the unmatched host gap must be retrieved. Viruses assigned to these matched hosts are used as guides for placement of the virus gap, with edge values scaled off the distances between the host gap and other matched hosts. These estimated virus gaps (**blue**) can be used to generate

the network described in **Figure 5.1.h**.

A secondary feature to add would be the ability to query gaps by host species or taxon, which would allow host or host range specific gaps to be used for virus discovery applications. Adjacent viral sequence to all virus gaps pertaining to a particular host could be selected for microarray design or create viral profile HMMs.

These profiles can be used to mine high-throughput sequencing data. Once high-scoring virus gaps are identified across all viral families, a reevaluation of ‘high-yield’ metagenomic samples such as the AIDS saliva in Chapter 2 should be performed for gap-centric virus discovery.

5.6. ACKNOWLEDGEMENTS

I would like to thank Patsy Babbitt for her always insightful advice and suggestions. I would also like to acknowledge Holly Atkinson for her initial introduction to sequence similarity networks.

FIGURES

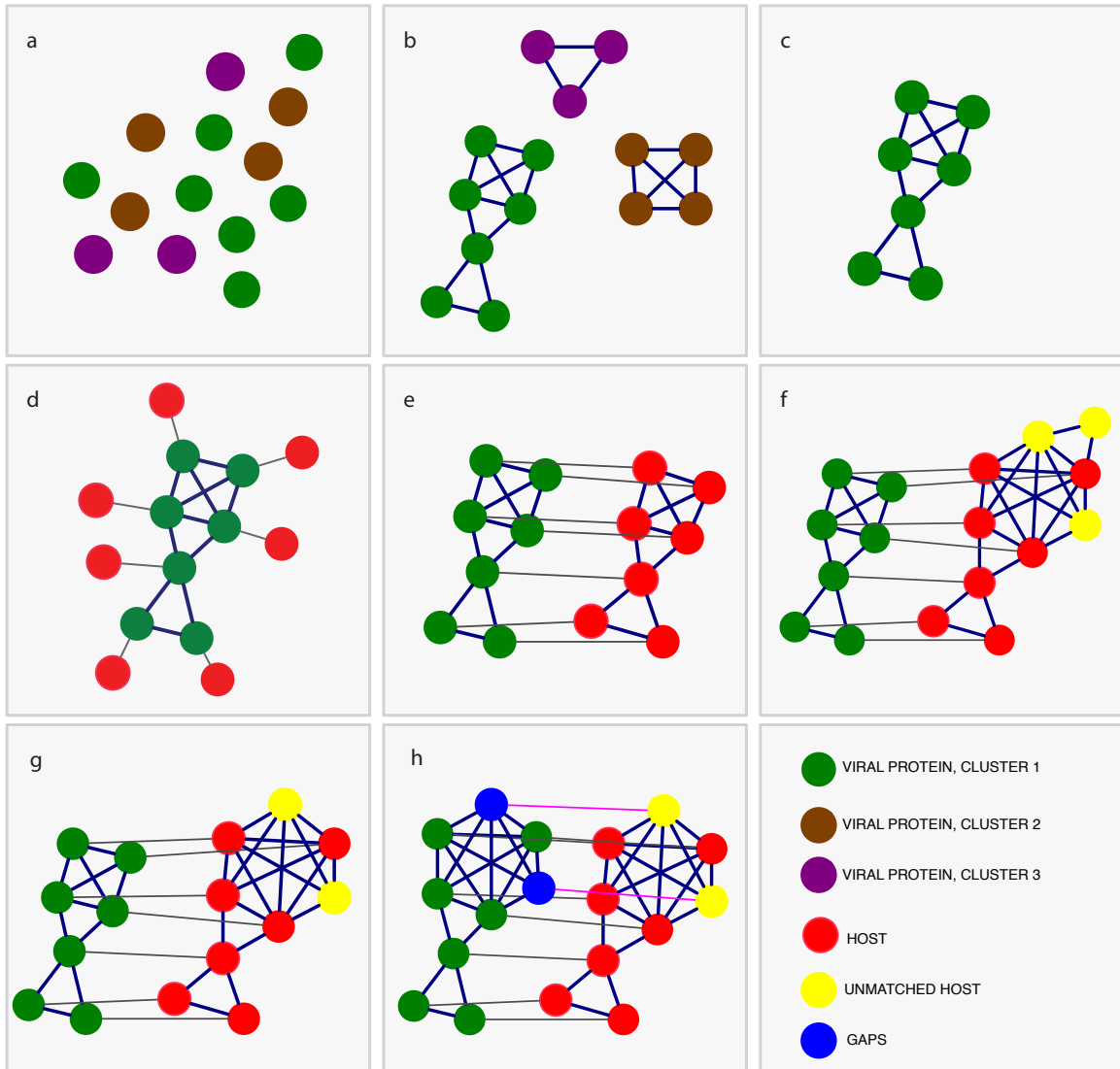


Figure 5.2. Virogap . Viral protein sequence (green, brown, purple) is clustered using MCL with all-against-all sequence similarity (b). The largest cluster is selected (c), and viruses are mapped to their hosts (d). Host sequence is then clustered (e), and unmatched host sequence is introduced when all organisms under the LCA of matched hosts is used as a lenient inclusion threshold (f). Only unmatched hosts that are well-connected with matched hosts are retained (g), and these gaps are mapped back to their relative position in the virus network (h).

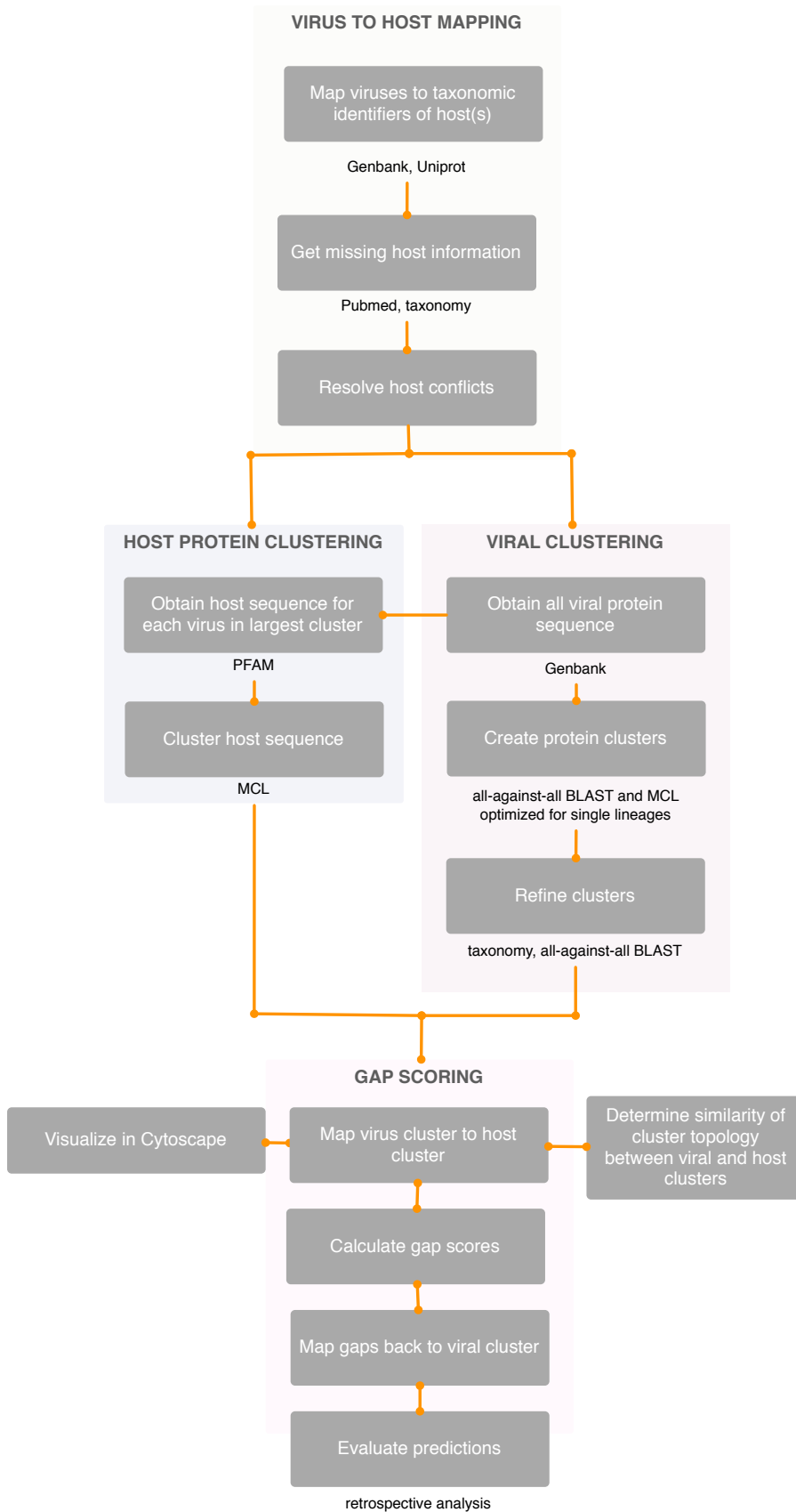


Figure 5.2. Virogap pipeline. Virus to host mapping (orange) is followed by host protein clustering (blue) and viral clustering (brown) performed in parallel, and finally gaps are recognized and scored (pink).

```

q - query sequence
s - subject sequence
Cq - Cluster containing q
Cs - Cluster containing s

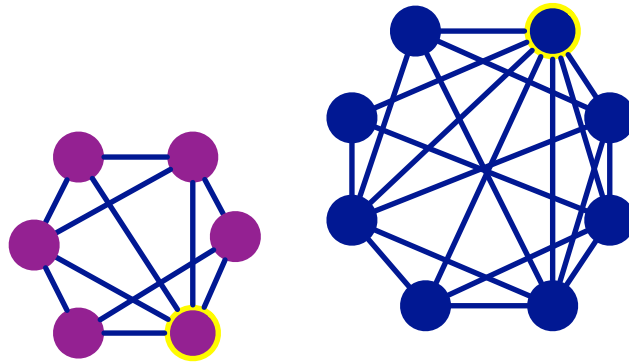
For GI q and GI s of all-against-all pairwise alignments for N nodes in order of increasing e-value:
  If neither q nor s are in a cluster:
    If the same source virus:
      Combine into a single new cluster.
    If different source viruses:
      Create separate clusters.
      Reset the e-value cutoff to e-value between q and s.
  Else if both q nor s are in clusters:
    If they are in different clusters:
      If the e-value is less than the e-value of one of the
      clusters:
        Move q or s into the same cluster.
  Else if q is in a cluster but not s:
    If the same source virus and e-value is not greater than cluster of q.e-value:
      Add s to Cq
    Else:
      If e-value is less than Cq.e-value:
        Cq.evalue = e-value
  Else if s is in a cluster but not q:
    If the same source virus and e-value is not greater than cluster of s.e-value:
      Add q to Cs
    Else:
      If e-value is less than Cs.e-value:
        Cs.evalue = e-value

For each cluster C:
  Retain only the node that has the strongest edges to all other nodes (bit score).

```

Figure 5.3. Clustering algorithm to reduce redundant viral sequence while maintaining source virus homogeneity.

Ovine herpesvirus 2 glycoprotein B



Ovine herpesvirus 2 DNA polymerase



Human herpesvirus 2 glycoprotein B

Figure 5.4. Collapsing isolates using iterative clustering modified to maintain source virus homogeneity within clusters. Sequence (nodes) with the highest degree of incident edges are selected as the representative sequence of the cluster (yellow border).

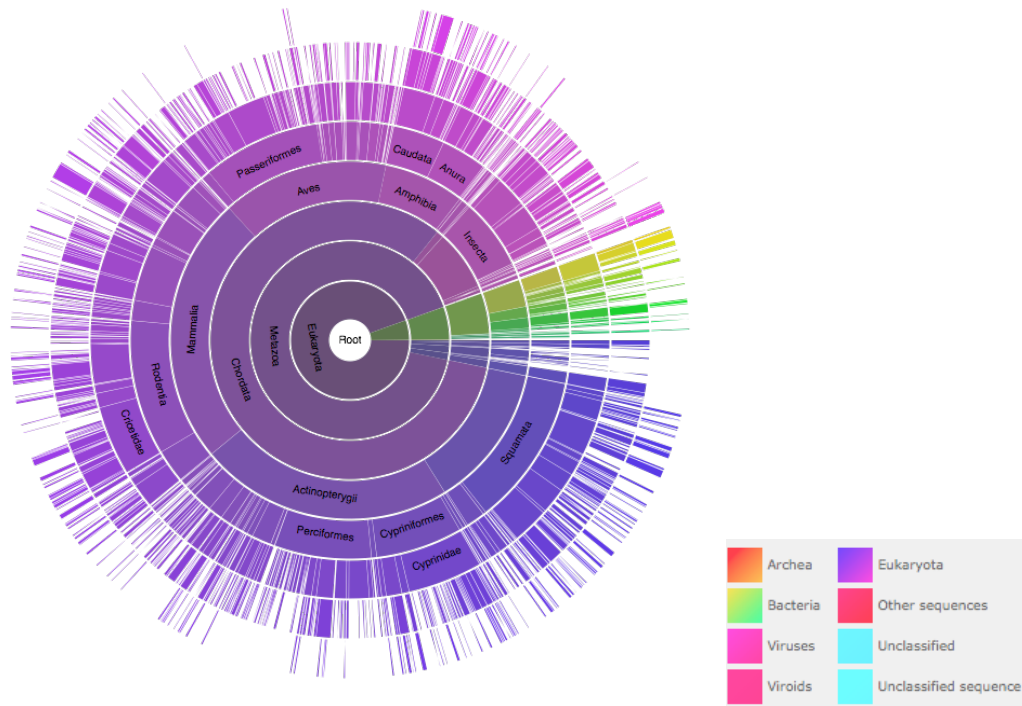


Figure 5.5. The universality of cytochrome B –N terminus across the tree of life, image by PFAM [120]. Segments are weighted by the number of sequences in the taxon group.

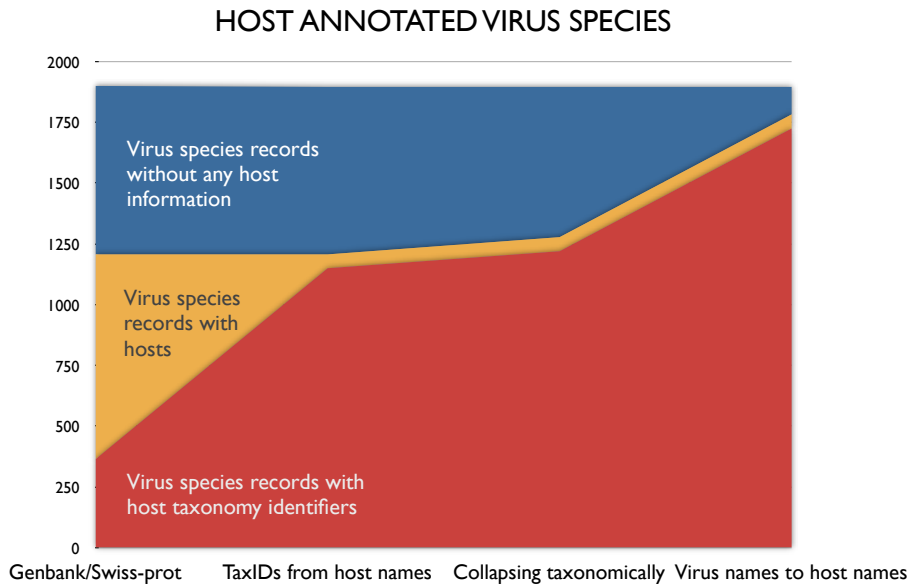


Figure 5.6. dsDNA virus species that have adequate host annotations after each iteration of virus host curation. Genbank and Swiss-prot records are used for the initial first pass. Host taxonomic identifiers are then filled in by translating the common, scientific, or misspelled name to a unique NCBI tax ID using the *names* lookup. Host annotations are propagated upwards from isolate and subspecies records to species records in a supervised automated manner. Next, records lacking host information are processed for indication of host within their virus names. Finally, the small remaining number of viruses lacking host data are manually annotated through literature review.

capsid proteins,
transcription regulators,
glycoproteins

DNA polymerase glycoprotein B

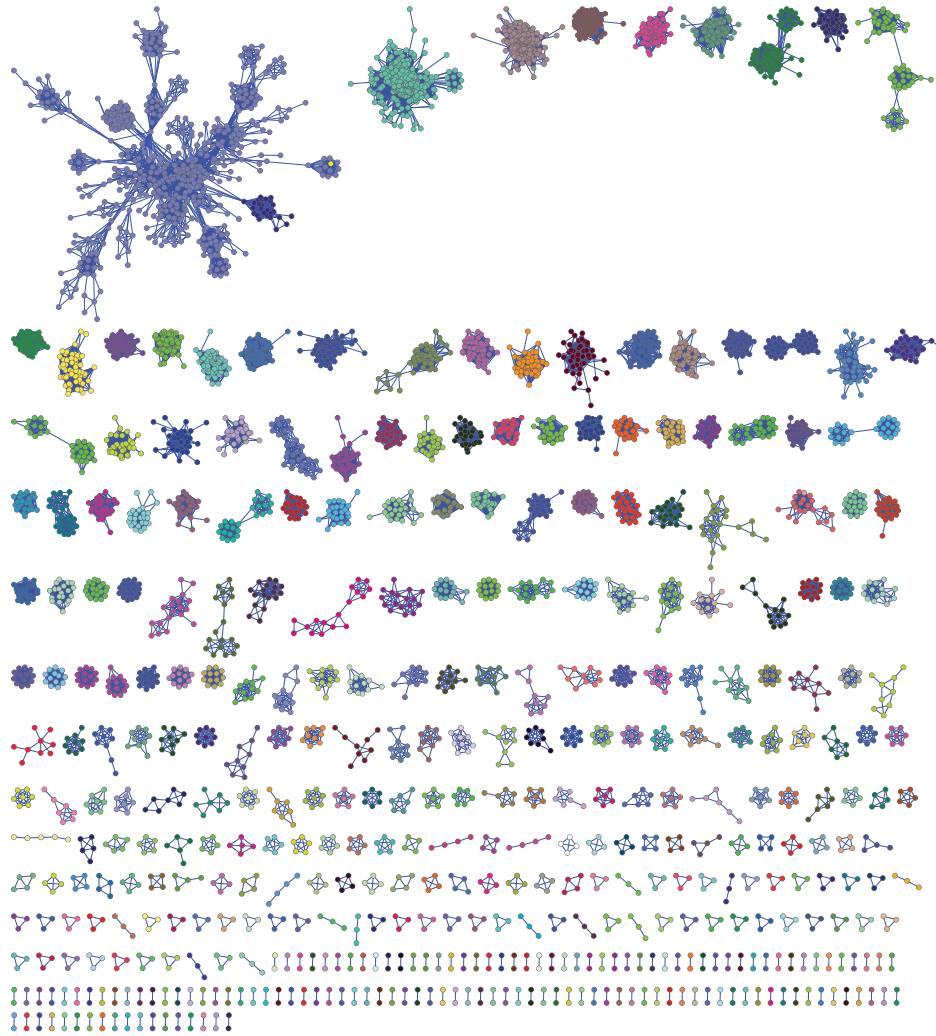


Figure 5.7a. All-against-all MCL clustering of all herpesvirus protein sequence (I=1.0).

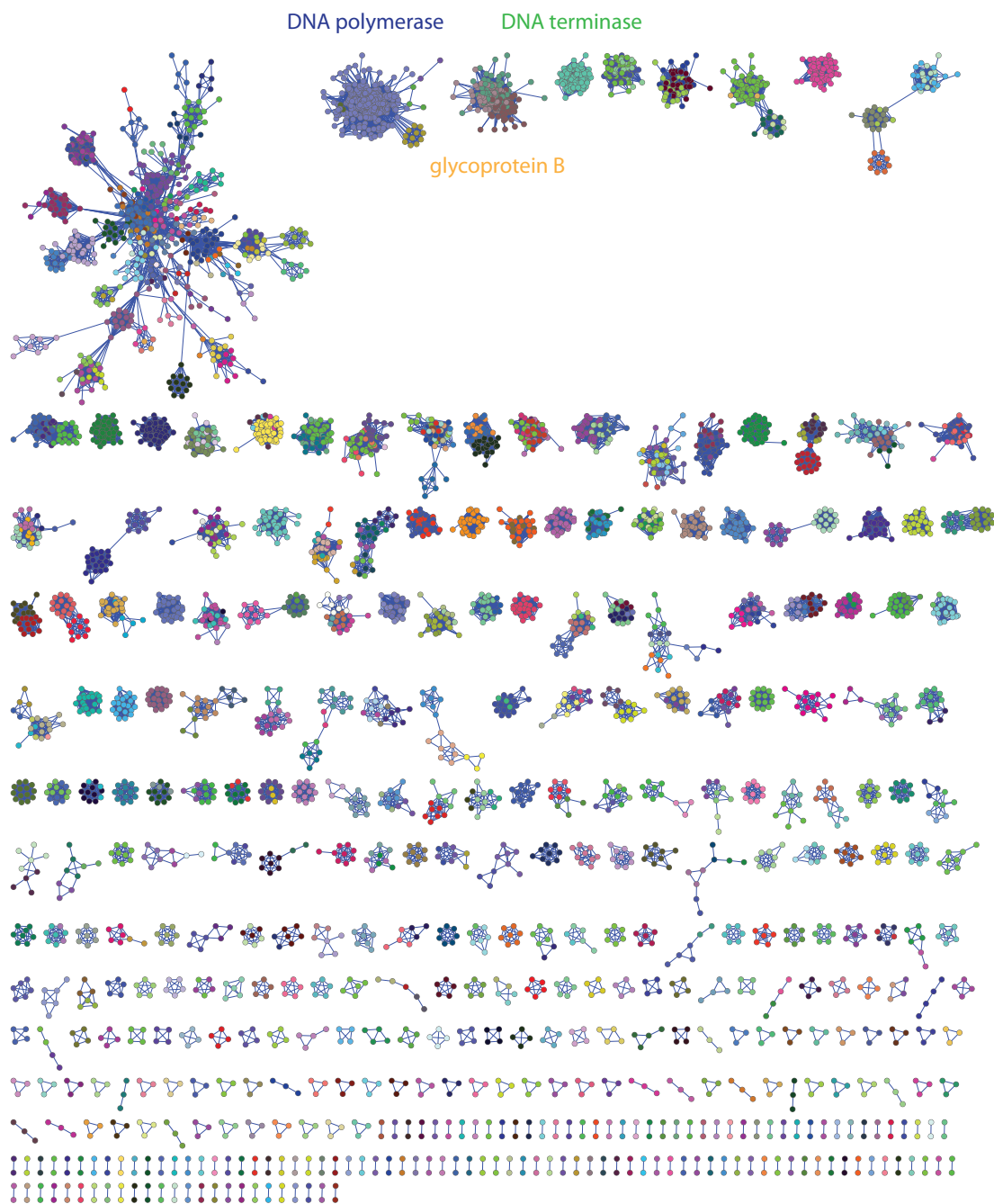


Figure 5.7b. All-against-all MCL clustering of all herpesvirus protein sequence

(I=4.0).

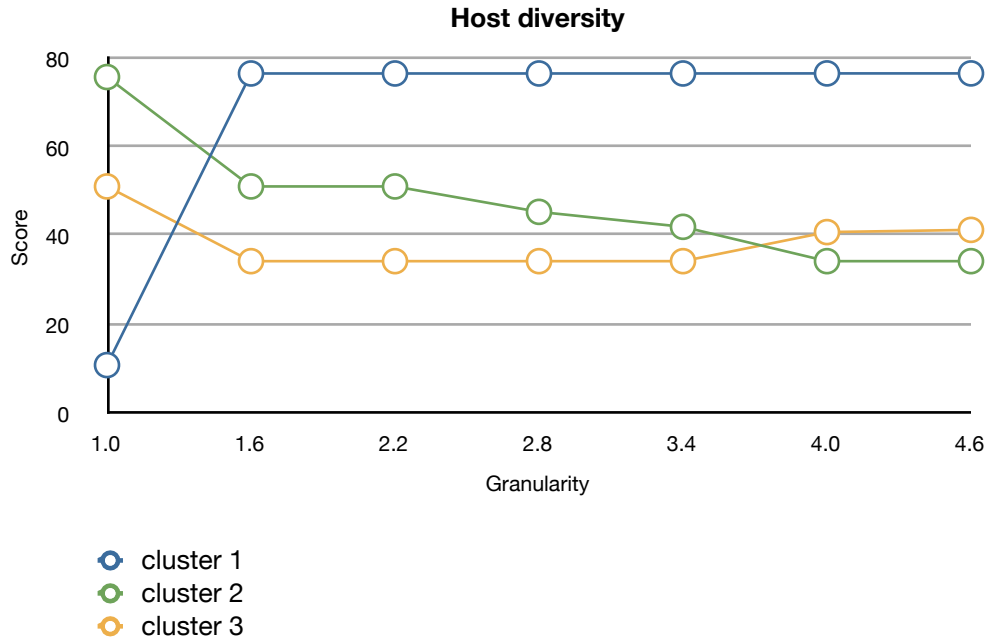
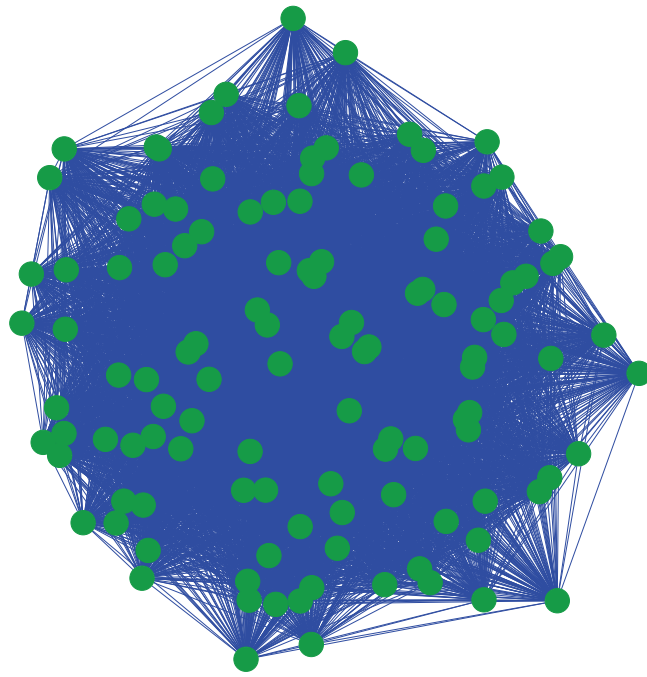


Figure 5.7c. Host diversity for three largest protein clusters (blue, green, gold) over a range of MCL inflation constants. The three clusters are identified by equivalently colored text in Figures 5.7a and 5.7b. The host diversity score is calculated as follows:

$$S = \sum_{i=1}^n \frac{1}{x_i}$$

where x_i is the number of times host i is represented by viruses in the cluster.

a



b

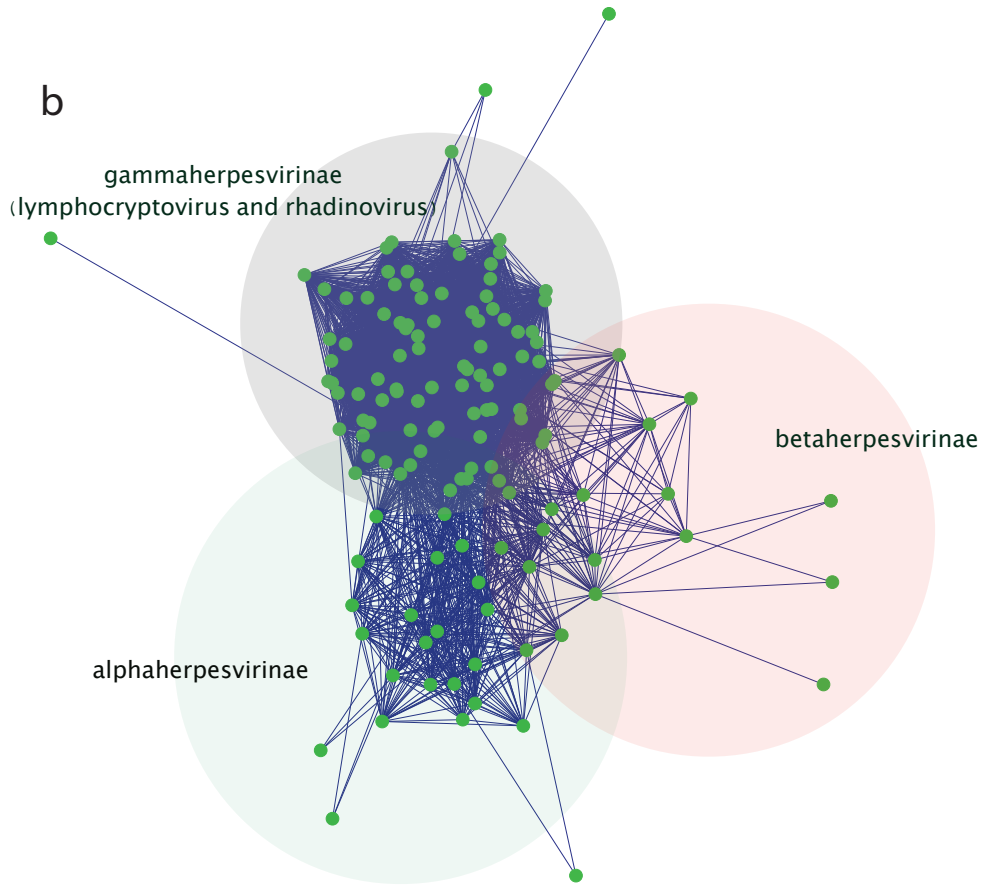


Figure 5.8. All-against-all sequence similarity network of herpesvirus DNA polymerase with an e-value of 10^{-15} (a) and 10^{-120} (b). Herpesvirinae subfamilies alpha-, beta-, and gammaherpesvirinae are only apparent in the latter network using a lower e-value threshold. The subfamilies are shaded accordingly in (b).

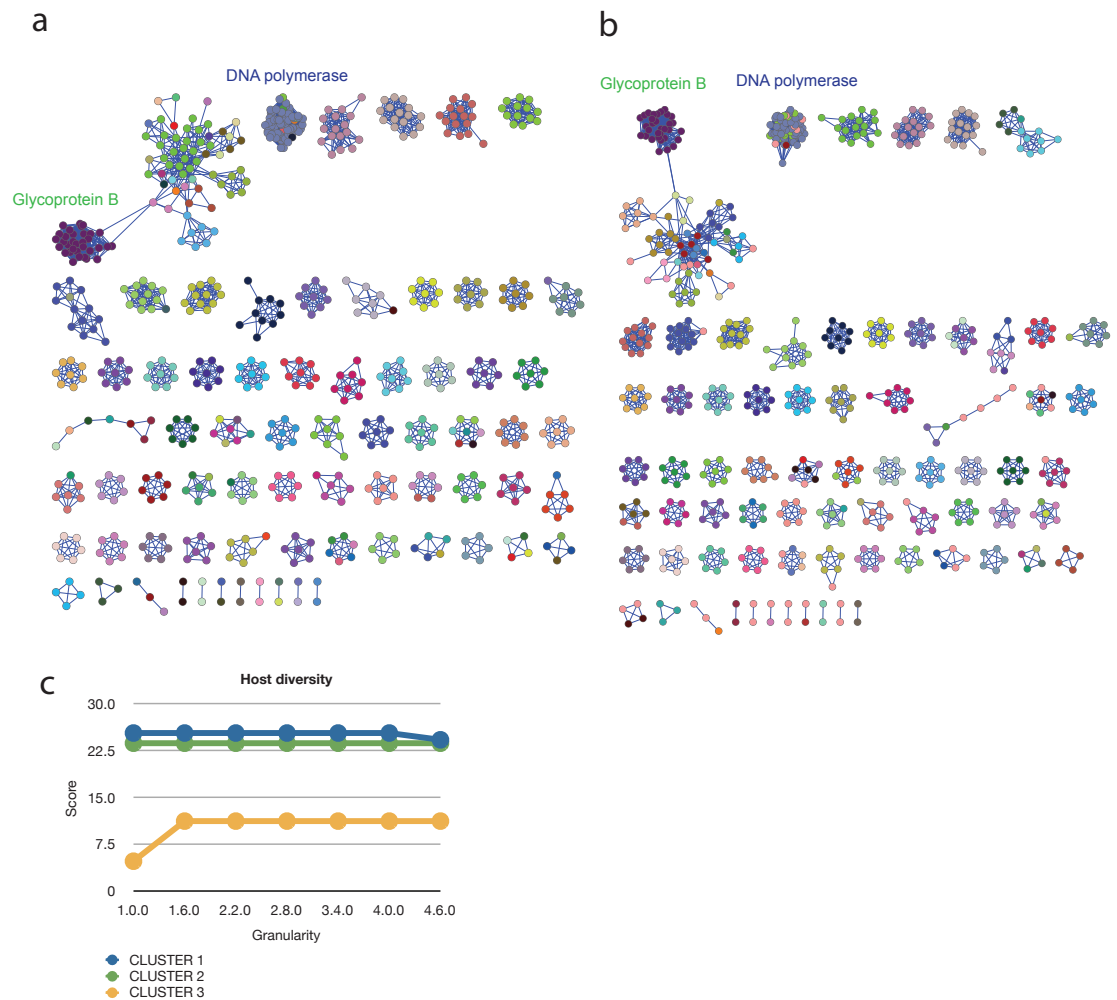


Figure 5.9. Sequence similarity networks for rhadinovirus proteins using an MCL granularity constant of 1.0 (a) to 4.6 (b). Host diversity, shown in (c) for the three largest clusters, is calculated as previously described.

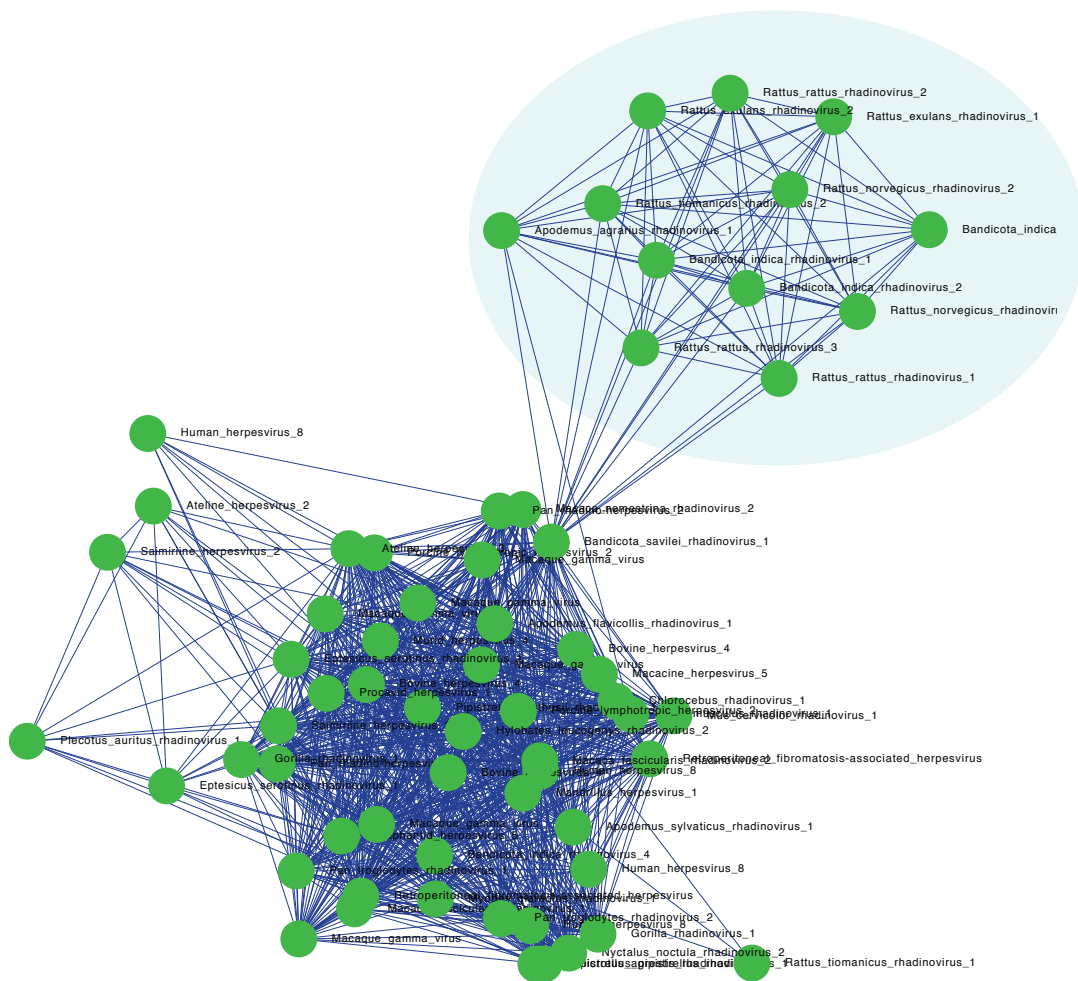


Figure 5.10. Sequence similarity network of rhadinovirus DNA polymerase using an e-value threshold of 10^{-15} . Partial DNA polymerase sequence is shaded in teal.

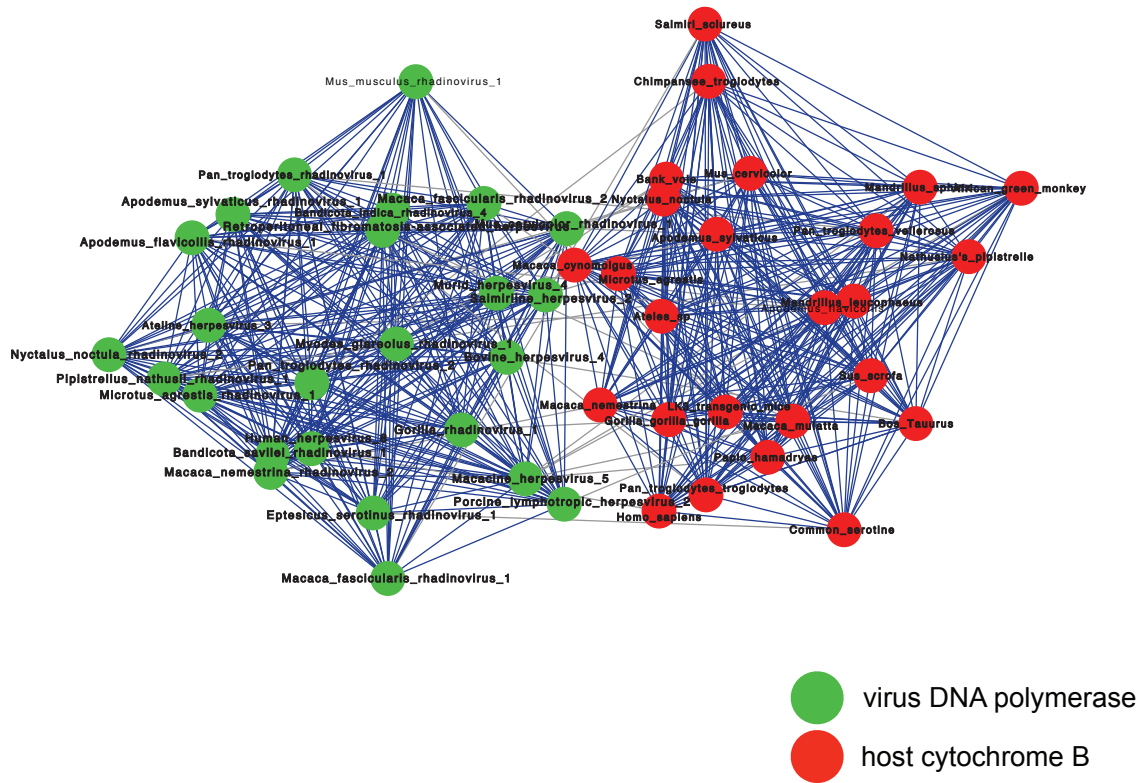


Figure 5.11. Host cytochrome b sequence (red) clustered in a sequence similarity network next to DNA polymerase from rhadinoviruses (green). The grey lines denote virus to host mappings. Host organisms ranged from primate species--mandrillus, macaques, chimpanzees, and gorillas--to hosts as far reaching as the bank vole, feral cattle, and the wild boar. The lowest common ancestor of the *rhadinovirus* hosts was the taxon Eutheria.

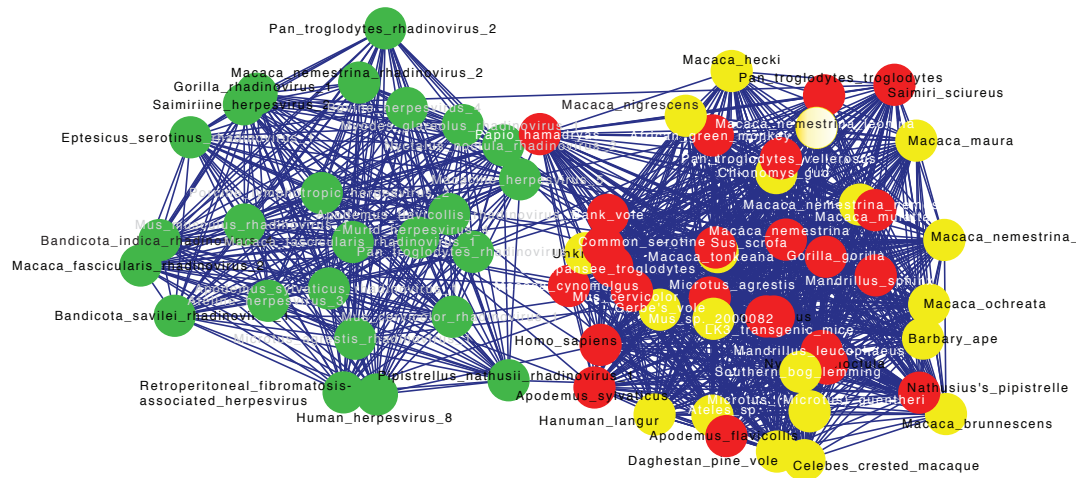


Figure 5.12. Top 20 unmatched hosts (yellow), evaluated by the sum of distance scores to matched hosts (red). Most are primates, reflecting the predominance of matched primate hosts. The highest ranking gap is *Macaca leonina* (white), a species in the *Macaca* genus that includes the host of *Macaca nemestrina* rhadinovirus. *Macaca leonina* was traditionally cast as a subspecies of the matched host *Macaca nemestrina*. Aside from primates, some interesting gaps include the southern bog lemming, which is more genetically similar to voles, the host of the vole rhadinoviruses, *Microtus agrestis* rhadinovirus and *Myodes glareolus* rhadinovirus.

TABLES

TABLE 5.1: VIRAL LIFESTYLES

ACUTE	PERSISTENT
↑ transient	↓ transience. not fully cleared
↑ virulence	↓ virulence
↑ transmissibility	↓ transmissibility
↑ wide range of hosts	↓ number of host species
↓ degree of cospeciation	↑ degree of cospeciation

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

Pathogen discovery in humans is an extensively canvassed field, where advances have been highly dependent on the headway made in molecular techniques and next-generation technologies. Traditional methods of pathogen detection have their limitations. In turn, microarrays and next-generation sequencing technologies have allowed for the detection of divergent viruses or scant amounts of virus, and have responded to many of the challenges that left viruses out of the initial metagenomics movement.

A few approaches to pathogen detection are explored in this dissertation--targeted viral discovery, disease-focused viral discovery, and viral metagenomics. Targeted viral discovery is of particular interest, as this approach can be broadened and applied as a scheme to generate hypotheses of viral discovery targets. Chapter 2 is an implementation of targeted viral discovery, where a gap in a specific lineage of herpesviruses is proposed, and all efforts of detection are focused on this gap. Construction of a phylogenetic tree of all gamma-2 herpesviruses revealed mirrored clades, which both comprise primate herpesviruses but only one of which includes a human herpesvirus, Kaposi's sarcoma-associated herpesvirus (KSHV). It was hypothesized that the clade lacking a human herpesvirus equivalent had an undiscovered virus that, like other gammaherpesviruses, caused lymphotropic disease and reactivated in immunocompromised hosts more frequently. These phenotypic motifs were used to select sample sets that would bias our efforts towards a greater likelihood of discovering a novel human herpesvirus. Saliva

from severely immunocompromised AIDS patients collected and banked prior to the distribution of antiretroviral therapies was examined by viral microarray and high throughput sequencing. This “high-yield” approach was motivated by attributes of the herpesvirus lifestyle--herpesviruses are known to be secreted in saliva, and they are persistent viruses that establish latency, thereby evading immunodetection. However, they are more likely to actively replicate when the host has a compromised immune system. Tonsillitis was another sample set chosen solely for the purpose of lymphotropic virus detection, as tonsillar tissue is lymphocyte-rich. Again, to increase the likelihood of detection, harvested cells were treated with PMA to induce the lytic cycle of any latent viruses present. In spite of using two sets of samples of known tropism and some of the most sensitive methods of detection, no novel human herpesvirus was detected. Finding a single large double-stranded DNA virus that has evaded detection through the discovery and characterization of eight other human herpesviruses is a difficult and risky undertaking, and this targeted approach to discovery yielded a more generalized approach to virus discovery using cophylogeny between virus and host.

The syndrome-centered approach to viral discovery is reproduced in Chapter 3 and 4, when idiopathic subsets of three disparate diseases--uveitis, acute liver failure, and acute exacerbation of idiopathic pulmonary fibrosis--are investigated for prospective viral etiologies. Divergent and unexpected viruses were discovered with the use of viral microarrays and high-throughput sequencing, and the pathogenic involvement of these viruses should be further studied. The findings of viruses in acute exacerbation of

idiopathic pulmonary fibrosis were published in American Journal of Respiratory and Critical Care Medicine (AJRCCM).

Finally, this dissertation circles back in Chapter 5 to the motivation behind the targeted herpesvirus approach taken in Chapter 2. The inferred human herpesvirus gap in the gamma-2 herpesvirus clade is used as a paradigm for future discovery projects. This chapter describes a method of gap-recognition in virus phylogenetic trees for motivating chip-based and high-throughput sequencing-based viral metagenomics. We begin by curating a map between viruses and their respective host species. This, in turn, is used to create a minimum spanning tree of host species from the lowest common ancestor of viruses represented in a subtree of orthologous viruses. We then find nodes represented in the host species tree and not the virus tree, and these gaps subsequently can be used to guide our experimental methods of virus discovery. It is our hope that this method will drive more detection and discovery applications designed with the directed approach used to describe the gap in herpesviruses. Additionally, with the scalability of this approach, new platforms such as host-specific pan-viral or gap-centric microarrays could be designed to aid in novel virus discovery.

As virus discovery in humans becomes increasingly surveyed, advances in the field will rely on a few paramount systems. First, advances in high-throughput sequencing and bioinformatic tools for sequence analysis will allow for detection and *de novo* assembly of novel viral genomes. Second, using virus gap information for targeted discovery will

allow for shrewd sample selection and collection and informed techniques to optimize discovery conditions. And finally, further detection of viruses in non-human hosts will not only add power to virus gap prediction in humans, but is also essential to gain insight to viral evolution and diversity. The methods, viruses and disease, and tools chronicled in this dissertation narrate a convergence of technology, molecular microbiology, and bioinformatics; when applied to a clinical setting, these systems can be used to study viral metagenomics and etiologies of some of the most devastating diseases.

REFERENCES

1. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* 99(24): 15687-15692.
2. Moore PS, Chang Y. (2010) Why do viruses cause cancer? highlights of the first century of human tumour virology. *Nat Rev Cancer* 10(12): 878-889.
3. Morse SS, Schluenderberg A. (1990) From the national institute of allergy and infectious diseases, the fogarty international center of the national institutes of health, and the rockefeller university. *emerging viruses: The evolution of viruses and viral diseases. J Infect Dis* 162(1): 1-7.
4. Storch GA. (2000) Diagnostic virology. *Clin Infect Dis* 31(3): 739-751.
5. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, et al. (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated kaposi's sarcoma. *Science* 266(5192): 1865-1869.
6. Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, et al. (2008) Identification of cardioviruses related to theiler's murine encephalomyelitis virus in human infections. *Proc Natl Acad Sci U S A* 105(37): 14124-14129.
7. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, et al. (2008) Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: Identification of a candidate etiologic agent. *Virology* 375: 88.
8. Hudnall SD, Chen T, Tyring SK. (2004) Species identification of all eight human herpesviruses with a single nested PCR assay. *J Virol Methods* 116(1): 19-26.
9. Alba MM, Das R, Orengo CA, Kellam P. (2001) Genomewide function conservation and phylogeny in the herpesviridae. *Genome Res* 11(1): 43-54.
10. Ruby JG. Paired-read iterative contig extension (PRICE). .
11. Metzker ML. (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1): 31-46.
12. Feng H, Shuda M, Chang Y, Moore PS. (2008) Clonal integration of a polyomavirus in human merkel cell carcinoma. *Science* 319(5866): 1096-1100.

13. Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, et al. (2010) Human enterovirus 109: A novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua. *J Virol* 84(18): 9047-9058.
14. Mindell DP, Rest JS, Villarreal LP. (2004) **Viruses and the tree of life**. In: Cracraft J, editor. **Assembling the tree of life**. . pp. 107.
15. Suttle CA. (2005) Viruses in the sea. *Nature* 437(7057): 356-361.
16. Kinross JM, von Roon AC, Holmes E, Darzi A, Nicholson JK. (2008) The human gut microbiome: Implications for future health care. *Curr Gastroenterol Rep* 10(4): 396-403.
17. Hamady M, Knight R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* 19(7): 1141-1152.
18. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, et al. (2009) The NIH human microbiome project. *Genome Res* 19(12): 2317-2323.
19. Lovisolo O, Hull R, Rosler O. (2003) Coevolution of viruses with hosts and vectors and possible paleontology. *Adv Virus Res* 62: 325-379.
20. Iyer LM, Aravind L, Koonin EV. (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 75(23): 11720-11734.
21. Zanotto PM, Gibbs MJ, Gould EA, Holmes EC. (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J Virol* 70(9): 6083-6096.
22. C. Fauquet, International Committee on Taxonomy of Viruses, International Union of Microbiological Societies., editor. (2005) *Virus taxonomy: Classification and nomenclature of viruses : Eighth report of the international committee on the taxonomy of viruses*. London: Academic Press.
23. Gibbs AJ, Calisher CH, García-Arenal F. (1995) *Molecular basis of virus evolution* . Cambridge, UK: Cambridge University Press.
24. Holland JJ, De La Torre JC, Steinhauer DA. (1992) RNA virus populations as quasispecies. *Curr Top Microbiol Immunol* 176: 1-20.
25. Best SM, Kerr PJ. (2000) Coevolution of host and virus: The pathogenesis of virulent and attenuated strains of myxoma virus in resistant and susceptible European rabbits. *Virology* 267(1): 36-48.

26. Lucht E, Brytting M, Bjerregaard L, Julander I, Linde A. (1998) Shedding of cytomegalovirus and herpesviruses 6, 7, and 8 in saliva of human immunodeficiency virus type 1-infected patients and healthy controls. *Clin Infect Dis* 27(1): 137-141.
27. Sinclair J, Sissons P. (2006) Latency and reactivation of human cytomegalovirus. *J Gen Virol* 87(Pt 7): 1763-1779.
28. Jenner RG, Alba MM, Boshoff C, Kellam P. (2001) Kaposi's sarcoma-associated herpesvirus latent and lytic gene expression as revealed by DNA arrays. *J Virol* 75(2): 891-902.
29. Miller G, Heston L, Grogan E, Gradoville L, Rigsby M, et al. (1997) Selective switch between latency and lytic replication of kaposi's sarcoma herpesvirus and epstein-barr virus in dually infected body cavity lymphoma cells. *J Virol* 71(1): 314-324.
30. McAllister SC, Hansen SG, Messaoudi I, Nikolich-Zugich J, Moses AV. (2005) Increased efficiency of phorbol ester-induced lytic reactivation of kaposi's sarcoma-associated herpesvirus during S phase. *J Virol* 79(4): 2626-2630.
31. Endo LH, Ferreira D, Montenegro MC, Pinto GA, Altemani A, et al. (2001) Detection of epstein-barr virus in tonsillar tissue of children and the relationship with recurrent tonsillitis. *Int J Pediatr Otorhinolaryngol* 58(1): 9-15.
32. Paradise JL, Bluestone CD, Colborn DK, Bernard BS, Rockette HE, et al. (2002) Tonsillectomy and adenotonsillectomy for recurrent throat infection in moderately affected children. *Pediatrics* 110(1 Pt 1): 7-15.
33. Myoung J, Ganem D. (2011) Infection of primary human tonsillar lymphoid cells by KSHV reveals frequent but abortive infection of T cells. *Virology* 413(1): 1-11.
34. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-predict: A computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* 6(9): R78.
35. Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95(25): 14863-14868.
36. Wootton SC, Kim DS, Kondoh Y, Chen E, Lee JS, et al. (2011) Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 183(12): 1698-1702.

37. Ziv J, Lempel A. (1977) A universal algorithm for sequential data compression. *IEEE Trans* 23: 337-338-343.
38. Kent WJ. (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12(4): 656-664.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.
40. Okamura A, Yoshioka M, Kubota M, Kikuta H, Ishiko H, et al. (1999) Detection of a novel DNA virus (TTV) sequence in peripheral blood mononuclear cells. *J Med Virol* 58(2): 174-177.
41. Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, et al. (2010) Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J Virol* 84(4): 1674-1682.
42. Levings RL, Wessman SJ. (1991) Bovine viral diarrhea virus contamination of nutrient serum, cell cultures and viral vaccines. *Dev Biol Stand* 75: 177-181.
43. Ling PD, Lednicky JA, Keitel WA, Poston DG, White ZS, et al. (2003) The dynamics of herpesvirus and polyomavirus reactivation and shedding in healthy adults: A 14-month longitudinal study. *J Infect Dis* 187(10): 1571-1580.
44. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. (2005) Defining the normal bacterial flora of the oral cavity *J Clin Microbiol* 43(11): 5721-5732.
45. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, et al. (2009) Metagenomic study of the oral microbiota by illumina high-throughput sequencing. *J Microbiol Methods* 79(3): 266-271.
46. Saito I, Nishimura S, Kudo I, Fox RI, Moro I. (1991) Detection of epstein-barr virus and human herpes virus type 6 in saliva from patients with lymphoproliferative diseases by the polymerase chain reaction. *Arch Oral Biol* 36(11): 779-784.
47. Li L, Barry P, Yeh E, Glaser C, Schnurr D, et al. (2009) Identification of a novel human gammapapillomavirus species. *J Gen Virol* 90(Pt 10): 2413-2417.
48. McKaig RG, Baric RS, Olshan AF. (1998) Human papillomavirus and head and neck cancer: Epidemiology and molecular biology. *Head Neck* 20(3): 250-265.

49. Levi JE, Kleter B, Quint WG, Fink MC, Canto CL, et al. (2002) High prevalence of human papillomavirus (HPV) infections and high frequency of multiple HPV genotypes in human immunodeficiency virus-infected women in brazil. *J Clin Microbiol* 40(9): 3341-3345.
50. Hino S, Miyata H. (2007) Torque teno virus (TTV): Current status. *Rev Med Virol* 17(1): 45-57.
51. Deng X, Terunuma H, Handema R, Sakamoto M, Kitamura T, et al. (2000) Higher prevalence and viral load of TT virus in saliva than in the corresponding serum: Another possible transmission route and replication site of TT virus. *J Med Virol* 62(4): 531-537.
52. West KH, Bystrom JM, Wojnarowicz C, Shantz N, Jacobson M, et al. (1999) Myocarditis and abortion associated with intrauterine infection of sows with porcine circovirus 2. *J Vet Diagn Invest* 11(6): 530-532.
53. Gritz DC, Wong IG. (2004) Incidence and prevalence of uveitis in northern california; the northern california epidemiology of uveitis study. *Ophthalmology* 111(3): 491-500; discussion 500.
54. Islam SM, Tabbara KF. (2002) Causes of uveitis at the eye center in saudi arabia: A retrospective review. *Ophthalmic Epidemiol* 9(4): 239-249.
55. Lee WM. (1993) Acute liver failure. *N Engl J Med* 329(25): 1862-1872.
56. Lee WM, Squires RH, Jr, Nyberg SL, Doo E, Hoofnagle JH. (2008) Acute liver failure: Summary of a workshop. *Hepatology* 47(4): 1401-1415.
57. Quillen DA, Davis JB, Gottlieb JL, Blodi BA, Callanan DG, et al. (2004) The white dot syndromes. *Am J Ophthalmol* 137(3): 538-550.
58. Koch WC, Adler SP. (1990) Detection of human parvovirus B19 DNA by using the polymerase chain reaction. *J Clin Microbiol* 28(1): 65-69.
59. Rothova A, Buitenhuis HJ, Meenken C, Brinkman CJ, Linssen A, et al. (1992) Uveitis and systemic disease. *Br J Ophthalmol* 76(3): 137-141.
60. Jap A, Chee SP. (2011) Viral anterior uveitis. *Curr Opin Ophthalmol* 22(6): 483-488.
61. Dong Q, Brulc JM, Iovieno A, Bates B, Garoutte A, et al. (2011) Diversity of bacteria at healthy human conjunctiva. *Invest Ophthalmol Vis Sci* 52(8): 5408-5413.

62. Chomel BB, Boulouis HJ. (2005) Zoonotic diseases caused by bacteria of the genus bartonella genus: New reservoirs ? new vectors? Bull Acad Natl Med 189(3): 465-77; discussion 477-80.
63. Terrada C, Bodaghi B, Conrath J, Raoult D, Drancourt M. (2009) Uveitis: An emerging clinical form of bartonella infection. Clin Microbiol Infect 15 Suppl 2: 132-133.
64. Loh J, Zhao G, Presti RM, Holtz LR, Finkbeiner SR, et al. (2009) Detection of novel sequences related to african swine fever virus in human serum and sewage. J Virol 83(24): 13019-13025.
65. Khashab M, Tector AJ, Kwo PY. (2007) Epidemiology of acute liver failure. Curr Gastroenterol Rep 9(1): 66-73.
66. Lee WM. (2008) Etiologies of acute liver failure. Semin Liver Dis 28(2): 142-152.
67. Stapleton JT. (2003) GB virus type C/Hepatitis G virus. Semin Liver Dis 23(2): 137-148.
68. Heringlake S, Osterkamp S, Trautwein C, Tillmann HL, Boker K, et al. (1996) Association between fulminant hepatic failure and a strain of GBV virus C. Lancet 348(9042): 1626-1629.
69. Linnen J, Wages J, Jr, Zhang-Keck ZY, Fry KE, Krawczynski KZ, et al. (1996) Molecular cloning and disease association of hepatitis G virus: A transfusion-transmissible agent. Science 271(5248): 505-508.
70. Young NS, Brown KE. (2004) Parvovirus B19. N Engl J Med 350(6): 586-597.
71. Dutta U, Mittal S, Ratho RK, Das A. (2005) Acute liver failure and severe hemophagocytosis secondary to parvovirus B19 infection. Indian J Gastroenterol 24(3): 118-119.
72. Dame C, Hasan C, Bode U, Eis-Hubinger AM. (2002) Acute liver disease and aplastic anemia associated with the persistence of B19 DNA in liver and bone marrow. Pediatr Pathol Mol Med 21(1): 25-29.
73. Lee WM, Brown KE, Young NS, Dawson GJ, Schlauder GG, et al. (2006) Brief report: No evidence for parvovirus B19 or hepatitis E virus as a cause of acute liver failure. Dig Dis Sci 51(10): 1712-1715.

74. [Anonymous]. (2000) American thoracic society. idiopathic pulmonary fibrosis: Diagnosis and treatment. international consensus statement. american thoracic society (ATS), and the european respiratory society (ERS). *Am J Respir Crit Care Med* 161(2 Pt 1): 646-664.
75. American Thoracic Society, European Respiratory Society. (2002) American thoracic Society/European respiratory society international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. this joint statement of the american thoracic society (ATS), and the european respiratory society (ERS) was adopted by the ATS board of directors, june 2001 and by the ERS executive committee, june 2001. *Am J Respir Crit Care Med* 165(2): 277-304.
76. King TE,Jr, Pardo A, Selman M. (2011) Idiopathic pulmonary fibrosis. *Lancet* .
77. Collard HR, Moore BB, Flaherty KR, Brown KK, Kaner RJ, et al. (2007) Acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 176(7): 636-643.
78. Kim DS, Collard HR, King TE,Jr. (2006) Classification and natural history of the idiopathic interstitial pneumonias. *Proc Am Thorac Soc* 3(4): 285-292.
79. Song JW, Hong SB, Lim CM, Koh Y, Kim DS. (2010) Acute exacerbation of idiopathic pulmonary fibrosis: Incidence, risk factors, and outcome. *Eur Respir J* .
80. Kim DS, Park JH, Park BK, Lee JS, Nicholson AG, et al. (2006) Acute exacerbation of idiopathic pulmonary fibrosis: Frequency and clinical features. *Eur Respir J* 27(1): 143-150.
81. Collard HR, Calfee CS, Wolters PJ, Song JW, Hong SB, et al. (2010) Plasma biomarker profiles in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Physiol Lung Cell Mol Physiol* 299(1): L3-7.
82. Huie TJ, Olson AL, Cosgrove GP, Janssen WJ, Lara AR, et al. (2010) A detailed evaluation of acute respiratory decline in patients with fibrotic lung disease: Aetiology and outcomes. *Respirology* .
83. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, et al. (2009) Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 180(2): 167-175.
84. Ware LB, Matthay MA. (2000) The acute respiratory distress syndrome. *N Engl J Med* 342(18): 1334-1349.

85. Bernard GR. (2005) Acute respiratory distress syndrome: A historical perspective. *Am J Respir Crit Care Med* 172(7): 798-806.
86. Lam WY, Yeung AC, Tang JW, Ip M, Chan EW, et al. (2007) Rapid multiplex nested PCR for detection of respiratory viruses. *J Clin Microbiol* 45(11): 3631-3640.
87. Aurelius E, Johansson B, Skoldenberg B, Staland A, Forsgren M. (1991) Rapid diagnosis of herpes simplex encephalitis by nested polymerase chain reaction assay of cerebrospinal fluid. *Lancet* 337(8735): 189-192.
88. Ikuta K, Saiga K, Deguchi M, Sairenji T. (2003) Epstein-barr virus DNA is detected in peripheral blood mononuclear cells of EBV-seronegative infants with infectious mononucleosis-like symptoms. *Virus Genes* 26(2): 165-173.
89. Welch TA. (1984) A technique for high-performance data compression. *IEEE Computer* 17(6): 9-19.
90. Mushahwar IK, Erker JC, Muerhoff AS, Leary TP, Simons JN, et al. (1999) Molecular and biophysical characterization of TT virus: Evidence for a new virus family infecting humans. *Proc Natl Acad Sci U S A* 96(6): 3177-3182.
91. Handa A, Dickstein B, Young NS, Brown KE. (2000) Prevalence of the newly described human circovirus, TTV, in united states blood donors. *Transfusion* 40(2): 245-251.
92. Okamoto H, Takahashi M, Nishizawa T, Ukita M, Fukuda M, et al. (1999) Marked genomic heterogeneity and frequent mixed infection of TT virus demonstrated by PCR with primers from coding and noncoding regions. *Virology* 259(2): 428-436.
93. Maggi F, Focosi D, Albani M, Lanini L, Vatteroni ML, et al. (2010) Role of hematopoietic cells in the maintenance of chronic human torquetenovirus plasma viremia. *J Virol* 84(13): 6891-6893.
94. Bando M, Ohno S, Oshikawa K, Takahashi M, Okamoto H, et al. (2001) Infection of TT virus in patients with idiopathic pulmonary fibrosis. *Respir Med* 95(12): 935-942.
95. Mariscal LF, Lopez-Alcorocho JM, Rodriguez-Inigo E, Ortiz-Movilla N, de Lucas S, et al. (2002) TT virus replicates in stimulated but not in nonstimulated peripheral blood mononuclear cells. *Virology* 301(1): 121-129.
96. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, et al. (2008) Metagenomic analysis of human diarrhea: Viral detection and discovery. *PLoS*

Pathog 4(2): e1000011.

97. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778): 1355-1359.
98. Friaza V, la Horra C, Rodriguez-Dominguez MJ, Martin-Juan J, Canton R, et al. (2010) Metagenomic analysis of bronchoalveolar lavage samples from patients with idiopathic interstitial pneumonia and its antagonistic relation with pneumocystis jirovecii colonization. *J Microbiol Methods* 82(1): 98-101.
99. Vannella KM, Moore BB. (2008) Viruses as co-factors for the initiation or exacerbation of lung fibrosis. *Fibrogenesis Tissue Repair* 1(1): 2.
100. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 462(7276): 1056-1060.
101. Duprez R, Boulanger E, Roman Y, Gessain A. (2004) Novel gamma-2-herpesvirus of the rhadinovirus 2 lineage in gibbons. *Emerg Infect Dis* 10(5): 899-902.
102. McGeoch DJ, Rixon FJ, Davison AJ. (2006) Topics in herpesvirus genomics and evolution. *Virus Res* 117(1): 90-104.
103. Lovisolo O, Hull R, Rosler O. (2003) Coevolution of viruses with hosts and vectors and possible paleontology. *Adv Virus Res* 62: 325-379.
104. Mindell DP, Rest JS, Villarreal LP. (2004) Viruses and the tree of life. In: Joel Cracraft, Michael J. Donoghue, editor. **Assembling the tree of life**. New York, New York: Oxford University Press. pp. 107-108-118.
105. Perez-Losada M, Christensen RG, McClellan DA, Adams BJ, Viscidi RP, et al. (2006) Comparing phylogenetic codivergence between polyomaviruses and their hosts. *J Virol* 80(12): 5663-5669.
106. Holmes EC. (2003) Molecular clocks and the puzzle of RNA virus origins. *J Virol* 77(7): 3893-3897.
107. Jackson AP, Charleston MA. (2004) A cophylogenetic perspective of RNA-virus evolution. *Mol Biol Evol* 21(1): 45-57.
108. Villarreal LP, Defilippis VR, Gottlieb KA. (2000) Acute and persistent viral life strategies and their relationship to emerging diseases. *Virology* 272(1): 1-6.

109. Switzer WM, Salemi M, Shanmugam V, Gao F, Cong ME, et al. (2005) Ancient co-speciation of simian foamy viruses and primates. *Nature* 434(7031): 376-380.
110. Libeskind-Hadas R, Charleston MA. (2009) On the computational complexity of the reticulate cophylogeny reconstruction problem. *J Comput Biol* 16(1): 105-117.
111. Ovadia Y, Fielder D, Conow C, Libeskind-Hadas R. (2011) The co phylogeny reconstruction problem is NP-complete. *J Comput Biol* 18(1): 59-65.
112. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* 4(2): e4345.
113. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. (2011) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* 27(3): 431-432.
114. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2011) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 39(Database issue): D38-51.
115. Pruitt KD, Tatusova T, Klimke W, Maglott DR. (2009) NCBI reference sequences: Current status, policy and new initiatives. *Nucleic Acids Res* 37(Database issue): D32-6.
116. Samuel Lattimore B, van Dongen S, Crabbe MJ. (2005) GeneMCL in microarray analysis. *Comput Biol Chem* 29(5): 354-359.
117. Enright AJ, Van Dongen S, Ouzounis CA. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7): 1575-1584.
118. Holm L, Sander C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* 14(5): 423-429.
119. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8): 1034-1050.
120. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The pfam protein families database. *Nucleic Acids Res* 38(Database issue): D211-22.

121. Irwin DM, Kocher TD, Wilson AC. (1991) Evolution of the cytochrome b gene of mammals. *J Mol Evol* 32(2): 128-144.
122. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2011) Database resources of the national center for biotechnology information. *Nucleic Acids Res* .

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

1/9/12

Date