UNIVERSITY OF CALIFORNIA, SAN DIEGO

SAN DIEGO STATE UNIVERSITY

Viral Metagenomics in Host-associated Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biology

by

Dana Leigh Willner

Committee in charge:

University of California, San Diego

Professor Eric Allen
Professor Phil Bourne

San Diego State University

Professor Forest Rohwer, Chair
Professor Stephanie Brodine
Professor David Lipson

2010

The Dissertation of Dana Leigh Willner is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

San Diego State University

2010

DEDICATION

To Shirley and Sam, and Carol and Stanley: For all that you are and all that you do

To Sasha, Alice, BallenTX, and CK: My friends, first and last and always

EPIGRAPH

Go, then. There are other worlds than these.

-Stephen King, *The Dark Tower*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Dr. Forest Rohwer, for his guidance and support during my doctoral studies. I would also like to acknowledge Dr. Matthew Haynes for helping me with anything and everything, and Linda Wegley for her kindness and unflagging optimism.

Chapter 1, in full, has been accepted for publication as part of the chapter Viral metagenomics: from fish slime to the world. Willner, D, Desnues, C, and Rohwer, F. Chapter in “Metagenomics and its applications in agriculture, biomedicine, and environmental studies,” Nova Scientific Publishers. The dissertation author was the primary investigator and author of this material.

Chapter 2 is a reprint of: Willner D**,** Thurber RV, and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology.* (Epub March 18, 2009) The dissertation author was the primary investigator and author of this material.

Chapter 3 has been accepted for publication in PNAS, and is currently in review as: Willner, D, Furlan, M, Schmieder, R, Grasis, J, Pride, DT, Relman, D, Angly, F, McDole, T, Mariella, RP, Rohwer, F, and Haynes, M. Metagenomic detection of phage-encoded platelet binding factors in the human oropharynx. The dissertation author was the primary investigator and author of this material.

Chapter 4 has been published in full as: Willner D, Furlan M, Haynes M, Schmieder R, Angly F, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. (2009) Metagenomic analysis of respiratory tract viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*. 4(10): e7370. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, has been accepted for publication as part of the chapter Viral metagenomics: from fish slime to the world. Willner, D, Desnues, C, and Rohwer, F. Chapter in “Metagenomics and its applications in agriculture, biomedicine, and environmental studies,” Nova Scientific Publishers. The dissertation author was the primary investigator and author of this material.

VITA

2000	Bachelor of Arts, Washington University, St. Louis
2000	Bachelor of Science, Washington University, St. Louis
2006	Master of Science, San Diego State University
2010	Doctor of Philosophy, University of California, San Diego/ San Diego State University

PUBLICATIONS

Willner, D, Furlan, M, Schmieder, R, Grasis, J, Pride, DT, Relman, DA, Angly, FE, McDole, T, Mariella Jr., RP, Rohwer, F, and Haynes, M. Metagenomic detection of phage-encoded platelet binding factors in the human oropharynx. (In press)

Willner, D, Desnues, C, and Rohwer, F. Viral metagenomics: from fish slime to the world. Chapter in “Metagenomics and its applications in agriculture, biomedicine, and environmental studies”. Nova Scientific Publishers. (In press)

Willner, D and Furlan, M. Deciphering the role of phage in the cystic fibrosis airway. *Virulence.* (In press)

Su, L, Willner, DL, and Segall, AM. (2010) An antimicrobial peptide that targets DNA repair intermediates *in vitro* inhibits *Salmonella* growth within murine macrophages. *Antimicrobial Agents and Chemotherapy.* (epub: doi: 10.1128/AAC.01610-09)

Rodriguez-Mueller, B, Li, L, Wegley, L, Furlan, M, Angly, F, Breitbart, M, Buchanan, J, Desnues, C, Dinsdale, E, Edwards, R, Felts, B, Haynes, M, Liu, H, Lipson, D, Mahaffy, J, Martin-Cuadrado, ABM, Mira, A, Nulton, J, Pasic, L, Rayhawk, S, Rodriguez-Mueller, J, Rodriguez-Valera, F, Salamon, P, Srinagesh, S, Thingstad, RF, Tran, T, Thurber, RV, Willner, D, Youle, M, Rohwer, F. (2010) Kill-the-winner in four aquatic environments. *ISME Journal.*(epub: doi:10.1038/ismej.2010.1)

Angly, F, Willner, D, Prieto-Davó, A, Edwards, RA, Schmieder, R, Vega-Thurber, R, Antonopoulos, DA, Barott, K, Cottrell, MT, Desnues, C, Dinsdale, EA, Furlan, M, Haynes, M, Henn, MR, Hu, Y, Kirchman, DL, McDole, T, McPherson, JD, Meyer, F, Miller, RM, Mundt, E, Naviaux, RK, Rodriguez-Mueller, B, Stevens, R, Wegley, L, Zhang, L, Zhu, B, Rohwer, F. (2009) The GAAS metagenomic tool and its application to the estimation of viral and microbial average genome size in four biomes. *PLoS Computational Biology.* 5(12): e1000593.

Allen, B, Willner, D, Oechel, W, and Lipson, D. (2009) Top-down control of microbial biomass and activity in an Arctic ecosystem. *Environmental Microbiology*. DOI:

1111/j.1462-2920.2009.02104.x

Willner D, Furlan M, Haynes M, Schmieder R, Angly F, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. (2009) Metagenomic analysis of respiratory tract viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*. 4(10): e7370.

Thurber RV, Willner-Hall, D, Angly F, Desnues C, Rodriguez-Brito B, Rohwer F. (2009) Metagenomic analysis of stressed coral holobionts. *Environmental Microbiology*. (Epub April 22, 2009)

Willner D**,** Thurber RV, and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology.* (Epub March 18, 2009)

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*. 452(7185); 340-3.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. *Nature.* 452(7187): 629-32.

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, et al. (2008) Microbial ecology of four coral atolls in the northern line islands. *PloS ONE.* 3(2): e1584.

Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, et al. (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *PNAS*105:18413-18418.

FIELD OF STUDY

Major Field: Molecular Biology, Professor Forest Rohwer

ABSTRACT OF THE DISSERTATION

Viral metagenomics in host-associated systems

by

Dana Leigh Willner

Doctor of Philosophy in Biology

University of California, San Diego, 2010
San Diego State University, 2010

Professor Forest Rohwer, Chair

Viruses are the most abundant and diverse entities on earth. Exploration of viral diversity has traditionally been limited by the lack of common marker genes, however, the advent of viral metagenomics has made it possible to characterize global viral communities. Viruses in host-associated systems, such as human and animal tissues, are of special interest as they may be causative agents of disease. Additionally, changes in the total viral consortium may be indicative of host health status, with opportunists and pathogens replacing normal viral flora in the disease state. This dissertation presents an introduction to viral metagenomics and explores use in both human and animal associated systems. Methods in viral metagenomics, including both molecular biology and bioinformatics are reviewed as well as viral metagenomic studies to date. The metagenomic signature technique is explored as a method to characterize metagenomes and to screen for contaminating host genomic DNA sequences in viral metagenomes. Three experimental studies are presented to demonstrate the utility of metagenomics in healthy and diseased individuals. A case study of oropharyngeal viruses revealed the presence of phage-encoded virulence genes in healthy individuals, and also provided the first ever characterization of oropharyngeal viral communities. In the second study, viral communities from the airways of individuals with and without cystic fibrosis (CF) were compared. There was a striking difference in metabolic functions encoded by phage in CF versus Non-CF individuals. Regardless of which taxa were present, CF-associated phage shared a common core metabolism that reflected the disease state and aberrant airway physiology. Viral communities in healthy and diseased fish were compared in the third study. In contrast to the airway viromes, fish-associated viromes were found to differ taxonomically but not in metabolic function in the disease state. Together these studies demonstrate the power of viral metagenomics for discovery and for deciphering how viral communities change in the face of disease.

CHAPTER 1: INTRODUCTION

Viruses are the most diverse biological entities on earth. Culturing-based approaches to study this diversity are not feasible, due to the scale of the problem and the inability to culture viruses and/or their hosts. Culture-independent methods such as 16S rDNA analysis, which have been so successful for changing our view of microbial diveristy, are not applicable to viruses because they do not share signature genes. Viral metagenomics circumvents these limitations, allowing for the direct isolation and sequencing of viral DNA and RNA from environmental samples. This chapter will review metagenomic methods and how they have been used to survey viral communities in most of the world's major biomes.

Viral metagenomics for discovery and ecological studies

Since the publication of the first two viral metagenomes in 2002, nearly 100 viromes have been sequenced and published (Table 1.1). This explosion in metagenomic data has been enabled by improvements in isolation techniques, reduced sequencing costs, and the development of bioinformatic tools. Metagenomic methods can be applied to study viruses in any system, including marine, terrestrial, and animal-associated environments. Viral metagenomics has been used for two main applications: 1) viral discovery and 2) ecological studies.

Viral metagenomics has been used as a diagnostic tool to discover etiologic agents in disease outbreaks and also to develop rapid screens for viral pathogens in plants and animals. Many traditional diagnostic techniques such as PCR and ELISA

Table 1.1: Viral metagenomic studies, Continued. Studies were classified by type as well as by sequencing and isolation methods.

Ecological studies

Name	**Goal**	**Methods used**		**Reference**
		Viral purification	**Sequencing**	
Offshore marine viruses	Characterize the diversity of two uncultured marine viral communities	Refined	Sanger	Breitbart *et al.*, 2002
Marine sediment viruses	Characterize a near-shore marine sediment viral community and compare to marine communities	Refined	Sanger	Breitbart *et al.*, 2003
Adult fecal DNA	Characterize the composition and population structure of a viral community from human feces	Refined	Sanger	Breitbart *et al.*, 2003
Horse fecal viruses	Characterize the composition and diversity of viruses in the equine gut	Intermediate	Sanger	Cann *et al.*, 2005
Four marine viromes	Characterize viral distribution and diversity in four oceanic regions	Refined	454 GS20 pyrosequencing	Angly *et al.*, 2006
Coastal RNA viruses	Characterize the diversity of marine RNA viruses from off the coast of Vancouver, BC	Intermediate	Sanger	Culley *et al.*, 2006
Adult fecal RNA viruses	Characterize RNA viruses isolated from human feces	Intermediate	Sanger	Zhang *et al.*, 2006
Chesapeake Bay virioplankton	Characterize an estaurine viral community	Intermediate	Sanger	Bench *et al.*, 2007
Soil viruses	Characterize viral diversity in desert, prairie, and rainforest soils and compare to bacterial, archaeal, and fungal diversity	Refined	Sanger	Fierer *et al.*, 2007
Tampa Bay lysogens	Identify the genes associated with lysogeny in temperate phage in an aquatic system	Refined	454 GS20 pyrosequencing	Breitbart *et al.*, 2008
Infant fecal viruses	Characterize viral communities in an infant gut	Refined	Sanger	Breitbart *et al.*, 2008
Microbialite viruses	Assess diversity of viruses in microbialites	Refined	454 GS20 pyrosequencing	Desnues *et al.*,
Southern Line Islands viruses	Evalute the effects of human disturbance on viral communities associated with four coral reefs	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008a
Aquaculture pond viruses	Characterize changes in the viral community in four aquaculture ponds	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008b

Table 1.1: Viral metagenomic studies, Continued.

Name	Goal	Methods used		Reference
		Viral purification	**Sequencing**	
Solar saltern viruses	Characterize changes in the viral communities over time and across a salinity gradient	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008b
Salton Sea viruses	Characterize viral communities in the Salton Sea in California	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008b
Skan Bay viruses	Characterize viral communities in Skan Bay in Alaska	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008b
Mosquito-associated viruses	Characterize viral communities associated with mosquitoes in San Diego, CA	Refined	454 GS20 pyrosequencing	Dinsdale *et al.*, 2008b
Human diarrhea viruses	Characterize viral communities in pediatric patients with diarrhea	Intermediate	Sanger	Finkbeiner *et al.*, 2008
Rice paddy viruses	Characterize ssDNA viral diversity in soil	Intermediate	Sanger	Kim *et al.*, 2008
Healthy and bleached coral viruses	Assess differences between viral communities associated with healthy and diseased corals	Refined	Sanger	Marhaver *et al.*, 2008
Hot spring viruses	Characterize the diversity, composition, and adaptations of two viral communities from hot springs	Intermediate	Sanger	Schoenfeld *et al.*, 2008
Porites compressa viruses	Determine the effects of stressors on viral communities associated with coral	Refined	454 GS20 pyrosequencing	Vega Thurber *et al.*, 2008
Hydrothermal vent viruses	Determine the prevalence of temperate viruses in hydrothermal vents	Intermediate	Sanger	Williamson *et al.*, 2008
Lake RNA viruses	Characterize RNA viral communities in a freshwater lake	Intermediate	Sanger and 454 GSFLX pyrosequencing	Djikeng *et al.*, 2009
Reclaimed water viruses	Characterize the viral community in reclaimed water and compare to potable water	Refined	454 GSFLX pyrosequencing	Rosario *et al.*, 2009
Human respiratory tract viruses	Compare respiratory viral communities from individuals with and without cystic fibrosis	Refined	454 GSFLX pyrosequencing	Willner *et al.*, 2009
Viral discovery				
Nasopharyngeal viruses	Develop a method for viral screening in human nasopharyngeal aspirates	Intermediate	Sanger	Allander *et al.*, 2005

Table 1.1: Viral metagenomic studies, Continued.

Name	Goal	Methods used		Reference
		Viral purification	**Sequencing**	
Human blood viruses	Develop a method for viral identification in human blood	Refined	Sanger	Breitbart *et al.*, 2005
Novel blood viruses	Identify viruses associated with acute viral infection and HIV	Intermediate	Sanger	Jones *et al.*, 2005
Novel human polyomavirus	Viral screening of human nasopharyngeal aspirates	Intermediate	Sanger	Allander *et al.*, 2007
Honeybee colony collapse disorder	Identify candidate pathogens in honeybee colony collapse disorder	Crude	454 GSFLX pyrosequencing	Cox-Foster *et al.*, 2007
Borna virus in psittacine birds	Identify etiologic agents in parrots with proventricular dilation disease	Crude	454 GSFLX pyrosequencing	Honkavouri *et al.*, 2008
Arenavirus in transplant patients	Identify the cause of death in two patients following transplant	Crude	454 GSFLX pyrosequencing	Palacios *et al.*, 2008
RNA viruses in animal tissue	Develop a method for rapid identification of RNA viral pathogens in animal tissue	Intermediate	Sanger	Victoria *et al.*, 2008
Poultry intestinal viruses	Identify viruses associated with enteric disease in poultry	Intermediate	Sanger	Zsak *et al.*, 2008
Plant pathogenic viruses	Develop a diagnostic tool to detect viral pathogens in plants	Crude	454 GSFLX pyrosequencing	Adams *et al.*, 2009
Syrah grapevine decline viruses	Identify etiologic agents of decline in Syrah grapevines	Crude	454 GSFLX pyrosequencing	Al Rawanih *et al.*, 2009
Hemorrhagic fever virus	Determine the etiologic agent of a hemorrhagic fever outbreak	Crude	454 GSFLX pyrosequencing	Briese *et al.*, 2009
Klassevirus discovery	Identify etiologic agents associated with diarrhea in children	Crude	454 GSFLX pyrosequencing	Greninger *et al.*, 2009
Fecal and nasopharyngeal viruses	Demonstrate the utility of a high-throughput sequencing approach to detect human pathogens	Crude	454 GSFLX pyrosequencing	Nakamura *et al.*, 2009
Sea lion viruses	Determine the etiologic agent of a mortality event in captive California sea lions	Refined	Sanger	Ng *et al.*, 2009
Sea turtle viruses	Identify viruses associated with fibropapillomas in sea turtles	Refined	Sanger	Ng *et al.*, 2009
AFP stool viruses	Identify etiologic agents in stool samples from children with acute flaccid paralysis	Intermediate	454 GSFLX pyrosequencing	Victoria *et al.*, 2009

rely on a priori knowledge to identify pathogens. In many cases, especially disease outbreaks and sudden mortality events, the etiologic agent is novel. Metagenomics provides rapid screening for both known and novel pathogens, since nucleic acids are isolated directly from the environment. Adams et al.(2009) identified a novel bromovirus in an infected plant using a methodology to rapidly screen for plant pathogens using metagenomics and high-throughput sequencing (1). Similarly, RNA viruses causing decline disease in Syrah grapevines were discovered by metagenomic analysis (2). Allander et al. (2001) identified two novel bovine parvoviruses in commercially available bovine serum while developing the DNAse-SISPA method for viral discovery (3). Viral metagenomics has also identified etiologic agents of disease outbreaks in animals including poultry and apicultured bees (discussed below), and a rapid assay has recently been developed to detect RNA viruses in animal tissue (4-6).

Analysis of viral, and especially phage, communities using metagenomics has provided insight into microbial ecology and environmental dynamics of host-associated and agricultural systems**.** Virus-encoded functional genes are indicative of the most important metabolic processes in a given ecosystem, and viral metabolic profiles are distinct in different environments (7). In marine environments, phage and algal viruses carry photosynthetic genes which enable their hosts to carry out photosynthesis even in intense sunlight, which is normally photo-inhibitive (8-11). Many phage also encode genes for antibiotic and host immune resistance, and virulence factors such as toxins which enhance the pathogenicity of their microbial hosts (12). Though not discussed in detail in this chapter, phage also influence biogeochemical cycles, control microbial populations, maintain microbial diversity through predation, and have been shown to exert top-down controls in marine, hyperthermal, and soil environments (13-17).

Viral taxonomy can also provide an environmental readout, as abundant viruses may represent thriving host communities, and viral diversity is correlated with environmental complexity. Metagenomic analysis of viral communities in Korean rice paddy soil revealed a high diversity of eukaryotic viruses, many of which seemed to be novel pathogens of indigenous plants and animals (18). Fierer et al. (2007) found similar results in prairie, desert, and rainforest soils (19). Soil is an extremely complex environment, containing many niches and microhabitats, and thus would be expected to support the high level of viral diversity observed in these studies (20).

Laboratory methods for viral metagenomics

Viral isolation and purification techniques in metagenomics range from crude to very refined. Crude protocols involve initial processing of samples via homogenization, shaking, or centrifugation of environmental or clinical samples with no further viral purification steps prior to nucleic acid extraction. Metagenomes generated from crude isolations often contain a large proportion of non-viral sequences, as no attempts have been made to remove eukaryotic and microbial cells. Refined protocols use filtration followed by density gradient centrifugation to specifically isolate and purify virions. Cesium chloride step gradients have been commonly used to separate viral particles from free DNA and cellular material based on buoyant density (21). While refined methods enrich and select for viruses, at each step virions can be lost to processing, which may decrease the overall viral diversity detected (21). Many viral metagenomes have been created using an intermediate approach, i.e. using filtration to remove larger non-viral particles, but with no subsequent density centrifugation step.

Following viral nucleic acid isolation, metagenomic protocols often include additional steps to remove host genomic DNA or amplify nucleic acids. Samples derived from plants or animals can be contaminated with DNA from host organisms and microbial flora, as well as microbial DNA from reagents (3). DNAse treatment is commonly used to degrade unwanted free DNA prior to sequencing, as viral nucleic acids are protected from degradation by their proteinaceous capsids (3; 21). While DNAse treatment does reduce the amount of contaminating DNA in viral samples, it does not completely remove it, and the use of bioinformatic filters may be necessary after sequencing (3).

The amount of total nucleic acids isolated from viral particles is often too low for sequencing, depending on the sequencing technology. Multiple displacement amplification with Phi29 polymerase can be used to amplify total viral DNA or cDNA, and generate adequate template for sequencing (21-23). Viral RNA can be amplified using whole transcriptome amplification methods, such as the TransPlex system (21; 24). Very small amounts of starting viral nucleic acid material may need to be cloned (e.g., LASLs; see below).

Metagenomic sequencing

Metagenomic sequencing technologies differ in library preparation methods and the length of reads produced. Sanger sequencing methods produce the longest reads (>600 base pairs), but require cloning. The earliest DNA viral metagenomes were linker-amplified shotgun libraries (LASLs), created by ligating dsDNA linkers to genomic DNA fragments and cloning them into a vector plasmid for subsequent Sanger sequencing (25-27). LASLs were initially limited to dsDNA viruses, however, the advent of strand-displacement amplification using Phi29 polymerases extended the technology to ssDNA viruses (18; 28). Allander et al. (2005) sequenced the first RNA viral metagenome by using random RT-PCR primers linked to adaptor sequences for cDNA synthesis (29). Similar to the LASL method, adaptor-ligated cDNAs could then be cloned and Sanger sequenced.

The development of high-throughput pyrosequencing by 454 Life Sciences revolutionized viral metagenomics, allowing for the rapid acquisition of large amount of sequence data with no cloning (30). The sequencing by synthesis methodology of 454 pyrosequencing is explained in detail in (30). The first pyrosequenced viral metagenomes were published by Angly et al. in 2006, which were sequenced using 454 GS20 technology, producing over 1.5 million metagenomic reads with an average length of 102 base pairs (31). In 2007, 454 Life Sciences released the GSFLX system, which extended read length to an average of 250 base pairs. Subsequent improvements in sequencing chemistry have extended read length to over 400 base pairs, which is nearly comparable to read lengths produced by Sanger sequencing. Other high-throughput sequencing methodologies are currently available, including reversible chain termination sequencing (e.g. Illumina) and sequencing by ligation (ABI SOLiD) (32). These methods produce shorter sequences (<100 base pairs) which are maybe less appropriate for metagenomic projects. However, Illumina's new generation will produce >200 base pair reads.

Bioinformatics for viral metagenomics

Bioinformatic analyses of viral metagenomes strive to answer three questions: how many viruses are there (diversity), what are they (taxonomy), and what are they doing (function)? Analyses often extend to compare diversity, taxonomy, and function between viromes, e.g. diseased versus healthy. Additionally, bioinformatic screens can be used to identify and filter out host and other non-viral sequences from metagenomes. Bioinformatic methods for viral metagenomics can be generally classified as similarity-dependent or similarity-independent. Similarity-dependent methods rely on comparisons with known sequences maintained in annotated databases to assign taxonomy and function to metagenomic sequences. Viral metagenomes typically contain a large number of sequences with no similarity to known sequences. For example, in microbialite viromes, unknown sequences accounted for 99% of metagenomic libraries (33). Similarity-based analyses exclude these sequences, and may disregard a significant fraction of the available data. Similarity-independent methods are free of this limitation, and are able to utilize all metagenomic sequences whether they are known or unknown. Here, we review both similarity-dependent and similarity-independent bioinformatic methods for the analysis of viral metagenomes.

Metagenomic sequences are typically assigned to taxonomic classes by BLAST-based comparisons to known sequences. Many target databases are available ranging from the very general, such as the non-redundant database at NCBI, to specific boutique databases containing subsets of viral or other sequences. Users can conduct BLAST from the command line interface or use annotation services such as MG-RAST (http://metagenomics.nmpdr.org/) or IMG (http://img.jgi.doe.gov/), and can select from a variety of BLAST algorithms (34; 35). BLASTn compares nucleotide sequences to a nucleotide database, while tBLASTx compares translated sequences to a translated nucleotide database in all six reading frames (34). tBLASTx can identify similarities at the amino acid level, and is useful for viral discovery, as novel viruses may by similar to known viruses at the protein but not the nucleotide level. BLASTx can also be used to compare translated metagenomic sequences to protein databases, such as the SEED (34; 36).

Several software packages are available to rapidly parse and visualize BLAST results, and also to assist in taxonomic assignment. CARMA (http://www.cebitec.uni-bielefeld.de/brf/carma/carma.html) classifies sequences taxonomically by searching for conserved Pfam domains and protein domains, and works well for even short sequence lengths (37). MEGAN (http://www-ab.informatik.uni-tuebingen.de/software/megan) assigns metagenomic sequences to NCBI taxonomic classes based on significant BLAST similarities, and assigns taxonomy at the lowest (i.e. most specific) level possible using a least common ancestor algorithm (38). The GAAS (http://www.sourceforge.net/GAAS) metagenomic tool also uses BLAST similarities to assign taxonomy. GAAS provides a set of viral community relative abundances based on all BLAST similarities for all sequences (39). GAAS also normalizes for the length of the target genome in the database, which provides more accurate estimates of community composition (39). BLAST analysis is biased towards larger genomes since they will produce more sequence fragments of a given size per genome than smaller genomes (39). Without normalization, viruses with larger genomes may appear more abundant than they truly are, and viruses with small genomes that are present in low abundances may be missed completely.

The phage proteomic tree (PPT) provides a framework to compare and visualize phage populations in metagenomes. Phylogenetic distances between phage in the PPT were determined by calculating distances between phage proteins using BLASTp (40). Initially, 105 phage genomes were incorporated into the tree, however, there are now 510 genomes available (http://phage.sdsu.edu/phage) (40). BLAST similarities to the phage genomes can be plotted against the PPT using the Bio-Metamapper (http://scums.sdsu.edu/Mapper) to obtain phage community signatures for a given environment. Using the phylogenetic relationships represented by the PPT, phage communities in different metagenomes can be compared using UniFrac. UniFrac provides statistical analyses to compare phylogenetic diversity between environments (41; 42). The inputs to UniFrac are a phylogenetic tree and counts of the relative abundances of each taxa in each environment, and outputs include distances between environments, statistical tests to tell which environments are significantly different, and environments clustered by phylogenetic similarity (41; 42).

Metabolic functions are assigned to viral metagenomic sequences using BLAST similarities. Specialized databases containing functionally annotated sequences are available such as the NCBI Cluster of Orthologous Groups (COGs), TIGR funcat, and SEED (36; 43; 44). For viral metagenomes, BLAST analyses may only assign functions to a small percentage of sequences, as many viral proteins are unknown, and homologies may be distant between viral proteins and microbial proteins in the database (45). There are other methods available for functional annotation, including profile Hidden Markov Model approaches (see (46)) and gene neighborhood analysis (see (47)). Once sequences are annotated, metabolic profiles (i.e. the cohort of metabolic functions in a viral community) are compared using non-parametric and multivariate statistical techniques. The non-parametric statistical program XIPE compares the metabolic profiles of metagenomes in a pairwise manner, using a bootstrap procedure (48). XIPE determines if metabolic profiles are different overall at a particular confidence level, and then identifies which functional groups are driving the observed difference (48). Metastats performs similar comparisons, but uses a different statistical methodology which allows for comparison of multiple samples within two groups (49). Multivariate methods such as canonical discriminant analysis (CDA) and non-metric multidimensional scaling have also been used to compare large sets of metagenomes and group them by functional similarities (45; 50; 51).

Viral diversity and community structure cannot usually be determined from BLAST comparisons, since many metagenomic sequences have no significant similarities to known organisms. Laboratory methods such as Pulsed Field Gel Electrophoresis (PFGE) and Randomly Amplified Polymorphic DNA (RAPD-PCR) assays have been used to quantify viral richness (52-55). These methods do not always provide an accurate assessment of diversity, e.g. one band can represent multiple genomes in PFGE (56). Additionally, they cannot be used for metagenomes which have already been sequenced unless sufficient starting material was saved *a priori*. Computational methods which do not rely on BLAST similarities (i.e. similarity-independent methods) have been developed to explore the diversity of viral communities. PHAge Communities from Contig Spectrum (PHACCS) implements mathematical models to determine viral community structure and calculate alpha diversity measures from contig spectra (25; 57). When metagenomic sequences are assembled, overlapping sequences are grouped together to form contigs, or contiguous sequences. A contig is defined by the number of sequences that compose it, i.e. a 2-contig was produced from two original sequences, a 10-contig from ten sequences. A contig spectrum is a sequential listing of the number of contigs of each type generated by assembly, e.g. the contig spectrum [5 2 1] corresponds to one 5-contig, two 2-contigs, and one 3-contig. PHACCs can be easily run from a web interface (http://biome.sdsu.edu/phaccs/) or from the command line. The program takes four inputs: the calculated contig spectrum, the average fragment size in the metagenomic library, the minimum overlap length, and the average genome size (39; 57). Contig spectra can be generated using the free software Circonspect (http://sourceforge.net/projects/circonspect/), and average genome size can be estimated using GAAS (39). PHACCs tests several viral community structure models, and outputs the best fit model, along with estimated species richness, evenness, and the Shannon diversity index (57).

Methods to compare diversity between environments (beta-diversity) using viral metagenomes range from very simple to very complex. Diversity comparisons seek to determine which species are shared and which are unique to individual environments. Finding the number of shared and unique sequences in metagenomes can be used as a simple proxy for species diversity. Bi-directional BLAST analysis has been applied to metagenomes to find shared sequences as well as to protein sequences in individual genomes to identify orthologs (33; 51; 58-60). Each metagenome is formatted as a BLAST database and is compared pairwise to all other metagenomes using either BLASTn or tBLASTx. Sequences from two metagenomes which are the best BLAST hit for each other in both directions (i.e when metagenome 1 was used as the database and metagenome 2 was used as the query and vice versa) were considered to be shared between metagenomes. A higher percentage of shared sequences between metagenomes indicate a higher degree of similarity, however, bi-directional BLAST does not give any information on the actual number of species shared. Monte Carlo methods have been used to provide more accurate assessments of beta-diversity (31; 33). In brief, metagenomic reads from different metagenomes are assembled with each other to generate cross-contig spectra representing shared species between two viral communities (31). Monte Carlo simulations based on the cross-contig spectra are then used to determine the percentage of species shared, and how their abundances change between the different communities (31). This methodology is described in detail in (31).

Bioinformatic screens for host DNA contamination

Molecular methods such as DNAse treatment may not completely remove host and other undesirable DNA from viral metagenomic samples. Contaminating sequences can be rapidly identified and removed using bioinformatic methods. Compositional analysis provides a rapid screen for eukaryotic DNA contamination in viral metagenomes as discussed in (61) and Chapter 2 of this dissertation.

Compositional analysis techniques examine patterns of oligonucleotide usage in DNA sequence, and can be used to classify genomic and metagenomic sequences (62-67). Eukaryotic genomes are depressed in CG dinucleotides (68). Dinucleotide odds ratios compare the frequency of a dinucleotide in a sequence, given the frequency of the individual nucleotides composing it (69). An odds ratio greater than one indicates that the dinucleotide is more abundant than expected, while an odds ratio less than one indicates that is it less abundant than expected (69). CG odds ratios which are significantly less than one for a metagenome suggest the presence of eukaryotic DNA and warrant further analysis, typically BLASTn comparisons, to find and remove contaminating sequences (61). The computational time for calculation of dinucleotide odds ratios is minimal, so pre-screening with compositional analysis saves time by preventing unnecessary BLAST analyses which are more computationally intensive (61).

Conclusions

Viral metagenomics provides a methodology to characterize viral ecology in natural systems. Analysis of viral communities can provide information on the health of an ecosystem as well as reveal novel pathogens and etiologic agents of disease. Metagenomic methods can also be even further refined to select for functional subsets in populations using methods such as DNA-SIP (70). Metagenomics gives an overview of how a system is working, but follow-up studies are required to extract more specific details about community structure and dynamics. Assembly of metagenomic sequences into contigs could provide more information, however, to truly validate the presence of a novel virus and obtain a complete genome, more specific methods such as PCR would be necessary. This approach has typically been used in viral discovery studies such as those of Allander et al. (2005 and 2007), i.e., identification of candidate viruses through metagenomics, followed by PCR confirmation and characterization (29; 71). Metagenomic data must also be carefully screened for contaminating sequences, as discussed in the following chapter. Viral metagenomics provides a starting point for the in-depth study of viral taxonomy, function, and diversity. With careful selection of follow-up studies and verification of results, the applications of viral metagenomics are only restricted by the limits of our imaginations.

References

1. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol 2009 Jul;10(4):537-545.

2. Al Rwahnih M, Daubert S, Golino D, Rowhani A. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 2009 May;387(2):395-401.

3. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc. Natl. Acad. Sci. U.S.A 2001 Sep;98(20):11609-11614.

4. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan P, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui L, Hutchison SK, Simons JF, Egholm M, Pettis JS, Lipkin WI. A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. Science 2007 Oct;318(5848):283-287.

5. Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL. Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog

2008;4(9):e1000163.

6. Zsak L, Strother KO, Kisary J. Partial genome sequence analysis of parvoviruses associated with enteric disease in poultry. Avian Pathol 2008 Aug;37(4):435-441.

7. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. Functional metagenomic profiling of nine biomes. Nature 2008 Apr;452(7187):629-632.

8. Seaton G, Lee K, Rohozinski J. Photosynthetic Shutdown in Chlorella NC64A Associated with the Infection Cycle of Paramecium bursaria Chlorella Virus-1. Plant Physiol 1995 Aug;108(4):1431-1438.

9. Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB, Yooseph S, Zeidner G, Golden SS, Mackey SR, Adir N, Weingart U, Horn D, Venter JC, Mandel-Gutfreund Y, Béjà O. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J 2007 Oct;1(6):492-501.

10. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 2006 Jul;4(8):e234.

11. Mann NH, Cook A, Millard A, Bailey S, Clokie M. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 2003 Aug;424(6950):741.

12. Wagner PL, Waldor MK. Bacteriophage control of bacterial virulence. Infect. Immun 2002 Aug;70(8):3985-3993.

13. Noble RT, Fuhrman JA. Rapid virus production and removal as measured with fluorescently labeled viruses as tracers. Appl. Environ. Microbiol 2000 Sep;66(9):3790-3797.

14. Fuhrman JA, Schwalbach M. Viral influence on aquatic bacterial communities. Biol. Bull 2003 Apr;204(2):192-195.

15. Breitbart M, Wegley L, Leeds S, Schoenfeld T, Rohwer F. Phage community dynamics in hot springs. Appl. Environ. Microbiol 2004 Mar;70(3):1633-1640.

16. Allen B, Willner D, Oechel WC, Lipson D. Top-down control of microbial activity and biomass in an Arctic soil ecosystem [Internet]. Environ Microbiol 2009 Nov; [cited 2010 Mar 18] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20002136

17. Rodriguez-Valera F, Martin-Cuadrado A, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol 2009 Nov;7(11):828-836.

18. Kim K, Chang H, Nam Y, Roh SW, Kim M, Sung Y, Jeon CO, Oh H, Bae J. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl. Environ. Microbiol 2008 Oct;74(19):5975-5985.

19. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl. Environ. Microbiol 2007 Nov;73(21):7059-7066.

20. Torsvik V, Øvreås L. Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol 2002 Jun;5(3):240-245.

21. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc 2009;4(4):470-483.

22. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 2002 Apr;99(8):5261-5266.

23. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics [date unknown];7:216-216.

24. Tomlins SA, Mehra R, Rhodes DR, Shah RB, Rubin MA, Bruening E, Makarov V, Chinnaiyan AM. Whole transcriptome amplification for gene expression profiling and development of molecular archives. Neoplasia 2006 Feb;8(2):153-162.

25. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U.S.A 2002 Oct;99(22):14250-14255.

26. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol 2003 Oct;185(20):6220-6223.

27. Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F. Production of shotgun libraries using random amplification. BioTechniques 2001 Jul;31(1):108-112, 114-

116, 118.

28. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 2005 Nov;39(5):729-736.

29. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proceedings of the National Academy of Sciences of the United States of America 2005;102(36):12891-12896.

30. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 Sep;437(7057):376-380.

31. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The Marine Viromes of Four Oceanic Regions. PLoS Biol 2006 Nov;4(11):e368.

32. Schuster SC. Next-generation sequencing transforms today's biology. Nat. Methods 2008 Jan;5(1):16-18.

33. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 2008 Mar;452(7185):340-343.

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990 Oct;215(3):403-410.

35. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 2008 Jan;36(Database issue):D534-538.

36. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H, Cohoon M, de

Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33(17):5691-5702.

37. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 2008 Apr;36(7):2230-2239.

38. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007 Mar;17(3):377-386.

39. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol 2009 Dec;5(12):e1000593.

40. Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol 2002 Aug;184(16):4529-4535.

41. Lozupone C, Hamady M, Knight R. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 2006;7:371.

42. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol 2005 Dec;71(12):8228-8235.

43. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003 Sep;4:41.

44. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004;32(18):5539-5545.

45. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C,

Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. Functional metagenomic profiling of nine biomes. Nature 2008 Apr;452(7187):629-632.

46. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 2007 Mar;5(3):e16.

47. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc. Natl. Acad. Sci. U.S.A 2007 Aug;104(35):13913-13918.

48. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative metagenomics. BMC Bioinformatics 2006;7:162.

49. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol 2009 Apr;5(4):e1000352.

50. Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, Joens LA, Konkel ME, Angly F, Dinsdale EA, Edwards RA, Nelson KE, White BA. Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. PLoS ONE 2008;3(8):e2945.

51. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS ONE 2009 Oct;4(10):e7370.

52. Helton RR, Wommack KE. Seasonal dynamics and metagenomic characterization of estuarine viriobenthos assemblages by randomly amplified polymorphic DNA PCR. Appl. Environ. Microbiol 2009 Apr;75(8):2259-2265.

53. Winget DM, Wommack KE. Diel and daily fluctuations in virioplankton production in coastal ecosystems. Environ. Microbiol 2009 Nov;11(11):2904-2914.

54. Wommack, Ravel, Hill, Colwell. Hybridization analysis of chesapeake bay virioplankton. Appl. Environ. Microbiol 1999 Jan;65(1):241-250.

55. Wommack, Ravel, Hill, Chun, Colwell. Population dynamics of chesapeake bay

virioplankton: total-community analysis by pulsed-field gel electrophoresis. Appl. Environ. Microbiol 1999 Jan;65(1):231-240.

56. Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? Environ. Microbiol 2004 Jan;6(1):1-11.

57. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 2005;6:41.

58. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W,

McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science 2001 Feb;291(5507):1304-1351.

59. Caffrey CR, Rohwer A, Oellien F, Marhöfer RJ, Braschi S, Oliveira G, McKerrow JH, Selzer PM. A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni. PLoS ONE 2009;4(2):e4413.

60. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JAE, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EGJ, Debets AJM, Dekker P, van Dijck PWM, van Dijk A, Dijkhuizen L, Driessen AJM, d'Enfert C, Geysens S, Goosen C, Groot GSP, de Groot PWJ, Guillemette T, Henrissat B, Herweijer M, van den Hombergh JPTW, van den Hondel CAMJJ, van der Heijden RTJM, van der Kaaij RM, Klis FM, Kools HJ, Kubicek CP, van Kuyk PA, Lauber J, Lu X, van der Maarel MJEC, Meulenberg R, Menke H, Mortimer MA, Nielsen J, Oliver SG, Olsthoorn M, Pal K, van Peij NNME, Ram AFJ, Rinas U, Roubos JA, Sagt CMJ, Schmoll M, Sun J, Ussery D, Varga J, Vervecken W, van de Vondervoort PJJ, Wedler H, Wösten HAB, Zeng A, van Ooyen AJJ, Visser J, Stam H. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. Nat. Biotechnol 2007 Feb;25(2):221-231.

61. Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes [Internet]. Environ. Microbiol 2009 Mar;[cited 2010 Mar 29] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19302541

62. Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol 1997 Jun;179(12):3899-3913.

63. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol 1999 Oct;16(10):1391-1399.

64. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res 2005;12(5):281-290.

65. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 2007 Jan;4(1):63-72.

66. Sandberg R, Winberg G, Bränden CI, Kaske A, Ernberg I, Cöster J. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. Genome Res 2001 Aug;11(8):1404-1409.

67. Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol 2004 Sep;6(9):938-947.

68. Gentles AJ, Karlin S. Genome-scale compositional comparisons in eukaryotes. Genome Res 2001 Apr;11(4):540-546.

69. Burge C, Campbell AM, Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. U.S.A 1992 Feb;89(4):1358-1362.

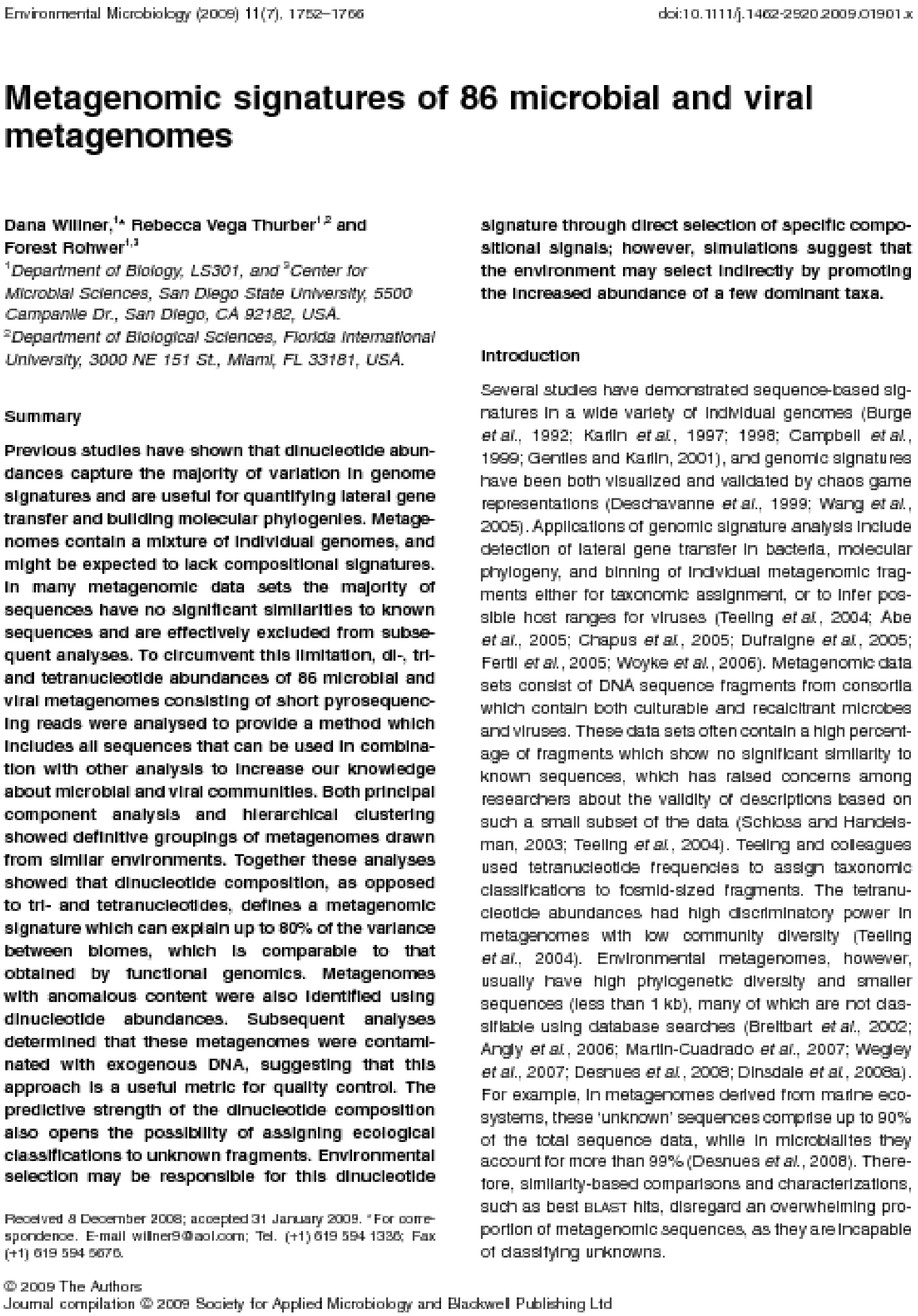
70. Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suciu D, Levine SR, Markowitz VM, Rigoutsos I, Tringe SG, Bruce DC, Richardson PM, Lidstrom ME, Chistoserdova L. High-resolution metagenomics targets specific functional types in complex microbial communities. Nat. Biotechnol 2008 Sep;26(9):1029-1034.

71. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, Persson MAA, Dalianis T, Ramqvist T, Andersson B. Identification of a third human polyomavirus. J. Virol 2007 Apr;81(8):4130-4136.

Acknowledgements

Chapter 1, in full, has been accepted for publication as part of the chapter entitled: Viral metagenomics: from fish slime to the world. Willner, D, Desnues, C, and Rohwer, F. Chapter in "Metagenomics and its applications in agriculture, biomedicine, and environmental studies," Nova Scientific Publishers. The dissertation author was the primary investigator and author of this material.

CHAPTER 2: METAGENOMIC SIGNATURES



Environmental Microbiology (2009) 11(7), 1752–1766 doi:10.1111/j.1462-2920.2009.01901.x

Metagenomic signatures of 86 microbial and viral metagenomes

Dana Willner,[1*] Rebecca Vega Thurber[1,2] and Forest Rohwer[1,3]

[1]Department of Biology, LS301, and [3]Center for Microbial Sciences, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182, USA.
[2]Department of Biological Sciences, Florida International University, 3000 NE 151 St., Miami, FL 33181, USA.

Summary

Previous studies have shown that dinucleotide abundances capture the majority of variation in genome signatures and are useful for quantifying lateral gene transfer and building molecular phylogenies. Metagenomes contain a mixture of individual genomes, and might be expected to lack compositional signatures. In many metagenomic data sets the majority of sequences have no significant similarities to known sequences and are effectively excluded from subsequent analyses. To circumvent this limitation, di-, tri- and tetranucleotide abundances of 86 microbial and viral metagenomes consisting of short pyrosequencing reads were analysed to provide a method which includes all sequences that can be used in combination with other analysis to increase our knowledge about microbial and viral communities. Both principal component analysis and hierarchical clustering showed definitive groupings of metagenomes drawn from similar environments. Together these analyses showed that dinucleotide composition, as opposed to tri- and tetranucleotides, defines a metagenomic signature which can explain up to 80% of the variance between biomes, which is comparable to that obtained by functional genomics. Metagenomes with anomalous content were also identified using dinucleotide abundances. Subsequent analyses determined that these metagenomes were contaminated with exogenous DNA, suggesting that this approach is a useful metric for quality control. The predictive strength of the dinucleotide composition also opens the possibility of assigning ecological classifications to unknown fragments. Environmental selection may be responsible for this dinucleotide signature through direct selection of specific compositional signals; however, simulations suggest that the environment may select indirectly by promoting the increased abundance of a few dominant taxa.

Introduction

Several studies have demonstrated sequence-based signatures in a wide variety of individual genomes (Burge *et al.*, 1992; Karlin *et al.*, 1997; 1998; Campbell *et al.*, 1999; Gentles and Karlin, 2001), and genomic signatures have been both visualized and validated by chaos game representations (Deschavanne *et al.*, 1999; Wang *et al.*, 2005). Applications of genomic signature analysis include detection of lateral gene transfer in bacteria, molecular phylogeny, and binning of individual metagenomic fragments either for taxonomic assignment, or to infer possible host ranges for viruses (Teeling *et al.*, 2004; Abe *et al.*, 2005; Chapus *et al.*, 2005; Dufraigne *et al.*, 2005; Fertil *et al.*, 2005; Woyke *et al.*, 2006). Metagenomic data sets consist of DNA sequence fragments from consortia which contain both culturable and recalcitrant microbes and viruses. These data sets often contain a high percentage of fragments which show no significant similarity to known sequences, which has raised concerns among researchers about the validity of descriptions based on such a small subset of the data (Schloss and Handelsman, 2003; Teeling *et al.*, 2004). Teeling and colleagues used tetranucleotide frequencies to assign taxonomic classifications to fosmid-sized fragments. The tetranucleotide abundances had high discriminatory power in metagenomes with low community diversity (Teeling *et al.*, 2004). Environmental metagenomes, however, usually have high phylogenetic diversity and smaller sequences (less than 1 kb), many of which are not classifiable using database searches (Breitbart *et al.*, 2002; Angly *et al.*, 2006; Martin-Cuadrado *et al.*, 2007; Wegley *et al.*, 2007; Desnues *et al.*, 2008; Dinsdale *et al.*, 2008a). For example, in metagenomes derived from marine ecosystems, these 'unknown' sequences comprise up to 90% of the total sequence data, while in microbialites they account for more than 99% (Desnues *et al.*, 2008). Therefore, similarity-based comparisons and characterizations, such as best BLAST hits, disregard an overwhelming proportion of metagenomic sequences, as they are incapable of classifying unknowns.

Received 8 December 2008; accepted 31 January 2009. *For correspondence. E-mail willner9@aol.com; Tel. (+1) 619 594 1336; Fax (+1) 619 594 5676.

Analysis of the occurrence of oligonucleotide frequencies in eukaryotic, microbial (*Bacteria* and *Archaea*), and viral genomes has demonstrated that individual genomes possess sequence-based signatures, which reflect the specific patterns of dinucleotide abundances (Burge *et al.*, 1992; Karlin and Ladunga, 1994; Karlin and Burge, 1995; Blaisdell *et al.*, 1996; Karlin and Mrazek, 1997; Karlin *et al.*, 1998; Campbell *et al.*, 1999; Gentles and Karlin, 2001). This dinucleotide signature has more phylogenetic signal than genomic GC content, since the percentage of G and C nucleotides can vary widely across genomes (Teeling *et al.*, 2004). The overabundance or relative absence of particular dinucleotides has been linked to DNA structural preferences as well as context-dependent cues, such as potential mutations and regulatory regions (Karlin *et al.*, 1998). For example, vertebrate genomes, and especially the human genome, show a depression of the frequency of CG dinucleotides. This phenomenon has been hypothesized to be driven by the propensity for CG to TA mutations that arise following methylation and subsequent deamination (Burge *et al.*, 1992; Karlin *et al.*, 1998). Additionally, dinucleotide signatures have been shown to be influenced by amino acid preferences and codon biases, which in turn may be driven by the environment (Karlin *et al.*, 1998; Singer and Hickey, 2003; Goodarzi *et al.*, 2007; Paul *et al.*, 2008). Nucleotide frequencies may also reflect environmental conditions such as temperature, pH, metal concentrations and other properties of an organism's habitat (Karlin *et al.*, 1998). Finally, obligate intracellular bacterial parasites and viruses that require endogenous cellular machinery to replicate have genomic signatures that tend to mirror those of the host (Karlin *et al.*, 1998; Campbell *et al.*, 1999).

Dinsdale and colleagues (2008b) recently showed that metagenomes derived from similar environments exhibit similar metabolic profiles, based on functional annotations of component genes. However, as with previous metagenomes, these comparisons were based exclusively on sequences that contained identifiable protein-encoding genes, excluding a large proportion of the sequences from the analysis. Here we present an alternative approach for profiling a nearly identical set of metagenomes based on GC and di-, tri- and tetranucleotide frequencies to determine if related metagenomes show similar oligomeric composition. This was not expected to occur because metagenomes contain a mixture of sequences derived from a variety of individual genomes. Since GC content has been shown to be a poor discriminatory tool in binning of metagenomic sequences, it was used here as a comparison to evaluate the effectiveness of dinucleotides in characterizing metagenomes (Teeling *et al.*, 2004). Both principal component analysis (PCA) and hierarchical clustering showed definitive groupings of metagenomes derived from similar environments based solely on dinucleotide abundances. We hypothesize that these groupings are driven by environmental selection, either for particular microbial and viral taxa with distinctive dinucleotide biases or by direct selection for DNA composition.

Results and discussion

Viral metagenomes have reduced GC content

GC content varies widely between individual genomes, as well as between metagenomes from different environments (Rocha and Danchin, 2002; Foerstner *et al.*, 2005; Raes *et al.*, 2007). Previous studies have used GC content to categorize both genomes and metagenomes and to bin sequences for taxonomic assignment, despite evidence that it performs poorly as a classification metric (Rocha and Danchin, 2002; Teeling *et al.*, 2004; Foerstner *et al.*, 2005; Raes *et al.*, 2007). Here, GC content was used to characterize metagenomes to provide a standard for comparison with the performance of oligomeric abundances in classifying and describing both microbiomes and viromes.

To assess trends in GC content among the 86 metagenomes, average metagenomic GC content and standard deviations were calculated (Tables S1 and S2) and descriptive statistics were compiled by metagenome type (i.e. microbiomes or viromes; Fig. 1) and by biome (Fig. 2). GC content in both the microbial and viral data sets follows an approximately normal distribution (Fig. 1). Overall, the viromes have a lower average GC content, with a mean of 45.19% versus 49.56%. This mirrors the 4% average difference between GC content in phage and their bacterial hosts reported by Rocha and Danchin (2002). Similarly, the average GC content of all microbial and viral genomes available from NCBI (http://www.ncbi.nlm.nih.gov) are 49.70% and 44.32%, respectively, an approximately 5% difference. Although there is no clear consensus, it has been suggested that the increased AT content of viruses may be due either to shorter genome lengths or to an increased energetic cost associated with G and C nucleotides (Rocha and Danchin, 2002).

A two-sample *t*-test to compare the mean GC percentages between viral and microbial metagenomes revealed that GC content was significantly different with a *P*-value less than 0.0001, and the 95% confidence interval for the true mean difference is (2.52, 6.97). The relatively lower overall GC content of the viromes supports the idea that viruses, and especially phage, tend to be more AT-rich than their hosts (Rocha and Danchin, 2002). A single virome from the Arctic was an outlier with a GC content of 62.10%. This result directly contradicts the prevailing

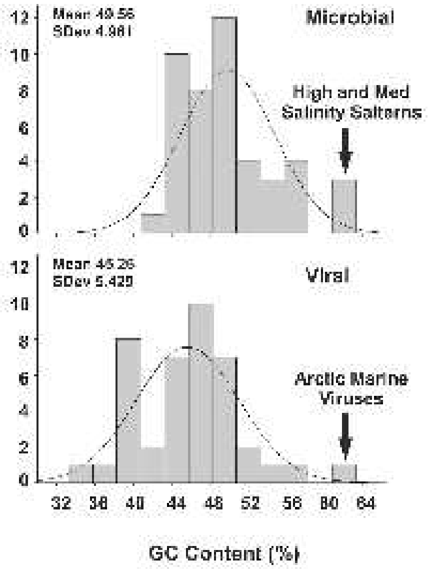


Fig. 1. Frequency histograms of per cent GC content for microbial and viral metagenomes with normal curve fitting and descriptive statistics. Extreme observations are indicated by arrows.

wisdom that genomes from colder environments would have higher AT content (Foerstner *et al.*, 2005).

GC content is not a strong predictor of biomes

GC content among environments/biomes was also determined. GC content generally varied across the biomes with no overall trends. Despite some apparent clusters in Fig. 2, GC content only explained 34.9% and 13.9% of variation in microbial and viral metagenomes, respectively, across biomes (based on adjusted r^2 values from regression analysis). This corroborates previous work which demonstrated that GC content has little discriminatory power for binning of metagenomic fragments (Teeling *et al.*, 2004). Although GC content has been shown to differ significantly between soil and marine metagenomes, no overall trends were observed among the 86 metagenomes in this study which encompassed a larger variety of environments (Foerstner *et al.*, 2005; Raes *et al.*, 2007).

Overview of dinucleotide relative abundances

Some microbial, viral and eukaryotic genomes show significant extremes in individual genomic dinucleotide abundances (Burge *et al.*, 1992; Karlin *et al.*, 1997; Campbell *et al.*, 1999). To assess the over- and under-representation of dinucleotides in individual metagenomes, relative abundance odds ratios, ρ^*_{XY}, were calculated (see *Experimental procedures*). Standard deviations for each relative abundance odds ratio were also calculated to assess the degree of variation in dinucleotide usage between metagenomic sequences.

To compare relative abundance variations in metagenomes with those in individual genomes, 10 random subsets of genomic fragments with an average length of 100 bp were created for two microbial genomes (*Escherichia coli* K-12 and *Halobacterium salinarum* R1). For each fragment set, dinucleotide relative abundance ratios were calculated and averaged and then compared with the calculated dinucleotide usage profile for the entire genome, using the δ^* metric, explained in Karlin and colleagues (1997) and *Experimental procedures*. The same process was also conducted using random sets of 1000 sequences from a medium-salinity solar saltern microbiome and the Christmas Island marine microbiome. The distance between average dinucleotide relative abundance profiles of genomic fragments and microbial genomes was smaller than the distance between relative abundances for metagenomes and their subsets, regardless of coverage (Table S3). This indicates that as

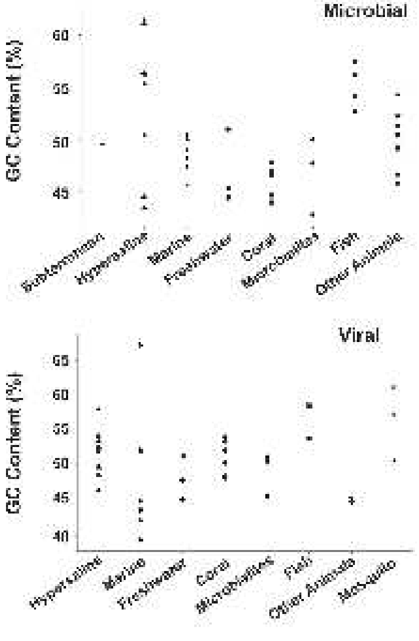


Fig. 2. Scatter plots of per cent GC content by biome for microbial and viral metagenomes.

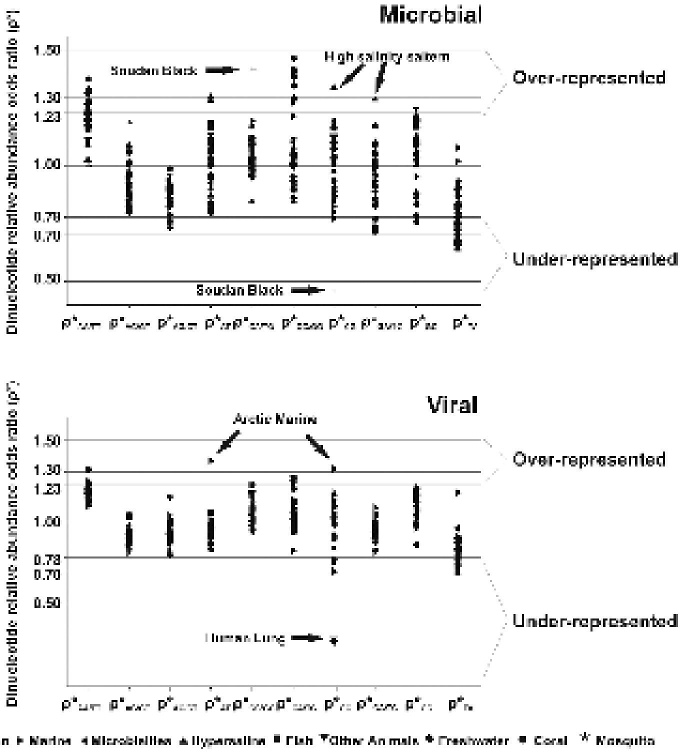


Fig. 3. Dinucleotide relative abundance odds ratios (ρ^*) for all microbial and viral metagenomes. Horizontal lines indicate cut-off points for dinucleotide abundance extremes as described by Karlin and colleagues (1997). Values outside of the normal range, $0.78 < \rho^* < 1.00$, indicate extreme dinucleotide abundances.

compared with individual genomes, metagenomes exhibit more variation in dinucleotide usage, which is expected, since metagenomes are comprised of sequence fragments from a variety of organisms.

The majority of metagenomes did not show any extremes in dinucleotide relative abundances. All microbiomes showed abundances of AC/GT dinucleotides in the normal range ($0.78 < \rho^*_{XY} < 1.00$) under-represent AG/CT, and over-represent AA/TT, although not necessarily in the significant range (Fig. 3; Table S4). Microbial metagenomes also tended to under-represent TA, except for two microbiomes from marine environments. The coral-associated microbiomes derived from *Porites compressa* displayed an overabundance of AA/TT and CC/GG dinucleotides, while the microbiome derived from a second coral species, *Porites astreoides*, did not. This may be due to differences in isolation techniques as discussed below.

As with the microbial metagenomes, the majority of viromes showed no extreme dinucleotide abundances (Fig. 3; Table S5). In general, AC/GT dinucleotides tended to be under-represented and all viromes showed TA depression, although most were outside the significant range. Three coral viromes, two marine viromes, one freshwater virome and the mosquito viromes show elevation of AA/TT, which has been shown to be common in genomes from a wide variety of organisms (Burge *et al.*, 1992).

Dinucleotide biases as a tool for evaluating human contamination in metagenomes

Anomalies in dinucleotide relative abundance odds ratios can be used to quickly identify discrepancies in metagenomes such as human genomic DNA contamination. The Soudan Black microbiome showed an elevation of CA/TG dinucleotides and a severe depression of CG dinucleotides not observed in any other microbiome. This CG depression was indicative of contaminating human DNA sequences in the metagenome. BLASTN analysis was then conducted on this metagenome and previously unidentified human sequences were found (data not shown) (Altschul *et al.*, 1990). This human genomic contamination was introduced during sequencing, as 18S PCR indicated no human genomic DNA in the original samples (data not shown).

As indicated by the dinucleotide biases in Table S5 the two animal virome samples also showed a depression of CG nucleotides in ($\rho^*_{CG} = 0.29$ and $\rho^*_{CG} = 0.27$). These samples were viral particles collected from human lung sputum. As with the Soudan Black microbiome these viral metagenomes have abundances of CG nucleotides in ranges normally exhibited by vertebrate, and especially human DNA (Gentles and Karlin, 2001). Human genomic DNA contamination was confirmed by BLASTN analysis (Altschul *et al.*, 1990). While BLAST analysis can require days to complete, the calculation of dinucleotide relative abundance odds ratios can be performed in a matter of minutes. We therefore suggest that this approach can be used as a quality control metric for human DNA contamination.

Dinucleotide relative abundance distances are large between unrelated metagenomes

To simultaneously compare abundance differences between metagenomes in all 16 dinucleotides, the δ^* statistic was calculated as previously described (Karlin *et al.*, 1997; van Passel *et al.*, 2006). A matrix of adjusted δ^* values for all pairwise comparisons of the 86 metagenomes was generated and scored using quartiles of the observed distribution of δ^* values to define similarity ranges (Fig. 4). The quartiles of the empirical distribution corresponded closely with the empirical cut-off values defined by Karlin and colleagues based on comparisons between reference genomes (Table S6), and therefore were given similar classifications (Karlin *et al.*, 1998).

When compared with all other metagenomes using δ^*, the Soudan Black microbiome appeared very different from any of the other metagenomes, including the other subterranean sample (adjusted δ^* greater than 135 in all cases; Fig. 4; upper arrows), as confirmed by BLASTN analysis. The Arctic viral metagenome was distant from the other viromes, yet it bore weak similarity to many microbiomes from a variety of environments (Fig. 4, lower arrows). Prophage often develop genomic content similar to that of their bacterial hosts (Blaisdell *et al.*, 1996). Consistent with this observation, Angly and colleagues (2006) showed that this virome contains a large prophage signal.

The six *P. compressa* coral microbiomes were more similar to each other than to all other metagenomes ($\delta^* < 90$ in all cases), and were different from the *P. astreoides* metagenome. These six *P. compressa* microbiomes where taken from corals treated with different stressors in aquaria (Vega Thurber *et al.*, 2008). This experiment was performed in Hawaii. In contrast, the *P. astreoides* coral was taken directly from a marine environment near Panama (Wegley *et al.*, 2007). Differences in overall dinucleotide frequencies may correspond to differences in the microbial consortia associated with different coral species and/or different coastal habitats. Alternatively, differences in isolation techniques may be responsible for the dissimilarity between the coral metagenomes, as the *P. astreoides* sample was known to contain mitochondrial DNA.

Three dimensions explain the majority of variance in metagenomes

For both viromes and microbiomes, PCA was conducted to reduce the dimensionality of the set of dinucleotide abundance predictors. While δ^* was used as a summary measure to simultaneously compare all differences between dinucleotide abundances, PCA combined the abundance variables into new variables which better explained dinucleotide differences between metagenomes. The eigenvalues for the first three principal components derived from ρ^*_{XY} values (Table 1) for microbial

Table 1. Eigenvalues and per cent of variance explained for the first three principal components derived from oligonucleotide composition of microbiomes and viromes.

	Dinucleotides			Trinucleotides			Tetranucleotides		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
Microbial									
Eigenvalue	4.77	1.95	1.35	23.86	10.24	7.99	94.37	58.94	26.94
Per cent variance explained	47.7%	19.5%	13.6%	37.3%	16.0%	12.5%	36.9%	23.0%	10.5%
Cumulative per cent variance explained	47.7%	67.2%	80.8%	37.3%	53.3%	65.7%	36.9%	59.9%	70.4%
Viral									
Eigenvalue	3.65	2.14	1.61	15.87	9.56	8.67	62.22	31.13	27.01
Per cent variance explained	36.5%	21.4%	16.1%	24.8%	14.9%	13.6%	24.3%	12.9%	10.6%
Cumulative per cent variance explained	36.5%	57.9%	74.0%	24.8%	39.7%	53.3%	24.3%	37.2%	47.8%

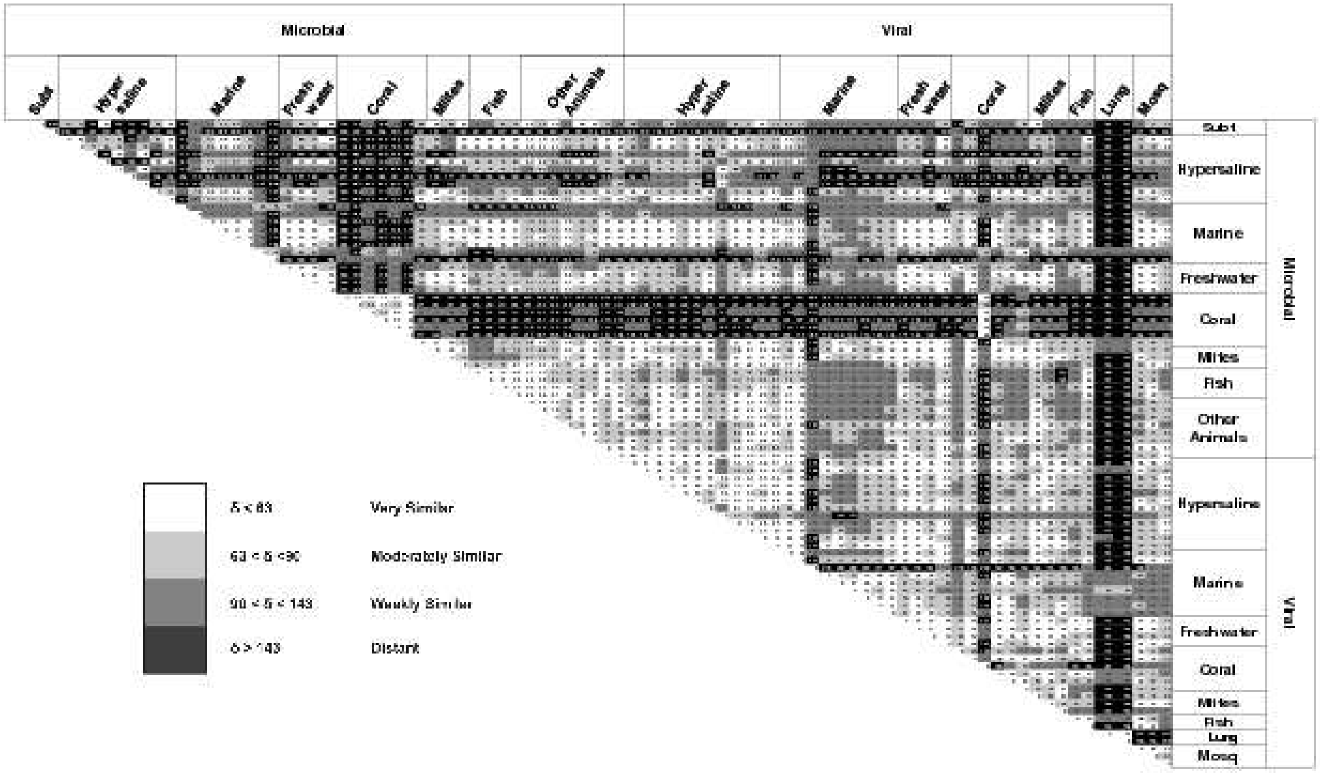


Fig. 4. Dinucleotide relative abundance distance (δ*) values for all pairwise comparisons of metagenomes. Values are shaded according to the degree of similarity between metagenomes based on relative dinucleotide abundances. Subt indicates subterranean microbiomes, Mites indicates microbialites and mosq indicates mosquito viromes.

and viral metagenomes were all greater than one and thus were retained (Quinn and Keough, 2002). For microbiomes, the first three principal components explained 80.8% of the variance between metagenomes. For the viromes, the fourth principal component also had an eigenvalue greater than one; however, this component was excluded, since the first three components explained nearly 74% of the variance. This is comparable to what was obtained by functional genomic analyses using two principal components (~70%; Dinsdale *et al.*, 2008b).

Principal component analysis was also conducted for tri- and tetranucleotide relative abundance odds ratios to determine whether differences in sequence composition of metagenomes could be better explained by higher-order oligonucleotides. Dinucleotides explained the highest proportion of variance between microbial and viral metagenomes (Table 1) as compared with other oligonucleotides. These results support the notion originally posited by Karlin that dinucleotides are sufficient to capture the majority of variations in sequence composition (Karlin *et al.*, 1997).

Previous work by Sandberg and colleagues (2003) demonstrated that longer oligonucleotides are more useful for classification of genomic fragments; however, it is important to note that Sandberg and colleagues utilized raw oligonucleotide frequencies, while this analysis uses a corrected measure, the relative abundance odds ratio (Burge *et al.*, 1992; Karlin *et al.*, 1997). This measure controls for underlying prevalences of lower-order terms, and allows for true determination of oligonucleotide biases (Burge *et al.*, 1992; Karlin *et al.*, 1997). To illustrate the overestimation of discriminatory power which occurs when unadjusted measures are used, PCA was conducted using di-, tri- and tetranucleotide raw frequences (Table S7). In all cases, the per cent of variability explained was inflated when raw frequences were used. However, regardless of whether raw or adjusted frequency measures were used, dinucleotides always explained more of the variability between metagenomes than tri- and tetranucleotides. These results differ from the work of Teeling and colleagues which showed the utility of tetranucleotides for binning of genomic fragments, and also that of Pride and colleagues which demonstrated strong tetranucleotide biases across microbial and phage genomes (Pride *et al.*, 2003; Teeling *et al.*, 2004). While both of these studies did use corrected abundance measures to control for lower-order terms, they considered much longer sequences (from 40 kb to entire genomes) than the metagenomic fragments used in this study, which averaged 100 bp in length (Pride *et al.*, 2003; Teeling *et al.*, 2004). Therefore, while for genomes and longer sequence fragments, higher-order oligonucleotides may provide more discriminatory power, dinucleotides perform best for the description of naturally occurring metagenomic signatures based on short sequence fragments.

Metagenomic clustering based on oligonucleotide relative abundances

Three-dimensional scatter plots of the first three principal components derived from dinucleotide relative abundance odds ratios showed that metagenomes derived from similar biomes cluster together (Fig. 5). All four fish microbiomes (top left, yellow circles) clustered tightly, as did two metagenomes derived from mice (top left, red circles enclosed in black ellipse). The Soudan Black subterranean microbiome fell distinctly outside any distinguishable cluster, reflecting its high level of compositional discordance with other metagenomes, as also demonstrated in frequency analysis using the δ^* statistic. The sample used to generate this metagenome was taken from the reduced mine sediments in Minnesota, an extreme, anoxic environment unlike any others included in this study (Edwards *et al.*, 2006) and it is highly likely that the composition of the microbial community greatly differs from those of the other microbiomes. Additionally, viromes from similar marine environments grouped together (bottom left, blue circles), while the Arctic marine viral metagenome was very distant from all other viromes, due to the high abundance of prophage sequences.

The scatter plots also revealed several interesting trends among metagenomes from related environments. Three of the four metagenomes from the Northern Line Islands (top left, blue circles enclosed in black ellipse) clustered, while the fourth, taken from Christmas Island, fell separately. This fourth microbiome represents Christmas Island, which is the mostly highly inhabited of all four islands (Dinsdale *et al.*, 2008a). Not only does Christmas Island have distinctly different water chemistry from the other three islands, but also the reef-associated microbes are more heterotrophic and pathogenic, as exhibited by the structural and functional 'metabolic profiles' (Dinsdale *et al.*, 2008a,b). Within the coral samples (green circles), all of the *P. compressa* microbial metagenomes clustered tightly near the PC3 axis, while the *P. astreoides* metagenome fell more centrally in the plot. These data again support the hypothesis that there are inherent differences in these samples despite the fact that they were classified *a priori* as belonging to the same biome. Two of the three mosquito viromes (bottom left, black circles) clustered separately from the third. BLAST analyses demonstrated that these two metagenomes were overwhelmingly dominated by sequences from a single-stranded virus of mosquitoes, *Aedes albopictus densovirus* (data not shown). The third mosquito sample was subjected to single-stranded DNA digestion prior to sequencing and containing no sequences with BLAST similarities to this virus.

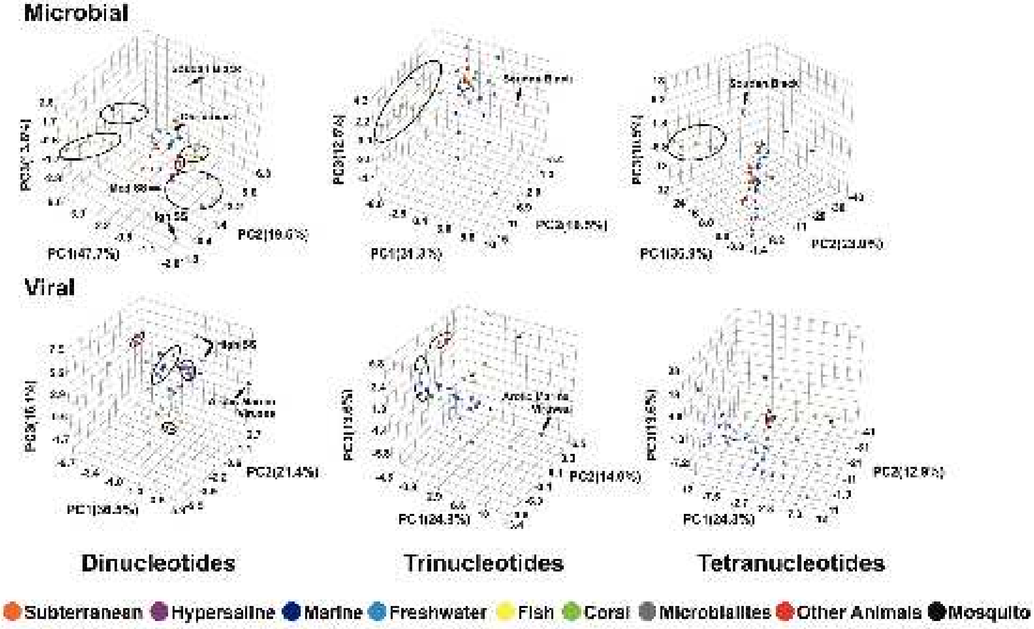


Fig. 5. Three-dimensional scatter plots of principal components from PCA of microbial and viral oligonucleotide frequencies. The per cent of variation each principal component explains is indicated in parentheses adjacent to the component axis.

Interestingly, for both microbiomes and viromes, hypersaline metagenomes (top and bottom left, purple circles) clustered according to a salinity gradient with the high-salinity saltern lying farthest, followed by the medium-salinity salterns, with the low-salinity samples lying more centrally. This corroborates results from previous studies in solar salterns, which show that diversity decreases as saltern ponds become more saline (Benlloch *et al.*, 2002; Casamayor *et al.*, 2002). The decrease in diversity leads to the marked dominance of extreme halophiles at high salinity, thus creating a metagenomic dinucleotide signature dominated by the inputs of relatively few species (Benlloch *et al.*, 2002; Casamayor *et al.*, 2002). This supports the hypothesis that dominant taxa in the environment may drive metagenomic dinucleotide signals.

To further test this dominant taxa hypothesis, we compared the genomic signatures of abundant taxa in medium- and high-salinity salterns to the metagenomic signatures. For each genome identified by BLASTN, dinucleotide relative abundance odds ratios were calculated. Weighted averages of odds ratios were calculated using subsets of the most abundant taxa as described in *Experimental procedures*. For each subset, the distance between the weighted average dinucleotide signature and the metagenomic signature was expressed using the δ^* metric of Karlin and colleagues (1997). As shown by the rank abundance curves, the high-salinity saltern metagenome had lower diversity and was less even than the medium-salinity saltern, with BLASTN hits to fewer known genomes (Fig. 6A and B, Table S8). At high salinity, over 70% of BLAST hits are attributable to only two microbial genomes (*Salinibacter ruber* and *Haloquadratum walsbyi*), while at medium salinity nearly 60 taxa must be included to account for 70% of BLAST hits. When the number of genomes used to calculate the weighted average abundance ratios versus δ^* was plotted, the curves mirrored the BLASTN-based rank–abundance curve for each metagenome (Fig. 6C and D). At high salinity, the distance between the estimated dinucleotide abundances and the true abundances changed little as more genomes were added, while the distance continuously decreased with the addition of genomes at medium salinity. This supports the hypothesis that dominant taxa are driving dinucleotide signatures, since each taxon seems to contribute to the metagenomic dinucleotide signature according to its relative abundance. It should be noted that even when all genomes identified by BLASTN hits are

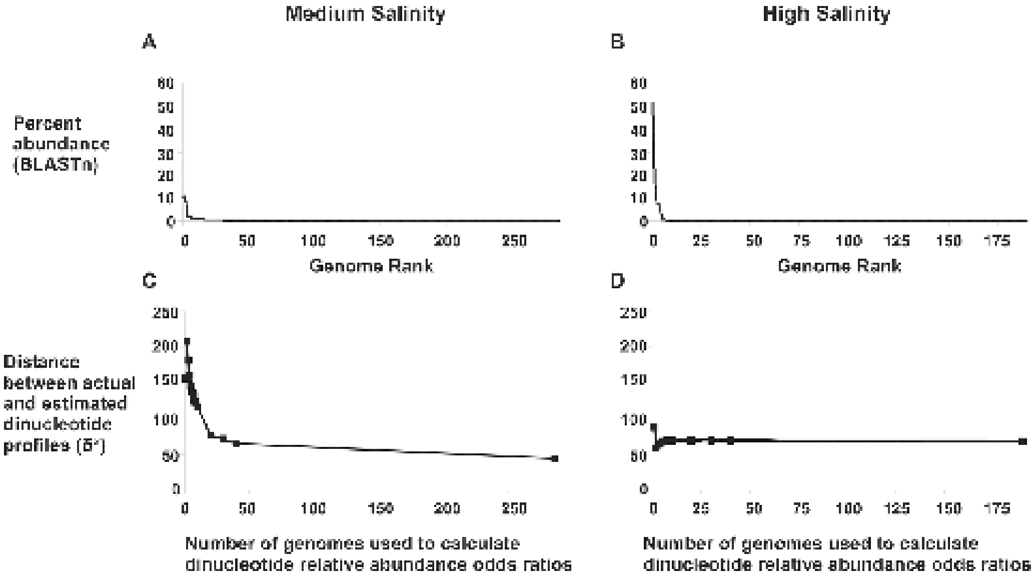


Fig. 6. Rank abundance curves (A and B) and dinucleotide relative abundance distances (C and D) for medium- and high-salinity solar saltern microbiomes.

considered, there is still considerable distance between dinucleotide relative abundance estimates and the true metagenomic signature (δ^* = 44 for medium salinity, and δ^* = 68 for low salinity). This is attributable to the large percentage of sequences in each metagenome with no similarity to known microbial genomes (88% at medium salinity and 68% at high salinity), which may have large contributions to the metagenomic signatures.

Scatter plots created using the first three principal components from PCA with tri- and tetranucleotide relative abundance odds ratios (Fig. 5, centre and right) did not provide any additional clustering of metagenomes. In fact, clusters and overall trends appeared to decline and/or disappear when longer oligonucleotides were used. Dinucleotide relative abundances exhibited substantial discriminatory power to cluster metagenomes by environment and also to provide biologically relevant information about similarities and differences between metagenomic libraries.

Dinucleotide clustering is robust to the addition of metagenomes containing longer sequences

The set of 86 microbial and viral metagenomes used in this study were selected because all samples were processed and sequenced analogously, producing approximately 100 base pair pyrosequencing reads. Since the calculated dinucleotide relative abundance ratios reflect average abundances over the entire metagenome, it might be assumed that the introduction of longer sequence reads would have little effect, although as previously stated, longer sequences may have stronger signatures using higher-order oligonucleotides. Other approaches to metagenomics, such as the use of fosmids and BACs, may be subject to differences in cloning efficiencies which could introduce bias. Additionally, the inclusion of further metagenomes could increase the discriminatory power of dinucleotides if more distinct environments were represented, but if additional data were to come from similar environments, it could also potentially decrease resolution. To determine the effects of the addition of metagenomic libraries on clustering behaviour, 11 microbial and 3 viral metagenomes were added to the analysis. GC content, dinucleotide relative abundance odds ratios and other characteristics of these additional metagenomes are provided in Table S9.

When the 11 microbial metagenomes were added, PCA produced comparable results to the analysis which included only pyrosequenced microbiomes. Previously, dinucleotide relative abundances accounted for 80.6%, with the addition of the other microbiomes; 76.4% of the variation was explained. Despite this reduction, metagenomes still exhibited nearly the same clustering behaviour as in the previous analysis, with the Soudan Black metagenome failing to associate with other microbiomes (Fig. S1, left, orange circle), and six of seven coral metagenomes (green circles) clustering together. The clustering of hypersaline metagenomes (purple circles) once again occurred along a salinity gradient, and the newly added metagenome from a Spanish saltern (purple square) clustered with pre-existing high-salinity solar saltern samples (Legault *et al.*, 2006). Two of the whale fall metagenomes (yellow circles) which were derived from whale bones clustered together, while the third sample, which was taken from a microbial mat, did not (Tringe *et al.*, 2005).

For viromes, the total per cent of variance explained by the first three principal components was 73.8%, a decrease from 78.4% in the previous analysis. Again, however, similar clustering behaviour was observed, with the Arctic viral metagenome and the two known contaminated lung samples (Fig. S1, right, red circles) appearing anomalous. The two hot springs metagenomes (green squares), which were derived from an environment unlike any other in the data set, lay distinctly outside the rest of the points on the scatter plot, displaying a close association with each other but a dramatic difference with the rest of the viromes (Schoenfeld *et al.*, 2008).

Hierarchical clustering by dinucleotide relative abundance odds ratios

Hierarchical clustering was used to quantify the grouping behaviour and trends of the 45 microbial and 41 viral metagenomes demonstrated visually by the three-dimensional scatter plots. Using dinucleotides, both the microbiomes (Fig. 7) and viromes (Fig. S2) formed many clusters containing metagenomes exclusively from identical or similar biomes. Consistent with the PCA results, the fish microbiomes clustered together, as did the mice microbiomes, similar marine viromes and the two contaminated human lung viromes. Additional associations which were not apparent in the scatter plots were clearly delineated in the dendrograms, such as the grouping of chicken and cow rumen microbiomes. Trends in metagenomic clustering, such as the appearance of a salinity gradient, were also consistent with the scatter plot results. Hierarchical clustering of viromes did not segregate metagenomes by environment as well as clustering for microbiomes. This reflects the higher total percentage of variance explained by dinucleotide relative abundances

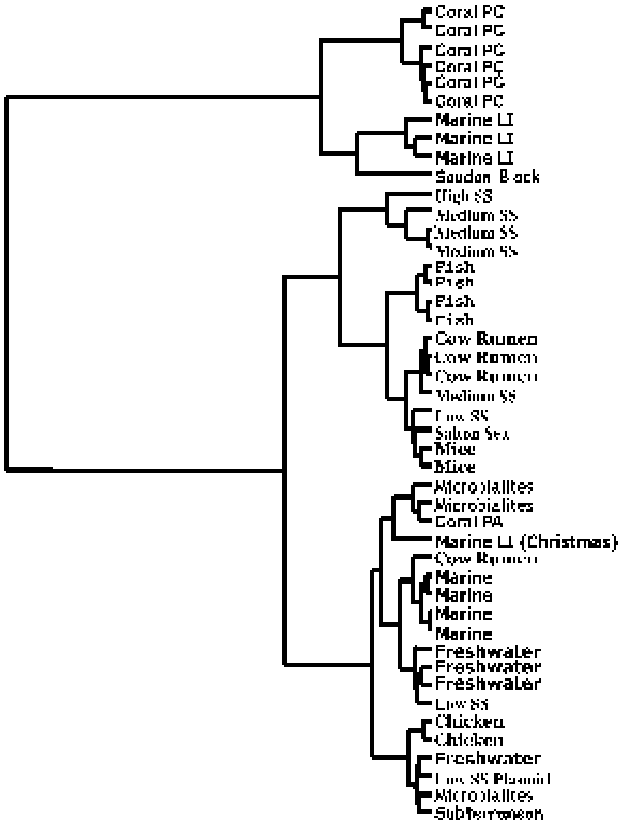


Fig. 7. Hierarchical clustering of microbial metagenomes by the first three principal components from dinucleotide relative abundances. Metagenomes are labelled according to biome. Coral PC indicates *P. compressa* coral microbiomes, Marine LI indicates microbiomes from the four Line Islands: Kiribati, Taburean, Palmyra and Christmas, and *SS* indicates solar salterns.

for microbiomes (80.8%) versus viromes (78.4%). Additionally, there may be unknown similarities between environments which are driving clustering behaviour, creating associations between biomes which have been classified *a priori* as different, and challenging traditional notions of what constitutes a biome.

Caveats

The majority of the metagenomic DNAs were amplified using multiple displacement amplification with Phi29 polymerase prior to sequencing, which could artificially inflate the occurrence of sequences from small circular as well as large linear genomes, and potentially exclude small linear viral genomes, thus biasing dinucleotide frequencies (Pinard *et al.*, 2006; Spits *et al.*, 2006). However, multiple displacement amplification generally provides an even representation of genomes except at the ends, and bias created by amplification would be away from dinucleotide extremes (Dean *et al.*, 2002). Additionally, all of the pyrosequenced metagenomes used in this study were collected and processed in an identical manner, thus equally exposing them to any potential biases due to sampling or amplification. Additionally, it should be noted that the sequences used here were generated from pyrosequencing using the GS20 platform, which has been reported to have an error rate as high as 4% (Huse *et al.*, 2007). However, in practice this error rate has been determined to be much lower, on the order of 0.25% (Huse *et al.*, 2007).

Conclusions

Previous work has demonstrated the presence of distinctive oligonucleotide signatures in a variety of prokaryotic, eukaryotic and viral genomes, as well as marked differences in dinucleotide abundances between the genomes of distantly related organisms (Burge *et al.*, 1992; Karlin and Ladunga, 1994; Blaisdell *et al.*, 1996; Karlin *et al.*, 1997; Gentles and Karlin, 2001; Teeling *et al.*, 2004). Metagenomes represent a diverse cross-section of a particular environmental community, and therefore are not composed of DNA from a single type of organism, but a mixture of DNA from a variety of organisms (Tringe and Rubin, 2005). Initially, we hypothesized that an individual metagenome would not have a characteristic signature, since it essentially represents an averaging of genomes from multiple species. Instead, dinucleotide compositional analysis showed that despite this high level of diversity, metagenomes do have distinct sequence-based signatures. These dinucleotide signatures are driven by environmental selection, in that environments may be dominated by a group of highly abundant taxa whose sequence composition accounts for trends in dinucleotide abundances. Alternatively, the environment itself might be selecting for particular patterns of dinucleotides, irrespective of taxonomy. With the current data set, the metagenomic dinucleotide profiling performs better than profiling using higher-order oligonucleotides, and explains approximately the same proportion of variance between metagenomes as functional genomic analyses. This approach also challenges preconceived notions regarding what constitutes a biome, indicating that the data may carry information that undermines *a priori* biome classifications. The predictive power of this approach suggests that it may be possible to identify anomalous sequences in a manner analogous to that used for individual genomes. The dinucleotide composition was also useful for determining subtle signals in sequence data, such as the presence of contamination, and should be used as a rapid quality check for metagenomes. Together these results show that using dinucleotide abundances allows for more complete characterization of metagenomic content and for rapid comparisons between metagenomes. These analyses could also be used in combination with functional analyses such as those presented in Dinsdale and colleagues (2008b). Since functional annotations rely on similarities to known sequences while compositional analyses do not, environmental clustering of metagenomes by function could be corroborated using dinucleotide signatures, thus providing greater power to discriminate between environments.

Experimental procedures

Data sets

The primary data used for this study consist of pyrosequencing (Roche/454 Life Sciences) reads for a total of 86 metagenomes (45 microbial and 41 viral) derived from nine different biomes, classified as in Dinsdale and colleagues (2008b). Table 2 shows the biome classifications along with how many metagenomes from each biome were used in the analysis. The metagenomic sequences are freely available from both the SEED platform and NCBI, and accession numbers as well as descriptions of the metagenomes are provided in

Table 2. The 86 microbial and viral metagenomes used in the study classified by biome.

Biome	Microbial metagenomes	Viral metagenomes
Subterranean	2	–
Hypersaline	9	12
Marine	8	9
Freshwater	4	4
Coral	7	6
Microbialites	3	3
Fish	4	2
Other animals	8	2
Mosquito	–	3
Total	45	41

Tables S10A and S10B. Data sets are classified as microbial or viral depending on whether sample DNA was extracted for sequencing from a whole microbial fraction or viral fraction derived from caesium chloride density gradient ultracentrifugation as previously described (Angly *et al.*, 2006; Wegley *et al.*, 2007; Desnues *et al.*, 2008; Dinsdale *et al.*, 2008b). Metagenomic sequences have an average length of 100.2 bp, and metagenomic sizes range from 4645 sequences to 699 590 sequences. Additional metagenomic data sets used in PCA were obtained from CAMERA (http://camera.calit2.net).

Frequency tabulation

GC content and mono- and di-, tri-, tetranucleotide counts over each entire metagenome were calculated using a self-written Perl script. This script and all other programs used in this analysis are available at http://sourceforge.net/projects/dinucleotidesig. Counts of N nucleotides (representing a poor sequencing read) as well as oligonucleotides containing N nucleotides were tabulated and in all cases comprised less than 1% of the total frequency data. All N mono- and di- and higher-order oligonucleotides were removed from the data sets prior to further data processing.

The odds ratio measure of dinucleotide bias ρ^*_{XY} was used to evaluate dinucleotides for over- and under-representation in the metagenomes (Burge *et al.*, 1992; Karlin *et al.*, 1997). This measure accounts both for the underlying frequencies of individual mononucleotides and for the complementary nature of double-stranded DNA (Burge *et al.*, 1992; Karlin *et al.*, 1997). Raw frequency values were corrected by averaging the frequency of each mono- or dinucleotide with the frequency of its reverse complement, and assigning the average frequency, f^*, to both (Burge *et al.*, 1992). Following frequency correction, ρ^*_{XY} was calculated for all possible dinucleotides over all metagenomes as $\rho^*_{XY} = \frac{f^*_{XY}}{f^*_X f^*_Y}$ which corrects the observed dinucleotide frequency for the lower-order mononucleotide frequency terms (Karlin *et al.*, 1997; 1998). As shown in *Results*, ρ^*_{XY} values were then classified as normal or extreme according to the given criteria (Karlin *et al.*, 1997; 1998). Standard deviations for dinucleotide relative abundance odds ratios were calculated by determining ρ^*_{XY} values for each individual sequence and then calculating the adjusted average difference between each sequence and ρ^*_{XY} for the whole metagenomes. Relative abundance odds ratios were calculated for trinucleotides using the third-order measure $\gamma^*_{XYZ} = \frac{f^*_{XYZ} f^*_X f^*_Y f^*_Z}{f^*_{XY} f^*_{YZ} f^*_{XNZ}}$, and for tetranucleotides using the fourth-order metric $\tau^*_{XYZW} = \frac{f^*_{XYZW} f^*_{XY} f^*_{XNZ} f^*_{XNNW} f^*_{YZ} f^*_{YNW} f^*_{ZW}}{f^*_{XYZ} f^*_{XYNW} f^*_{XNZW} f^*_X f^*_Y f^*_Z f^*_W}$, where *N* and *M* represent any nucleotide (Karlin *et al.*, 1997).

Calculation of relative abundance differences

Calculation of ρ^*_{XY} allows for comparison of dinucleotide frequencies across an individual metagenome, and identifies potential dinucleotide bias. To compare the overall frequencies between metagenomes, the average dinucleotide relative abundance difference, δ^*, was calculated for all pairwise combinations of the 86 metagenomes (Karlin *et al.*, 1997) using a self-written Java program. The value of δ^* was calculated as $\delta^*(f,g) = \frac{1}{16}\sum_{XY} |\rho^*_{XY}(f) - \rho^*_{XY}(g)|$ where *f* and *g* represent two different metagenomes (Karlin *et al.*, 1997). Descriptive statistics and a graphical summary of the distribution of δ^* values were generated using Minitab Version 15 software (Minitab State College, PA, USA), and values were classified by quartile. The average relative abundance differences reported in *Results* are multiplied by 1000 for easier comparison.

Comparison of dinucleotide relative abundance variation in genomes versus metagenomes

Ten randomly selected sets of 1000 sequences each were selected from a medium-salinity saltern microbiome and the Christmas Island microbiome using a self-written Perl script, to give 0.125× and 0.005× coverage of the metagenomes respectively. The genomes of *E. coli* K-12 substrain MG1655 (NC_000913) and *H. salinarum* R1 (NC_010364) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov), and dinucleotide relative abundance odds ratios were calculated as described above. Random sets of genomic fragments were generated using a self-written Perl script which allows the user to specify the desired genomic coverage as well as an average fragment length. Ten set of fragments with an average length of 100 bp were generated for both 0.125× and 0.005× coverage. For each set of genomic and metagenomic sequences, dinucleotide relative abundance odds ratios were calculated as described above. Relative abundance odds ratios were averaged over 10 repetitions and compared with the dinucleotide profiles for the original genome or metagenome using the δ^* metric of Karlin and colleagues (1997) described above.

BLAST analysis of metagenomes and rank–abundance calculations

Metagenomes were compared with the database of all microbial genomes available from NCBI (http://www.ncbi.nlm.nih.gov) using BLASTN with an *e*-value cut-off of 10^{-5} (Altschul *et al.*, 1990). Complete genomes for each organism detected by BLAST were downloaded from NCBI and dinucleotide relative abundance odds ratios were calculated as described above. To determine the relative contribution of each genome to the metagenomic signature, weighted averages of dinucleotide relative abundances were calculated by multiplying each dinucleotide relative abundance odds ratio for a genome by the percentage of total BLAST hits to that genome in the subset of genomes considered, summing the weighted odds ratios, and then dividing by the number of genomes in the subset.

Statistical analysis

All statistical analysis was performed using Minitab Version 15 software (Minitab, State College, PA, USA). Simple descriptive statistics as well as histograms and box plots of

GC content for microbial and viral genomes were created using the Minitab Descriptive Statistics option. GC content was also explored by biome using a scatter plot generated using the 2D Plot routine. Trends in di-, tri- and tetranucleotide frequencies were examined in Minitab using PCA as well as hierarchical clustering. For all analyses, relative abundance odds ratios were used instead of raw frequency values, because the raw values were too highly correlated to provide meaningful results. Principal component analysis takes a set of correlated variables and reduces them to a smaller set of uncorrelated variables, which can then be used for further analysis (Quinn and Keough, 2002). Principal component analysis was performed in this study separately on the microbial and viral metagenomes, using the correlation association matrix to compensate for unequal variances of predictor variables (Quinn and Keough, 2002). Predictor sets were reduced to three principal components based on eigenvalues in both cases, and these principal components were used to generate three-dimensional scatter plots with the software package Graphis (KyleBank Software, Ayr, UK). Eigenvalues greater than one indicate that a principal component explains more of the variance than would be expected by chance. Since PCA conducted using a correlation matrix standardizes each original variable to have a mean of zero and a variance of 1, the total variance equals the total number of original predictors. Thus, the larger the eigenvalue, the larger proportion of the variance a principal component is explaining (Quinn and Keough, 2002).

Hierarchical clustering using Euclidean distances with Ward linkage was performed on standardized data generated from both the raw values of ρ^{*}_{XY}, γ^{*}_{XYZ} and τ^{*}_{XYZW} and the principal component values generated from the PCA. Both methods assigned the same number of optimal clusters of identical composition.

Acknowledgements

We thank Steve Rayhawk for helpful discussion, and Katie Barott, Elizabeth Dinsdale, Matt Haynes and Linda Wegley for critical readings of the manuscript. This work was supported by the Marine Microbiology Initiative at the Gordon and Betty Moore Foundation (to F.R.) and the Canadian Institute for Advanced Research (CIFAR) Integrated Microbial Biodiversity Program (to F.R.).

References

Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., and Ikemura, T. (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res* **12:** 281–290.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Benlloch, S., Lopez-Lopez, A., Casamayor, E.O., Ovreas, L., Goddard, V., Daae, F.L., *et al.* (2002) Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environ Microbiol* **4:** 349–360.

Blaisdell, B.E., Campbell, A.M., and Karlin, S. (1996) Similarities and dissimilarities of phage genomes. *Proc Natl Acad Sci USA* **93:** 5854–5859.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250–14255.

Burge, C., Campbell, A.M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* **89:** 1358–1362.

Campbell, A., Mrazek, J., and Karlin, S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA* **96:** 9184–9189.

Casamayor, E.O., Massana, R., Benlloch, S., Ovreas, L., Diez, B., Goddard, V.J., *et al.* (2002) Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environ Microbiol* **4:** 338–348.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., and Deschavanne, P. (2005) Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol* **5:** 63.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99:** 5261–5266.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16:** 1391–1399.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452:** 340–343.

Dinsdale, E.A., Pantos, O., Smriga, S., Edwards, R.A., Angly, F., Wegley, L., *et al.* (2008a) Microbial ecology of four coral atolls in the northern line islands. *PLoS ONE* **3:** e1584.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008b) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., and Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* **33:** e6.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7:** 57.

Fertil, B., Massin, M., Lespinats, S., Devic, C., Dumee, P., and Giron, A. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res* **33:** W512–W515.

Foerstner, K.U., von Mering, C., Hooper, S.D., and Bork, P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep* **6:** 1208–1213.

Gentles, A.J., and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* **11:** 540–546.

Goodarzi, H., Torabi, N., Najafabadi, H.S., and Archetti, M. (2007) Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene* **407:** 30–41.

Huse, S., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8:** R:143.

Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283–290.

Karlin, S., and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* **91:** 12832–12836.

Karlin, S., and Mrazek, J. (1997) Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA* **94:** 10227–10232.

Karlin, S., Mrazek, J., and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179:** 3899–3913.

Karlin, S., Campbell, A.M., and Mrazek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32:** 185–225.

Legault, B., Lopez-Lopez, A., Alba-Casado, J.C., Doolittle, W.F., Bolhuis, H., Rodriguez-Valera, F., and Papke, R.T. (2006) Environmental genomics of '*Haloquadratum walsbyi*' in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7:** 171.

Martin-Cuadrado, A.B., Lopez-Garcia, P., Alba, J.C., Moreira, D., Monticelli, L., Strittmatter, A., *et al.* (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2:** e914.

van Passel, M.W., Bart, A., Luyf, A.C., van Kampen, A.H., and van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7:** 26.

Paul, S., Bag, S.K., Das, S., Harvill, E.T., and Dutta, C. (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* **9:** R70.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7:** 216.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13:** 145–158.

Quinn, G.P., and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge, UK: Cambridge University Press.

Raes, J., Foerstner, K.U., and Bork, P. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10:** 490–498.

Rocha, E.P., and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet* **18:** 291–294.

Sandberg, R., Bränden, C.I., Emberg, I., and Cöster, J. (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, animo acid usage and G+C content. *Gene* **11:** 35–42.

Schloss, P.D., and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14:** 303–310.

Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M., and Mead, D. (2008) Assembly of viral metagenomes from Yellowstone Hot Springs. *Appl Environ Microbiol* **74:** 4164–4174.

Singer, G.A.C., and Hickey, D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317:** 39–47.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006) Whole-genome multiple displacement amplification from single cells. *Nat Protoc* **1:** 1965–1970.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6:** 938–947.

Tringe, S.G., and Rubin, E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6:** 805–814.

Tringe, S., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Vega Thurber, R., Barott, K.L., Hall, D., Liu, H., Rodriguez-Mueller, B., Desnues, C., *et al.* (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa. PNAS* **105:** 18413–18418.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* **346:** 173–185.

Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H., and Rohwer, F. (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides. Environ Microbiol* **9:** 2707–2719.

Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443:** 950–955.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Three-dimensional scatter plot of principal components from PCA of microbial and viral dinucleotide frequencies for original and additional metagenomes. The per cent of variation each principal component explains is indicated in parentheses.

Fig. S2. Hierarchical clustering of viromes by the first three principal components from dinucleotide relative abundances. Metagenomes are labelled according to biome. Coral PC indicates *P. compressa* coral viromes, Marine LI indicates viromes from the Line Islands and SS indicates solar salterns.

Table S1. GC content with standard deviations for all microbiomes.

Table S2. GC content with standard deviations for all viromes.

Table S3. Results of simulation to evaluate the degree of variation in dinucleotide relative abundance profiles of genomes versus metagenomes. Metagenomes included are a medium-salinity solar saltern microbiome with SEED ID 4440416.4 and the Christmas Island microbiome with SEED ID 4440041.3.

Table S4. Over- and under-represented dinucleotides in microbiomes. Values outside the normal range are shaded and standard deviations are in parentheses. Values less than 0.78 correspond to under-representations, ρ^*_{XY} greater than 1.23 indicates over-representation. Light grey indicates ρ^*_{XY} less than 0.50, medium grey between 0.50 and 0.70, dark grey between 0.70 and 0.78, charcoal grey between 1.23 and 1.30, and black between 1.30 and 1.50.

Table S5. Over- and under-represented dinucleotides in viral metagenomes. Values of ρ^*_{XY} outside the normal range are shaded and standard deviations are given in parentheses. Values less than 0.78 correspond to under-representations while ρ^*_{XY} greater than 1.23 corresponds to over-representation. Light grey shading indicates ρ^*_{XY} less than 0.50, medium grey between 0.50 and 0.70, dark grey between 0.70 and 0.78, charcoal grey between 1.23 and 1.30, and black between 1.30 and 1.50.

Table S6. Ranges and classifications for the four quartiles of δ^* values (δ^* given as multiplied by 1000) compared with classifications given by Karlin and colleagues (1998).

Table S7. Eigenvalues and per cent of variance explained for the first three principal components derived from raw oligonucleotide frequencies of microbiomes and viromes.

Table S8. Results of BLASTN analysis (*e*-value < 0.00001) comparing a medium-salinity solar saltern microbiome (SEED ID: 444.0416.4) and a high-salinity solar saltern microbiome (SEED ID: 4440419.3) to all microbial genomes in NCBI.

Table S9. Characteristics of additional metagenomes containing long sequence reads used in PCA. All metagenomic sequences were obtained from CAMERA (http://camera.calit2.net). M indicates microbial metagenomes while V indicates viral metagenomes.

Table S10A. Description of microbial metagenomes including SEED and NCBI accession numbers and references.

Table S10B. Description of viral metagenomes including SEED and NCBI accession numbers and references.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

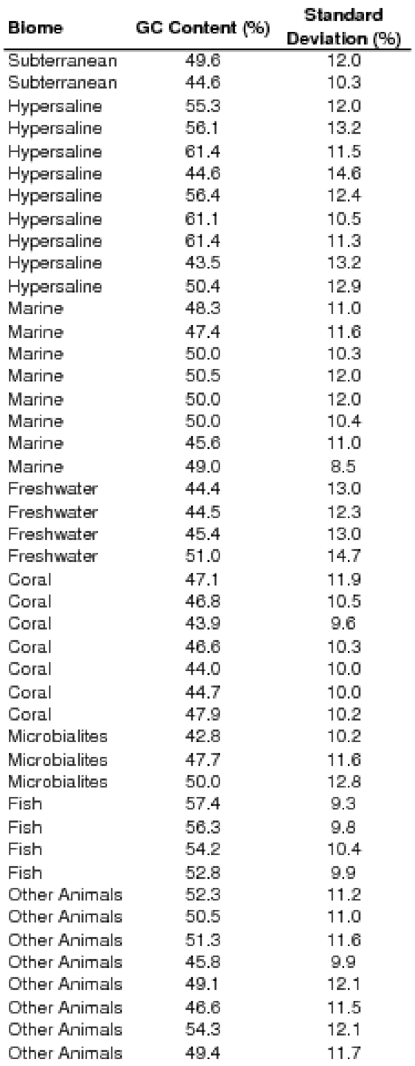
Acknowledgements

Chapter 2 is a reprint of: Willner D**,** Thurber RV, and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology.* (Epub March 18, 2009) The dissertation author was the primary investigator and author of this material. Supporting information for this chapter is presented in the Appendix to this chapter.

Appendix

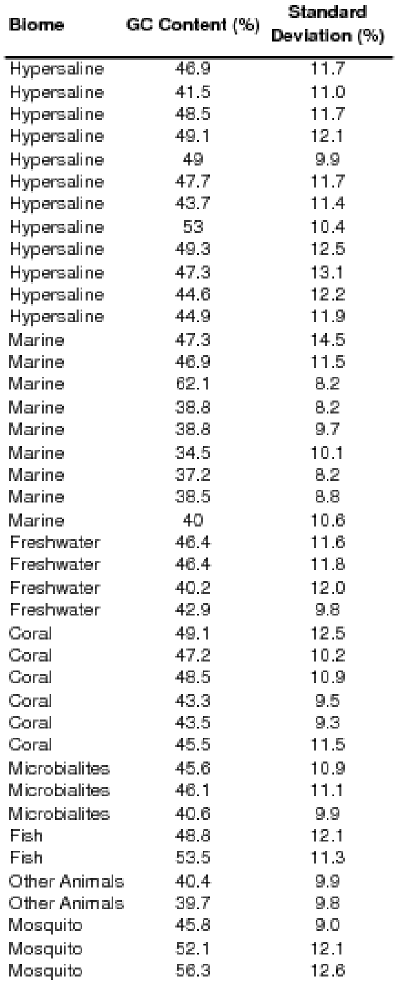
This appendix contains the supporting material published with Willner D**,** Thurber RV, and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology.* (Epub March 18, 2009).

Supplementary Table 1. GC content with standard deviations for all microbiomes.



Biome	GC Content (%)	Standard Deviation (%)
Subterranean	49.6	12.0
Subterranean	44.6	10.3
Hypersaline	55.3	12.0
Hypersaline	56.1	13.2
Hypersaline	61.4	11.5
Hypersaline	44.6	14.6
Hypersaline	56.4	12.4
Hypersaline	61.1	10.5
Hypersaline	61.4	11.3
Hypersaline	43.5	13.2
Hypersaline	50.4	12.9
Marine	48.3	11.0
Marine	47.4	11.6
Marine	50.0	10.3
Marine	50.5	12.0
Marine	50.0	12.0
Marine	50.0	10.4
Marine	45.6	11.0
Marine	49.0	8.5
Freshwater	44.4	13.0
Freshwater	44.5	12.3
Freshwater	45.4	13.0
Freshwater	51.0	14.7
Coral	47.1	11.9
Coral	46.8	10.5
Coral	43.9	9.6
Coral	46.6	10.3
Coral	44.0	10.0
Coral	44.7	10.0
Coral	47.9	10.2
Microbialites	42.8	10.2
Microbialites	47.7	11.6
Microbialites	50.0	12.8
Fish	57.4	9.3
Fish	56.3	9.8
Fish	54.2	10.4
Fish	52.8	9.9
Other Animals	52.3	11.2
Other Animals	50.5	11.0
Other Animals	51.3	11.6
Other Animals	45.8	9.9
Other Animals	49.1	12.1
Other Animals	46.6	11.5
Other Animals	54.3	12.1
Other Animals	49.4	11.7

Supplementary Table 2. GC content with standard deviations for all viromes.



Biome	GC Content (%)	Standard Deviation (%)
Hypersaline	46.9	11.7
Hypersaline	41.5	11.0
Hypersaline	48.5	11.7
Hypersaline	49.1	12.1
Hypersaline	49	9.9
Hypersaline	47.7	11.7
Hypersaline	43.7	11.4
Hypersaline	53	10.4
Hypersaline	49.3	12.5
Hypersaline	47.3	13.1
Hypersaline	44.6	12.2
Hypersaline	44.9	11.9
Marine	47.3	14.5
Marine	46.9	11.5
Marine	62.1	8.2
Marine	38.8	8.2
Marine	38.8	9.7
Marine	34.5	10.1
Marine	37.2	8.2
Marine	38.5	8.8
Marine	40	10.6
Freshwater	46.4	11.6
Freshwater	46.4	11.8
Freshwater	40.2	12.0
Freshwater	42.9	9.8
Coral	49.1	12.5
Coral	47.2	10.2
Coral	48.5	10.9
Coral	43.3	9.5
Coral	43.5	9.3
Coral	45.5	11.5
Microbialites	45.6	10.9
Microbialites	46.1	11.1
Microbialites	40.6	9.9
Fish	48.8	12.1
Fish	53.5	11.3
Other Animals	40.4	9.9
Other Animals	39.7	9.8
Mosquito	45.8	9.0
Mosquito	52.1	12.1
Mosquito	56.3	12.6

Supplementary Table 3. Results of simulation to evaluate the degree of variation in dinucleotide relative abundance profiles of genomes versus metagenomes. Metagenomes included are a medium salinity solar saltern microbiome with SEED ID 4440416.4 and the Christmas Island microbiome with SEED ID 4440041.3.

Metagenome/Genome name	Type	Coverage	Sequences per repetition (avg)	δ^*
Medium salinity saltem microbiome	Metagenome	0.125X	1000	2.44
E. coli K-12 (NC_000913)	Genome	0.125X	5827	0.75
Halobacterium R1 (NC_010364)	Genome	0.125X	2521	0.64
Christmas microbiome	Metagenome	0.005X	1000	2.81
E. coli K-12 (NC_000913)	Genome	0.005X	234	1.64
Halobacterium R1 (NC_010364)	Genome	0.005X	100	2.31

Supplementary Table 4. Over- and under-represented dinucleotides in microbiomes. Values outside the normal range are shaded and standard deviations are in parentheses. Values less than 0.78 correspond to under-representations, ρ*$_{XY}$ greater than 1.23 indicates over-representation. Light grey indicates ρ*$_{XY}$ less than 0.50, medium gray between 0.50 and 0.70, dark gray between 0.70 and 0.78, charcoal gray between 1.23 and 1.30, and black between 1.30 and 1.50.

Biome	ρ*AA/TT	ρ*AC/GT	ρ*AG/CT	ρ*AT	ρ*CA/TG	ρ*CC/GG	ρ*CG	ρ*GA/TC	ρ*GC	ρ*TA
Subterranean	1.33 (0.31)	0.80 (0.25)	0.84 (0.26)	1.02 (0.45)	0.96 (0.37)	1.09 (0.39)	1.10 (0.33)	1.00 (0.32)	1.10 (0.26)	0.72 (0.25)
Subterranean	1.00 (0.42)	1.19 (0.71)	0.94 (0.45)	0.90 (0.35)	1.42 (0.40)	1.06 (0.46)	0.46 (0.44)	0.85 (0.34)	0.88 (0.40)	0.79 (0.66)
Hypersaline	1.25 (0.37)	0.87 (0.31)	0.87 (0.37)	1.06 (0.55)	1.02 (0.55)	0.94 (0.36)	1.12 (0.27)	1.06 (0.39)	1.10 (0.33)	0.67 (0.30)
Hypersaline	1.20 (0.42)	0.88 (0.28)	0.84 (0.26)	1.16 (0.56)	1.01 (0.56)	0.95 (0.36)	1.16 (0.25)	1.06 (0.38)	1.10 (0.31)	0.73 (0.26)
Hypersaline	1.02 (0.44)	0.98 (0.29)	0.83 (0.25)	1.29 (0.61)	1.03 (0.61)	0.90 (0.31)	1.18 (0.23)	1.19 (0.40)	1.00 (0.31)	0.66 (0.27)
Hypersaline	1.27 (0.30)	0.80 (0.28)	0.91 (0.31)	0.95 (0.41)	0.99 (0.41)	1.07 (0.52)	1.08 (0.40)	0.97 (0.33)	1.19 (0.32)	0.76 (0.29)
Hypersaline	1.03 (0.40)	1.00 (0.36)	0.89 (0.32)	1.11 (0.59)	1.12 (0.59)	0.64 (0.43)	1.17 (0.26)	1.11 (0.40)	1.06 (0.36)	0.68 (0.40)
Hypersaline	1.08 (0.36)	1.09 (0.28)	0.85 (0.25)	1.00 (0.49)	0.84 (0.49)	0.86 (0.26)	1.34 (0.20)	1.29 (0.40)	0.88 (0.30)	0.71 (0.28)
Hypersaline	1.03 (0.44)	0.96 (0.29)	0.83 (0.26)	1.30 (0.61)	1.04 (0.61)	0.90 (0.31)	1.20 (0.22)	1.17 (0.39)	1.00 (0.31)	0.64 (0.27)
Hypersaline	1.24 (0.32)	0.93 (0.45)	0.92 (0.42)	0.87 (0.43)	1.08 (0.43)	1.04 (0.54)	0.98 (0.43)	0.92 (0.37)	1.13 (0.42)	0.76 (0.47)
Hypersaline	1.21 (0.31)	0.83 (0.26)	0.89 (0.26)	1.07 (0.41)	0.96 (0.41)	1.06 (0.39)	1.11 (0.31)	1.05 (0.35)	1.05 (0.29)	0.78 (0.26)
Marine	1.11 (0.29)	0.97 (0.26)	0.83 (0.26)	1.09 (0.29)	1.12 (0.29)	1.22 (0.52)	0.83 (0.39)	0.76 (0.27)	1.06 (0.26)	1.02 (0.27)
Marine	1.21 (0.45)	1.06 (0.53)	0.91 (0.41)	0.82 (0.50)	1.20 (0.50)	1.01 (0.56)	0.85 (0.53)	0.96 (0.45)	0.96 (0.42)	0.67 (0.51)
Marine	1.16 (0.29)	0.91 (0.28)	0.90 (0.28)	1.04 (0.42)	1.10 (0.42)	0.94 (0.41)	1.04 (0.28)	0.96 (0.36)	1.19 (0.29)	0.80 (0.28)
Marine	1.18 (0.31)	0.94 (0.30)	0.86 (0.28)	1.01 (0.45)	1.05 (0.45)	0.94 (0.43)	1.10 (0.29)	0.95 (0.37)	1.17 (0.310	0.84 (0.31)
Marine	1.18 (0.32)	0.92 (0.32)	0.86 (0.29)	1.03 (0.46)	1.06 (0.46)	0.96 (0.43)	1.07 (0.31)	0.95 (0.37)	1.17 (0.32)	0.82 (0.32)
Marine	1.17 (0.30)	0.91 (0.29)	0.89 (0.28)	1.03 (0.42)	1.10 (0.42)	0.93 (0.42)	1.04 (.29)	0.95 (0.35)	1.20 (0.30)	0.79 (0.29)
Marine	1.18 (0.35)	0.91 (0.36)	0.86 (0.33)	1.00 (0.40)	1.04 (0.40)	1.23 (0.60)	0.80 (0.46)	0.84 (0.42)	1.01 (0.34)	0.93 (0.34)
Marine	1.17 (0.29)	0.86 (0.27)	0.79 (0.27)	1.17 (0.36)	1.04 (0.36)	1.39 (0.36)	0.77 (0.34)	0.72 (0.39)	1.04 (0.27)	1.08 (0.30)
Freshwater	1.16 (0.37)	0.96 (0.54)	0.82 (0.33)	1.01 (0.38)	1.17 (0.38)	1.02 (0.60)	1.01 (0.46)	0.86 (0.36)	1.19 (0.35)	0.82 (0.50)
Freshwater	1.22 (0.30)	0.90 (0.33)	0.86 (0.29)	0.95 (0.37)	1.09 (0.37)	1.05 (0.51)	1.01 (0.37)	0.89 (0.33)	1.19 (0.30)	0.79 (0.37)
Freshwater	1.24 (0.26)	0.89 (0.32)	0.86 (0.28)	0.96 (0.40)	1.03 (0.40)	1.04 (0.48)	1.10 (0.35)	0.89 (0.34)	1.20 (0.30)	0.83 (0.33)
Freshwater	1.32 (0.23)	0.84 (0.26	0.85 (0.28)	1.00 (0.29)	1.04 (0.29)	1.01 (0.49)	1.12 (0.39)	0.91 (0.29)	1.22 (0.27)	0.73 (0.26)
Coral	1.32 (0.40)	1.06 (0.43)	0.76 (0.32)	0.84 (0.34)	1.00 (0.34)	1.39 (0.51)	0.89 (0.54)	0.73 (0.49)	0.83 (0.340	0.93 (0.38)
Coral	1.38 (0.34)	1.06 (0.36)	0.73 (0.31)	0.80 (0.28)	0.97 (0.26)	1.47 (0.44)	0.87 (0.47)	0.72 (0.48)	0.76 (0.33)	0.91 (0.30)
Coral	1.29 (0.29)	1.02 (0.35)	0.84 (0.30)	0.81 (0.29)	1.01 (0.29)	1.30 (0.48)	0.91 (0.44)	0.84 (0.42)	0.87 (0.32)	0.84 (0.29)
Coral	1.34 (0.33)	1.07 (0.37)	0.76 (0.31)	0.81 (0.29)	1.00 (0.29)	1.42 (0.46)	0.87 (0.48)	0.74 (0.47)	0.78 (0.32)	0.91 (0.32)
Coral	1.32 (0.32)	0.97 (0.32)	0.84 (0.29)	0.83 (0.28)	0.99 (0.26)	1.35 (0.46)	0.87 (0.44)	0.83 (0.42)	0.88 (0.30)	0.84 (0.29)
Coral	1.33 (0.32)	1.01 (0.34)	0.79 (0.31)	0.83 (0.29)	1.00 (0.29)	1.41 (0.45)	0.87 (0.46)	0.77 (0.42)	0.84 (0.31)	0.87 (0.29)
Coral	1.17 (0.30)	0.88 (0.26)	0.99 (0.27)	0.95 (0.34)	1.01 (0.34)	1.04 (0.42)	0.94 (0.36)	1.05 (0.33)	1.03 (0.29)	0.79 (0.28)
Microbialites	1.22 (0.23)	0.84 (0.26)	0.94 (0.27)	0.94 (0.32)	1.01 (0.32)	1.10 (0.56)	0.93 (0.48)	0.98 (0.29)	1.12 (0.28)	0.79 (0.27)
Microbialites	1.18 (0.28)	0.94 (0.26)	0.94 (0.27)	0.92 (0.37)	1.00 (0.37)	1.07 (0.47)	1.03 (0.31)	0.92 (0.33)	1.06 (0.28)	0.90 (0.27)
Microbialites	1.25 (0.32)	0.88 (0.37)	0.84 (0.28)	1.03 (0.42)	1.00 (0.42)	1.02 (0.45)	1.17 (0.35)	0.98 (0.36)	1.10 (0.32)	0.78 (0.38)
Fish	1.18 (0.32)	0.83 (0.24)	0.89 (0.24)	1.19 (0.50)	1.12 (0.50)	0.90 (0.30)	1.11 (0.21)	1.02 (0.36)	1.21 (0.25)	0.64 (0.27)
Fish	1.18 (0.31)	0.83 (0.25)	0.89 (0.24)	1.17 (0.49)	1.11 (0.49)	0.90 (0.30)	1.13 (0.22)	1.01 (0.36)	1.21 (0.25)	0.67 (0.27)
Fish	1.16 (0.30)	0.84 (0.25)	0.92 (0.25)	1.11 (0.46)	1.10 (0.46)	0.93 (0.34)	1.07 (0.25)	1.02 (0.35)	1.17 (0.26)	0.70 (0.27)
Fish	1.18 (0.29)	0.84 (0.25)	0.90 (0.24)	1.10 (0.44)	1.07 (0.44)	0.93 (0.33)	1.12 (0.24)	1.02 (0.35)	1.19 (0.25)	0.73 (0.27)
Other Animals	1.19 (0.32)	0.80 (0.25)	0.90 (0.25)	1.14 (0.43)	0.99 (0.43)	0.99 (0.36)	1.13 (0.27)	1.09 (0.35)	1.08 (0.27)	0.73 (0.26)
Other Animals	1.17 (0.30)	0.81 (0.25)	0.91 (0.25)	1.11 (0.40)	1.01 (0.40)	1.00 (0.36)	1.09 (0.27)	1.08 (0.33)	1.08 (0.27)	0.74 (0.26)
Other Animals	1.19 (0.32)	0.81 (0.25)	0.89 (0.26)	1.11 (0.42)	0.98 (0.42)	1.00 (0.36)	1.16 (0.28)	1.08 (0.35)	1.08 (0.27)	0.77 (0.27)
Other Animals	1.15 (0.26)	0.84 (0.24)	0.96 (0.26)	1.00 (0.36)	1.03 (0.36)	1.00 (0.46)	1.03 (0.30)	0.93 (0.32)	1.25 (0.29)	0.89 (0.26)
Other Animals	1.24 (0.32)	0.80 (0.25)	0.89 (0.26)	1.05 (0.41)	0.95 (0.41)	1.10 (0.45)	1.09 (0.31)	0.99 (0.34)	1.11 (0.26)	0.83 (0.27)
Other Animals	1.20 (0.28)	0.81 (0.24)	0.89 (0.25)	1.06 (0.36)	1.01 (0.36)	1.12 (0.42)	1.02 (0.30)	0.99 (0.31)	1.09 (0.25)	0.81 (0.26)
Other Animals	1.23 (0.41)	0.92 (0.49)	0.86 (0.39)	1.00 (0.55)	1.11 (0.55)	0.95 (0.60)	1.07 (0.61)	1.00 (0.42)	1.11 (0.40)	0.65 (0.49)
Other Animals	1.12 (0.32)	0.93 (0.27)	0.91 (0.29)	1.03 (0.45)	1.03 (0.45)	0.95 (0.42)	1.12 (0.32)	1.05 (0.36)	1.05 (0.32)	0.81 (0.28)

Supplementary Table 5. Over- and under-represented dinucleotides in viral metagenomes. Values of ρ^*_{XY} outside the normal range are shaded and standard deviations are given in parentheses. Values less than 0.78 correspond to under-representations while ρ^*_{XY} greater than 1.23 corresponds to over-representation. Light grey shading indicates ρ^*_{XY} less than 0.50, medium gray between 0.50 and 0.70, dark gray between 0.70 and 0.78, charcoal gray between 1.23 and 1.30, and black between 1.30 and 1.50.

Biome	ρ*AA/TT	ρ*AC/GT	ρ*AG/CT	ρ*AT	ρ*CA/TG	ρ*CC/GG	ρ*CG	ρ*GA/TC	ρ*GC	ρ*TA
Hypersaline	1.18 (0.28)	0.94 (0.27)	0.94 (0.27)	0.94 (0.37)	1.02 (0.43)	0.99 (0.43)	1.05 (0.32)	0.97 (0.34)	1.12 (0.27)	0.85 (0.27)
Hypersaline	1.14 (0.24)	0.94 (0.24)	0.98 (0.27)	0.92 (0.30)	1.05 (0.47)	1.01 (0.48)	0.92 (0.35)	0.96 (0.30)	1.12 (0.27)	0.87 (0.27)
Hypersaline	1.13 (0.31)	0.94 (0.27)	0.92 (0.29)	1.01 (0.43)	1.02 (0.43)	0.95 (0.43)	1.10 (0.31)	1.04 (0.35)	1.07 (0.31)	0.83 (0.27)
Hypersaline	1.14 (0.32)	0.92 (0.27)	0.91 (0.29)	1.01 (0.40)	1.01 (0.43)	0.97 (0.41)	1.13 (0.32)	1.03 (0.35)	1.06 (0.31)	0.82 (0.27)
Hypersaline	1.10 (0.30)	0.91 (0.30)	0.92 (0.27)	1.05 (0.37)	0.99 (0.36)	1.03 (0.38)	1.08 (0.29)	1.09 (0.36)	0.96 (0.26)	0.83 (0.28)
Hypersaline	1.13 (0.28)	0.94 (0.28)	0.90 (0.27)	1.01 (0.37)	1.05 (0.48)	0.98 (0.40)	1.10 (0.29)	1.05 (0.33)	1.01 (0.29)	0.78 (0.29)
Hypersaline	1.18 (0.25)	0.92 (0.26)	0.95 (0.27)	0.91 (0.34)	1.03 (0.46)	0.99 (0.48)	1.05 (0.33)	0.91 (0.31)	1.21 (0.28)	0.87 (0.26)
Hypersaline	1.09 (0.31)	1.01 (0.28)	0.86 (0.26)	1.05 (0.43)	1.02 (0.38)	0.90 (0.35)	1.22 (0.24)	1.12 (0.36)	0.97 (0.30)	0.75 (0.29)
Hypersaline	1.16 (0.33)	0.93 (0.27)	0.91 (0.29)	0.99 (0.46)	0.98 (0.45)	0.97 (0.41)	1.16 (0.33)	1.04 (0.36)	1.04 (0.32)	0.82 (0.28)
Hypersaline	1.19 (0.33)	0.90 (0.27)	0.91 (0.29)	0.97 (0.45)	0.98 (0.48)	0.98 (0.45)	1.16 (0.36)	1.03 (0.34)	1.07 (0.32)	0.80 (0.28)
Hypersaline	1.16 (0.27)	0.86 (0.26)	0.93 (0.27)	1.00 (0.35)	1.01 (0.46)	1.08 (0.45)	1.03 (0.36)	1.00 (0.32)	1.08 (0.28)	0.84 (0.27)
Hypersaline	1.14 (0.27)	0.88 (0.27)	0.94 (0.28)	0.99 (0.36)	1.01 (0.47)	1.05 (0.45)	1.03 (0.35)	1.00 (0.33)	1.08 (0.28)	0.85 (0.27)
Marine	1.12 (0.32)	0.88 (0.29)	0.92 (0.32)	1.06 (0.49)	0.95 (0.49)	0.97 (0.53)	1.16 (0.34)	1.03 (0.38)	1.13 (0.31)	0.91 (0.27)
Marine	1.16 (0.28)	0.90 (0.27)	0.93 (0.28)	0.99 (0.40)	1.01 (0.45)	1.00 (0.43)	1.04 (0.32)	0.99 (0.34)	1.12 (0.28)	0.86 (0.27)
Marine	1.24 (0.41)	0.82 (0.27)	0.81 (0.25)	1.37 (0.60)	0.94 (0.30)	0.82 (0.27)	1.32 (0.20)	1.17 (0.39)	1.18 (0.28)	0.58 (0.27)
Marine	1.12 (0.21)	0.94 (0.26)	1.03 (0.27)	0.90 (0.27)	1.06 (0.46)	1.06 (0.52)	0.75 (0.35)	0.96 (0.27)	1.09 (0.26)	0.89 (0.26)
Marine	1.20 (0.29)	0.88 (0.26)	1.02 (0.26)	0.87 (0.29)	1.02 (0.46)	1.13 (0.52)	0.77 (0.39)	0.95 (0.27)	1.12 (0.26)	0.82 (0.27)
Marine	1.18 (0.22)	0.84 (0.27)	0.99 (0.26)	0.91 (0.25)	1.03 (0.47)	1.23 (0.60)	0.70 (0.45)	0.99 (0.26)	1.09 (0.26)	0.83 (0.26)
Marine	1.11 (0.20)	0.93 (0.26)	1.02 (0.26)	0.92 (0.25)	1.08 (0.44)	1.09 (0.52)	0.70 (0.38)	0.98 (0.26)	1.07 (0.27)	0.87 90.26)
Marine	1.10 (0.21)	0.95 (0.26)	1.00 (0.26)	0.93 (0.27)	1.08 (0.45)	1.07 (0.51)	0.77 (0.37)	0.99 (0.27)	1.03 (0.27)	0.87 (0.27)
Marine	1.10 (0.23)	0.89 (0.26)	0.97 (0.28)	0.98 (0.29)	1.01 (0.48)	1.13 (0.49)	0.91 (0.39)	0.98 (0.29)	1.05 (0.27)	0.91 (0.26)
Freshwater	1.18 (0.26)	0.89 (0.26)	0.92 (0.27)	0.98 (0.36)	1.06 (0.44)	0.98 (0.46)	(0.33)0.3263	0.95 (0.34)	1.20 (0.28)	0.83 (0.27)
Freshwater	1.14 (0.26)	0.93 (0.27)	0.94 (0.28)	0.96 (0.35)	1.05 (0.44)	1.02 (0.45)	1.01 (0.33)	0.94 (0.36)	1.12 (0.27)	0.88 (0.28)
Freshwater	1.17 (0.24)	0.93 (0.26)	0.96 (0.28)	0.90 (0.32)	1.01 (0.53)	1.06 (0.54)	1.02 (0.39)	0.90 (0.31)	1.20 (0.27)	0.90 (0.29)
Freshwater	1.21 (0.29)	0.90 (0.27)	0.92 (0.27)	0.96 (0.40)	1.04 (0.42)	0.96 (0.42)	1.10 (0.31)	0.96 (0.36)	1.16 (0.28)	0.80 (0.27)
Coral	1.12 (0.34)	1.05 (0.46)	0.89 (0.33)	0.92 (0.36)	1.23 (0.56)	0.94 (0.54)	0.92 (0.45)	0.88 (0.35)	1.14 (0.33)	0.80 (0.47)
Coral	1.23 (0.28)	0.89 (0.25)	0.91 (0.28)	0.94 (0.37)	1.09 (0.41)	1.04 (0.41)	0.98 (0.30)	0.93 (0.34)	1.14 (0.28)	0.75 (0.28)
Coral	1.32 (0.34)	1.03 (0.35)	0.80 (0.32)	0.84 (0.35)	0.96 (0.53)	1.27 (0.47)	1.00 (0.45)	0.83 (0.46)	0.86 (0.40)	0.88 (0.31)
Coral	1.14 (0.24)	0.88 (0.26)	1.00 (0.27)	0.94 (0.31)	1.07 (0.45)	1.03 (0.48)	0.90 (0.36)	0.97 (0.30)	1.15 (0.28)	0.83 (0.27)
Coral	1.19 (0.28)	0.95 (0.34)	0.91 (0.31)	0.91 (0.31)	1.09 (0.47)	1.12 (0.45)	0.91 (0.39)	0.92 (0.34)	1.04 (0.32)	0.80 (0.39)
Coral	1.24 (0.30)	0.93 (0.30)	0.88 (0.30)	0.93 (0.32)	0.99 (0.50)	1.20 (0.47)	0.98 (0.41)	0.92 (0.38)	0.97 (0.33)	0.85 (0.29)
Microbialites	1.20 (0.27)	0.85 (0.26)	0.93 (0.27)	0.99 (0.35)	0.97 (0.43)	1.08 (0.43)	1.02 (0.32)	1.01 (0.32)	1.09 (0.28)	0.84 (0.26)
Microbialites	1.18 (0.25)	0.98 (0.29)	0.87 (0.25)	0.96 (0.34)	1.06 (0.41)	1.02 (0.43)	1.05 (0.30)	0.90 (0.31)	1.12 (0.26)	0.88 (0.25)
Microbialites	1.11 (0.23)	0.96 (0.27)	0.97 (0.29)	0.94 (0.30)	0.95 (0.55)	1.12 (0.50)	1.02 (0.39)	0.95 (0.31)	1.00 (0.29)	0.96 (0.30)
Fish	1.12 (0.31)	0.93 (0.41)	0.97 (0.31)	0.97 (0.45)	1.24 (0.28)	0.98 (0.50)	0.84 (0.33)	0.95 (0.34)	1.14 (0.30)	0.73 (0.39)
Fish	1.17 (0.32)	0.87 (0.34)	0.92 (0.28)	1.06 (0.49)	1.16 (0.43)	0.93 (0.37)	1.02 (0.26)	0.97 (0.35)	1.19 (0.27)	0.70 (0.33)
Other Animals	1.13 (0.25)	0.85 (0.29)	1.16 (0.34)	0.86 (0.29)	1.17 (0.33)	1.24 (0.50)	0.29 (0.39)	1.00 (0.28)	0.97 (0.33)	0.76 (0.30)
Other Animals	1.13 (0.23)	0.85 (0.26)	1.14 (0.28)	0.87 (0.27)	1.18 (0.30)	1.26 (0.49)	0.27 (0.37)	0.98 (0.27)	0.98 (0.27)	0.77 (0.27)
Mosquito	1.20 (0.26)	0.96 (0.33)	0.86 (0.27)	0.96 (0.34)	1.13 (0.43)	0.98 (0.45)	1.02 (0.31)	0.92 (0.31)	1.16 (0.28)	0.77 (0.32)
Mosquito	1.24 (0.33)	0.90 (0.34)	0.86 (0.29)	1.01 (0.47)	1.12 (0.46)	0.99 (0.39)	1.05 (0.29)	0.90 (0.36)	1.18 (0.31)	0.73 (0.37)
Mosquito	1.25 (0.37)	0.97 (0.30)	0.27)0.2694!	1.00 (0.48)	0.99 (0.47)	0.98 (0.57)	1.18 (0.35)	0.99 (0.47)	1.03 (0.30)	0.79 (0.34)

Supplementary Table 6. Ranges and classifications for the four quartiles of δ* values (δ* given as multiplied by 1000) compared with classifications given by Karlin et al. (Karlin, 1998).

Quartile Range	Classification	Reference Range(s) and Classification (Karlin 1998)
$\delta^* < 63$	Very Similar	$\delta^* \leq 50$ Close
$63 \leq \delta^* < 90$	Moderately Similar	$55 \leq \delta^* \leq 85$ Moderately Similar
$90 \leq \delta^* < 143$	Weakly Similar	$90 \leq \delta^* \leq 120$ Weakly Similar $125 \leq \delta^* \leq 145$ Distantly Similar
$\delta^* \geq 143$	Distant	$150 \leq \delta^* \leq 180$ Distant $125 \leq \delta^* \leq 145$ Very Distant

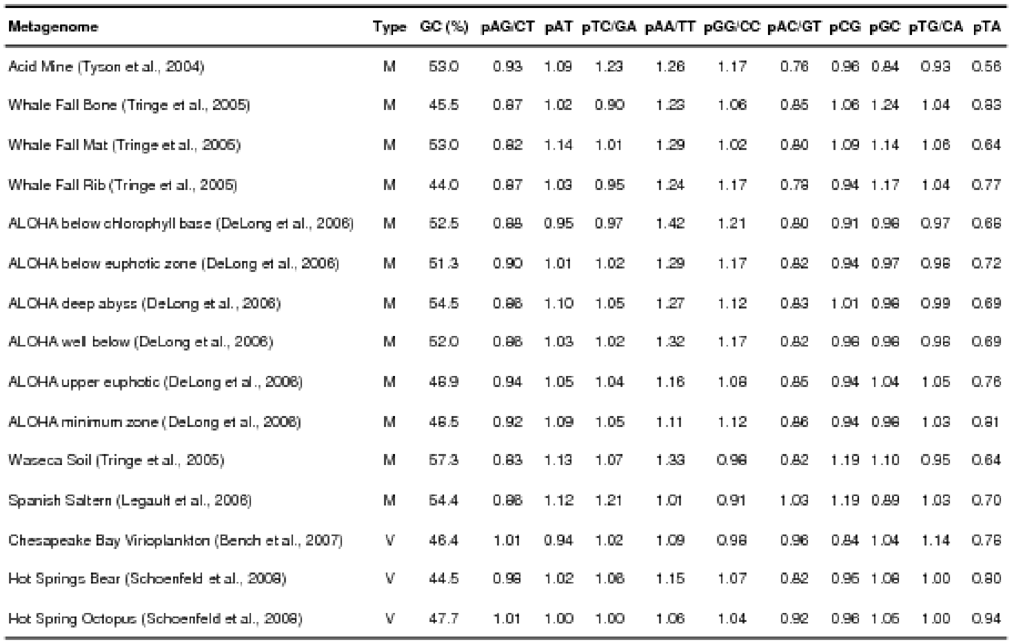
Supplementary Table 7. Eigenvalues and percent of variance explained for the first three principal components dervied from raw oligonucleotide frequencies of microbiomes and viromes.

	Dinucleotides			Trinucleotides			Tetranucleotides		
Microbial	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
Eigenvalue	8.58	2.92	2.59	29.56	13.85	5.82	102.87	58.88	20.5
Percent Variance Explained	44.0%	26.1%	15.5%	46.2%	21.6%	9.1%	40.2%	23.0%	8.0%
Cumulative Percent Variance Explained	44.0%	70.1%	85.6%	46.2%	67.8%	76.9%	40.2%	63.2%	71.2%
Viral									
Eigenvalue	7.04	4.18	2.48	35.84	8.05	6.27	126.37	33.41	26.34
Percent Variance Explained	53.6%	18.2%	16.2%	56.0%	12.6%	9.8%	49.4%	13.1%	10.3%
Cumulative Percent Variance Explained	53.6%	71.8%	88.1%	56.0%	68.6%	78.4%	49.4%	62.5%	72.7%

Supplementary Table 8. Results of BLASTN analysis (e-value<0.00001) comparing a medium salinity solar saltern microbiome (SEED ID: 444.0416.4) and a high salinity solar saltern microbiome (SEED ID: 4440419.3) to all microbial genomes in NCBI.

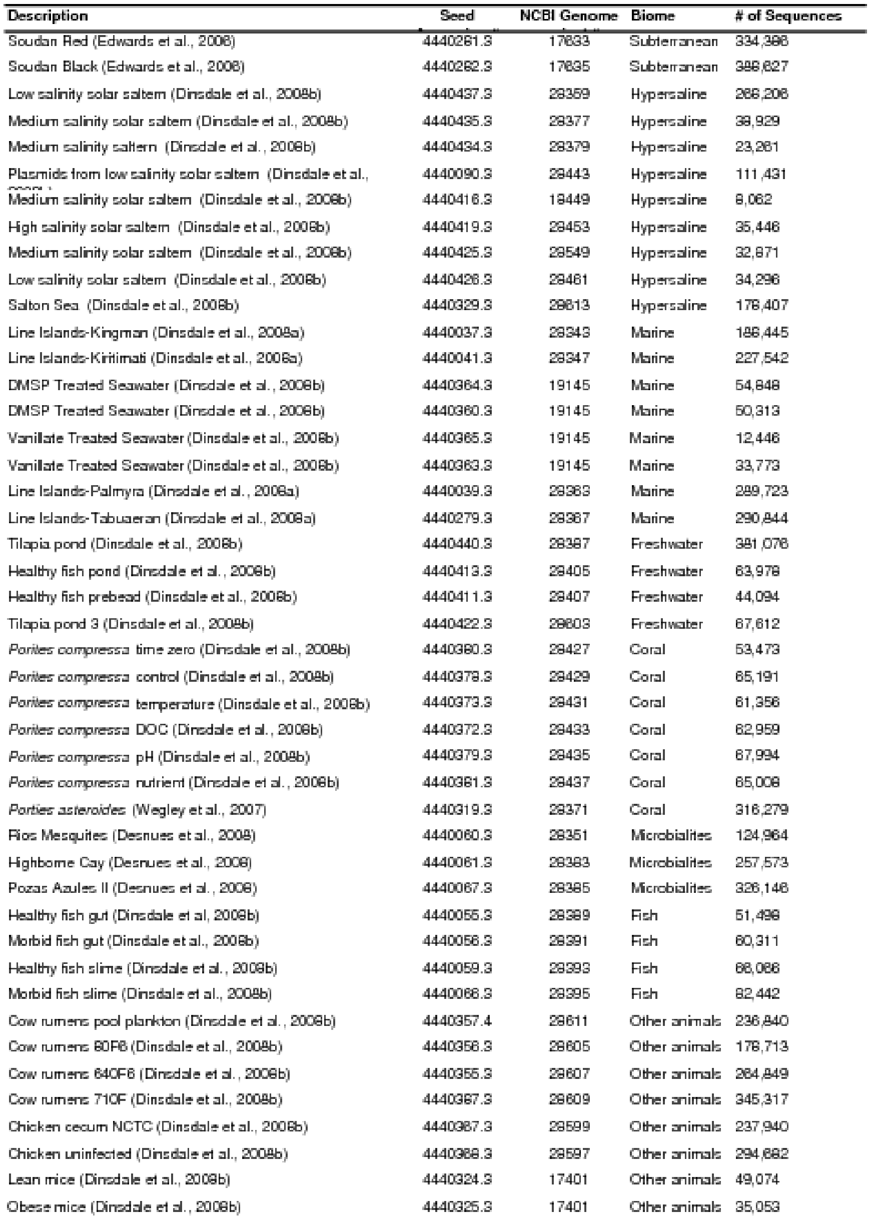
Metagenome	Number of hits (% of metagenome)	Number of unique taxa
Medium salinity solar saltern microbiome	947 (12%)	286
High salinity solar saltern microbiome	11478 (32%)	190

Supplementary Table 9. Characteristics of additional metagenomes containing long sequence reads used in PCA. All metagenomic sequences were obtained from CAMERA (http://camera.calit2.net). M indicates microbial metagenomes while V indicates viral metagenomes.



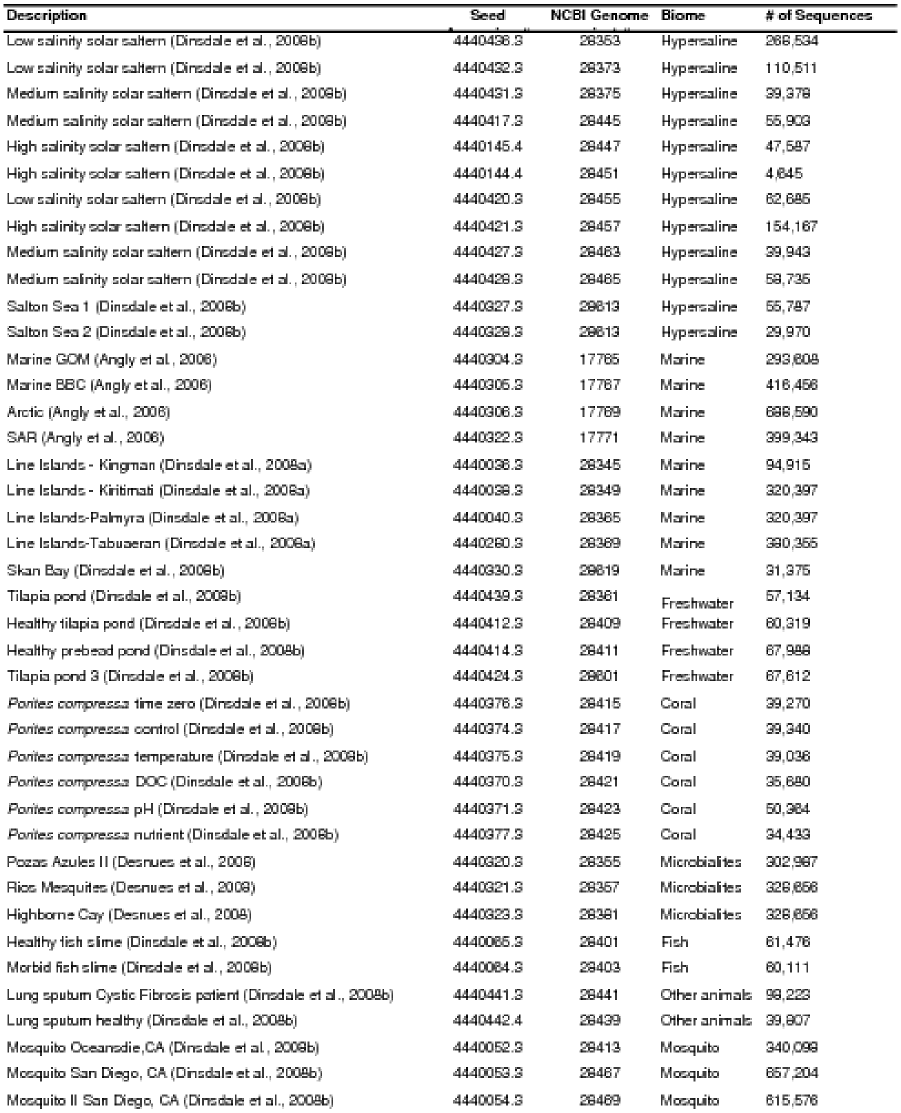
Metagenome	Type	GC (%)	pAG/CT	pAT	pTC/GA	pAA/TT	pGG/CC	pAC/GT	pCG	pGC	pTG/CA	pTA
Acid Mine (Tyson et al., 2004)	M	53.0	0.93	1.09	1.23	1.26	1.17	0.76	0.96	0.84	0.93	0.56
Whale Fall Bone (Tringe et al., 2005)	M	45.5	0.87	1.02	0.90	1.23	1.06	0.85	1.06	1.24	1.04	0.83
Whale Fall Mat (Tringe et al., 2005)	M	53.0	0.82	1.14	1.01	1.29	1.02	0.80	1.09	1.14	1.06	0.64
Whale Fall Rib (Tringe et al., 2005)	M	44.0	0.87	1.03	0.95	1.24	1.17	0.78	0.94	1.17	1.04	0.77
ALOHA below chlorophyll base (DeLong et al., 2006)	M	52.5	0.88	0.95	0.97	1.42	1.21	0.80	0.91	0.98	0.97	0.66
ALOHA below euphotic zone (DeLong et al., 2006)	M	51.3	0.90	1.01	1.02	1.29	1.17	0.82	0.94	0.97	0.96	0.72
ALOHA deep abyss (DeLong et al., 2006)	M	54.5	0.86	1.10	1.05	1.27	1.12	0.83	1.01	0.98	0.99	0.69
ALOHA well below (DeLong et al., 2006)	M	52.0	0.86	1.03	1.02	1.32	1.17	0.82	0.98	0.98	0.96	0.69
ALOHA upper euphotic (DeLong et al., 2006)	M	48.9	0.94	1.05	1.04	1.16	1.08	0.85	0.94	1.04	1.05	0.76
ALOHA minimum zone (DeLong et al., 2006)	M	48.5	0.92	1.09	1.05	1.11	1.12	0.86	0.94	0.98	1.03	0.81
Waseca Soil (Tringe et al., 2005)	M	57.3	0.83	1.13	1.07	1.33	0.98	0.82	1.19	1.10	0.95	0.64
Spanish Saltern (Legault et al., 2006)	M	54.4	0.86	1.12	1.21	1.01	0.91	1.03	1.19	0.89	1.03	0.70
Chesapeake Bay Virioplankton (Bench et al., 2007)	V	46.4	1.01	0.94	1.02	1.09	0.98	0.96	0.84	1.04	1.14	0.76
Hot Springs Bear (Schoenfeld et al., 2008)	V	44.5	0.98	1.02	1.06	1.15	1.07	0.82	0.95	1.08	1.00	0.80
Hot Spring Octopus (Schoenfeld et al., 2008)	V	47.7	1.01	1.00	1.00	1.06	1.04	0.92	0.96	1.05	1.00	0.94

Supplementary Table 10A. Description of microbial metagenomes including SEED and NCBI accession numbers and references.



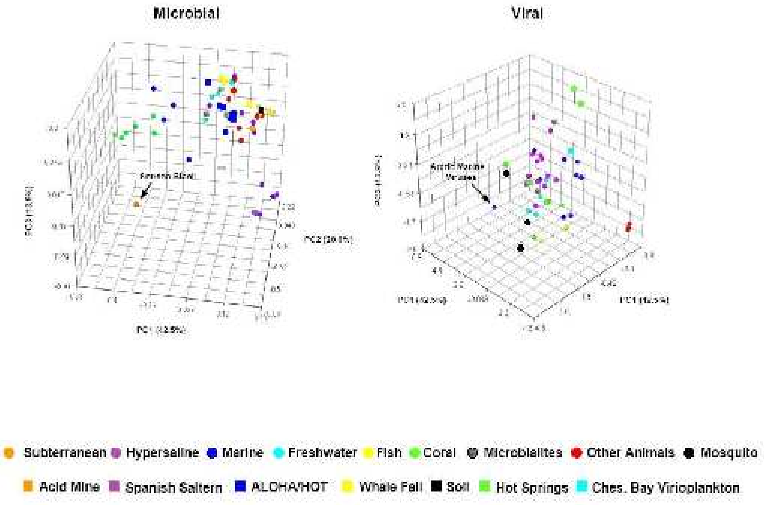
Description	Seed	NCBI Genome	Biome	# of Sequences
Soudan Red (Edwards et al., 2006)	4440281.3	17633	Subterranean	334,386
Soudan Black (Edwards et al., 2006)	4440282.3	17635	Subterranean	386,627
Low salinity solar saltem (Dinsdale et al., 2008b)	4440437.3	28359	Hypersaline	266,206
Medium salinity solar saltem (Dinsdale et al., 2008b)	4440435.3	28377	Hypersaline	38,929
Medium salinity saltern (Dinsdale et al., 2008b)	4440434.3	28379	Hypersaline	23,261
Plasmids from low salinity solar saltem (Dinsdale et al.,	4440090.3	28443	Hypersaline	111,431
Medium salinity solar saltem (Dinsdale et al., 2008b)	4440416.3	18449	Hypersaline	8,062
High salinity solar saltem (Dinsdale et al., 2008b)	4440419.3	28453	Hypersaline	35,446
Medium salinity solar saltem (Dinsdale et al., 2008b)	4440425.3	28549	Hypersaline	32,871
Low salinity solar saltem (Dinsdale et al., 2008b)	4440426.3	28461	Hypersaline	34,296
Salton Sea (Dinsdale et al., 2008b)	4440329.3	28613	Hypersaline	176,407
Line Islands-Kingman (Dinsdale et al., 2008a)	4440037.3	28343	Marine	186,445
Line Islands-Kiritimati (Dinsdale et al., 2008a)	4440041.3	28347	Marine	227,542
DMSP Treated Seawater (Dinsdale et al., 2008b)	4440364.3	19145	Marine	54,848
DMSP Treated Seawater (Dinsdale et al., 2008b)	4440360.3	19145	Marine	50,313
Vanillate Treated Seawater (Dinsdale et al., 2008b)	4440365.3	19145	Marine	12,446
Vanillate Treated Seawater (Dinsdale et al., 2008b)	4440363.3	19145	Marine	33,773
Line Islands-Palmyra (Dinsdale et al., 2008a)	4440039.3	28363	Marine	289,723
Line Islands-Tabuaeran (Dinsdale et al., 2008a)	4440279.3	28367	Marine	290,844
Tilapia pond (Dinsdale et al., 2008b)	4440440.3	28387	Freshwater	381,076
Healthy fish pond (Dinsdale et al., 2008b)	4440413.3	28405	Freshwater	63,978
Healthy fish prebead (Dinsdale et al., 2008b)	4440411.3	28407	Freshwater	44,094
Tilapia pond 3 (Dinsdale et al., 2008b)	4440422.3	28603	Freshwater	67,612
Porites compressa time zero (Dinsdale et al., 2008b)	4440380.3	28427	Coral	53,473
Porites compressa control (Dinsdale et al., 2008b)	4440378.3	28429	Coral	65,191
Porites compressa temperature (Dinsdale et al., 2008b)	4440373.3	28431	Coral	61,356
Porites compressa DOC (Dinsdale et al., 2008b)	4440372.3	28433	Coral	62,959
Porites compressa pH (Dinsdale et al., 2008b)	4440379.3	28435	Coral	67,994
Porites compressa nutrient (Dinsdale et al., 2008b)	4440381.3	28437	Coral	65,008
Porties asteroides (Wegley et al., 2007)	4440319.3	28371	Coral	316,279
Rios Mesquites (Desnues et al., 2008)	4440060.3	28351	Microbialites	124,964
Highborne Cay (Desnues et al., 2008)	4440061.3	28383	Microbialites	257,573
Pozas Azules II (Desnues et al., 2008)	4440067.3	28385	Microbialites	326,146
Healthy fish gut (Dinsdale et al, 2008b)	4440055.3	28389	Fish	51,498
Morbid fish gut (Dinsdale et al., 2008b)	4440056.3	28391	Fish	60,311
Healthy fish slime (Dinsdale et al., 2008b)	4440059.3	28393	Fish	66,066
Morbid fish slime (Dinsdale et al., 2008b)	4440066.3	28395	Fish	82,442
Cow rumens pool plankton (Dinsdale et al., 2008b)	4440357.4	28611	Other animals	236,840
Cow rumens 80F6 (Dinsdale et al., 2008b)	4440356.3	28605	Other animals	176,713
Cow rumens 640F6 (Dinsdale et al., 2008b)	4440355.3	28607	Other animals	264,849
Cow rumens 710F (Dinsdale et al., 2008b)	4440367.3	28609	Other animals	345,317
Chicken cecum NCTC (Dinsdale et al., 2008b)	4440367.3	28599	Other animals	237,940
Chicken uninfected (Dinsdale et al., 2008b)	4440368.3	28597	Other animals	294,682
Lean mice (Dinsdale et al., 2008b)	4440324.3	17401	Other animals	49,074
Obese mice (Dinsdale et al., 2008b)	4440325.3	17401	Other animals	35,053

Supplementary Table 10B. Description of viral metagenomes including SEED and NCBI accession numbers and references.

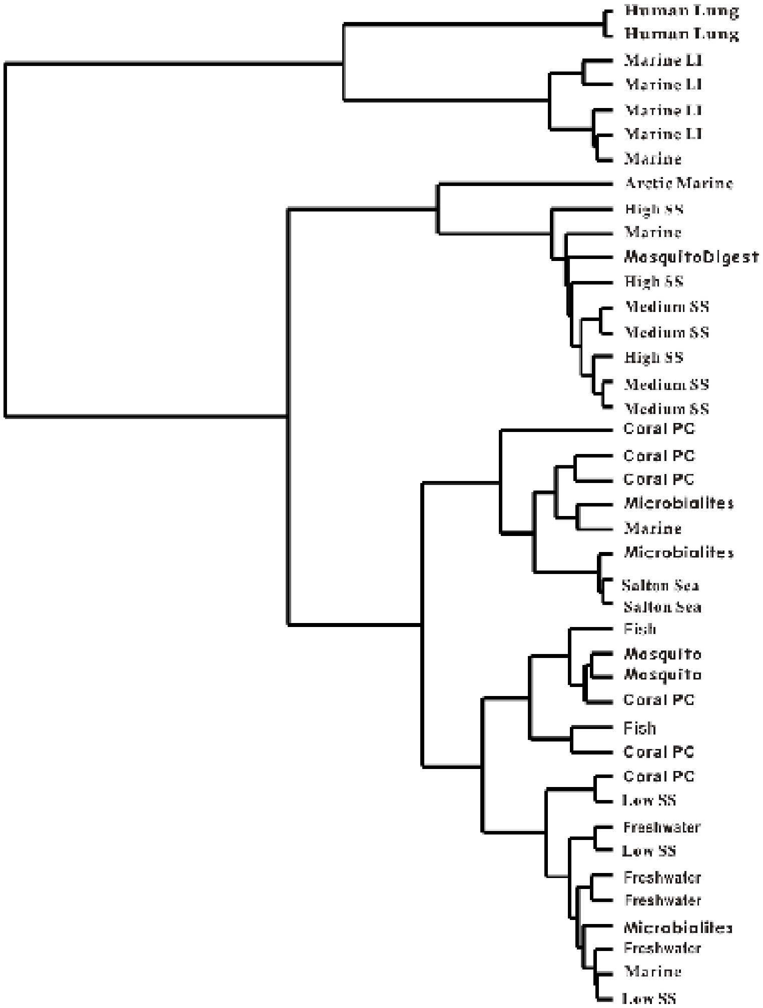


Description	Seed	NCBI Genome	Biome	# of Sequences
Low salinity solar saltern (Dinsdale et al., 2008b)	4440436.3	28353	Hypersaline	266,534
Low salinity solar saltern (Dinsdale et al., 2008b)	4440432.3	28373	Hypersaline	110,511
Medium salinity solar saltern (Dinsdale et al., 2008b)	4440431.3	28375	Hypersaline	39,378
Medium salinity solar saltern (Dinsdale et al., 2008b)	4440417.3	28445	Hypersaline	55,903
High salinity solar saltern (Dinsdale et al., 2008b)	4440145.4	28447	Hypersaline	47,587
High salinity solar saltern (Dinsdale et al., 2008b)	4440144.4	28451	Hypersaline	4,645
Low salinity solar saltern (Dinsdale et al., 2008b)	4440420.3	28455	Hypersaline	62,685
High salinity solar saltern (Dinsdale et al., 2008b)	4440421.3	28457	Hypersaline	154,167
Medium salinity solar saltern (Dinsdale et al., 2008b)	4440427.3	28463	Hypersaline	39,943
Medium salinity solar saltern (Dinsdale et al., 2008b)	4440428.3	28465	Hypersaline	58,735
Salton Sea 1 (Dinsdale et al., 2008b)	4440327.3	29613	Hypersaline	55,787
Salton Sea 2 (Dinsdale et al., 2008b)	4440328.3	29613	Hypersaline	29,970
Marine GOM (Angly et al., 2006)	4440304.3	17765	Marine	293,608
Marine BBC (Angly et al., 2006)	4440305.3	17767	Marine	416,456
Arctic (Angly et al., 2006)	4440306.3	17769	Marine	686,590
SAR (Angly et al., 2006)	4440322.3	17771	Marine	399,343
Line Islands - Kingman (Dinsdale et al., 2008a)	4440036.3	28345	Marine	94,915
Line Islands - Kiritimati (Dinsdale et al., 2008a)	4440038.3	28349	Marine	320,397
Line Islands-Palmyra (Dinsdale et al., 2008a)	4440040.3	28365	Marine	320,397
Line Islands-Tabuaeran (Dinsdale et al., 2008a)	4440280.3	28369	Marine	380,355
Skan Bay (Dinsdale et al., 2008b)	4440330.3	29619	Marine	31,375
Tilapia pond (Dinsdale et al., 2008b)	4440439.3	28361	Freshwater	57,134
Healthy tilapia pond (Dinsdale et al., 2008b)	4440412.3	28409	Freshwater	60,319
Healthy prebead pond (Dinsdale et al., 2008b)	4440414.3	28411	Freshwater	67,988
Tilapia pond 3 (Dinsdale et al., 2008b)	4440424.3	29601	Freshwater	67,612
Porites compressa time zero (Dinsdale et al., 2008b)	4440376.3	28415	Coral	39,270
Porites compressa control (Dinsdale et al., 2008b)	4440374.3	28417	Coral	39,340
Porites compressa temperature (Dinsdale et al., 2008b)	4440375.3	28419	Coral	39,036
Porites compressa DOC (Dinsdale et al., 2008b)	4440370.3	28421	Coral	35,680
Porites compressa pH (Dinsdale et al., 2008b)	4440371.3	28423	Coral	50,364
Porites compressa nutrient (Dinsdale et al., 2008b)	4440377.3	28425	Coral	34,433
Pozas Azules II (Desnues et al., 2008)	4440320.3	28355	Microbialites	302,987
Rios Mesquites (Desnues et al., 2008)	4440321.3	28357	Microbialites	328,656
Highborne Cay (Desnues et al., 2008)	4440323.3	28381	Microbialites	328,656
Healthy fish slime (Dinsdale et al., 2008b)	4440065.3	28401	Fish	61,476
Morbid fish slime (Dinsdale et al., 2008b)	4440064.3	28403	Fish	60,111
Lung sputum Cystic Fibrosis patient (Dinsdale et al., 2008b)	4440441.3	28441	Other animals	98,223
Lung sputum healthy (Dinsdale et al., 2008b)	4440442.4	28439	Other animals	39,807
Mosquito Oceansdie,CA (Dinsdale et al., 2008b)	4440052.3	28413	Mosquito	340,098
Mosquito San Diego, CA (Dinsdale et al., 2008b)	4440053.3	28467	Mosquito	657,204
Mosquito II San Diego, CA (Dinsdale et al., 2008b)	4440054.3	28469	Mosquito	615,576

Supplementary Figure 1. Three-dimensional scatter plot of principal components from PCA of microbial and viral dinucleotide frequencies for original and additional metagenomes. The percent of variation each principal component explains is indicated in parentheses.



Supplementary Figure 2. Hierarchical clustering of viromes by the first three principal components from dinucleotide relative abundances. Cluster numbers are indicated above the branch for the cluster node, and metagenomes are labeled according to biome. Coral PC indicates *P. compressa* coral viromes, Marine LI indicates viromes from the Line Islands, and SS indicates solar salterns.



CHAPTER 3: VIRAL METAGENOMICS IN THE HEALTHY HOST

To most effectively use viral metagenomics to identify emerging pathogens and to delineate the role of viruses in disease, it is first necessary to establish baseline viral communities in healthy individuals. Many studies have focused on identifying specific viruses in disease outbreaks, but few have characterized the normal viral flora in the non-diseased. This chapter provides a case study of oropharyngeal viruses in the healthy adult population, and demonstrates the use of metagenomics as a rapid screening tool, and a vehicle for the discovery of novel and unexpected viruses. Additionally, the effects of sample preparation techniques on metagenomes are discussed.

Introduction

The human oropharynx is constantly exposed to a wide variety of viruses and microbes from the environment, from both inhaled air and ingested food and water. The oropharynx serves as a niche for commensal bacteria, some of which (e.g. *Streptococcus* and *Neisseria* spp.) can be pathogenic when introduced into other body sites (1-3). In healthy individuals, these normal flora prevent colonization by invading organisms by changing the local pH, producing bacteriocins, and providing a mechanical barrier which prevents adherence to mucosal surfaces (3-5). The oropharynx is also a reservoir for several viruses, including HIV, as well as papillomaviruses and Epstein-Barr virus, which are associated with oropharyngeal carcinomas (6-9). While oropharyngeal and oral microbes in general have been studied extensively using culturing and 16S sequencing, little is known about viral communities in the oropharyngeal spaces of healthy individuals (1; 10-13). The advent of viral metagenomics, i.e. culture-independent sequencing of viral nucleic acids, has made it possible to rapidly screen human samples for both known and novel viruses (14-17). For example, a number of known pathogenic viruses, as well as previously unknown types, were detected in nasopharyngeal aspirates from patients with respiratory infections (14; 16).

Here, we present the first description of oropharyngeal viral communities in healthy individuals. The initial purpose of this study was to evaluate the feasibility of viral screening using metagenomics in asymptomatic human subjects. Characterization of viral communities in healthy individuals is critical, as it establishes a baseline for comparison with samples from diseased individuals. However, in healthy individuals, viruses are likely to present in very small numbers, presenting a greater challenge for detection. We demonstrate that oropharyngeal swabs coupled with high-throughput sequencing are an effective method for sampling and characterizing oropharyngeal viral communities. Viral metagenomic sequences from a pool of 19 oropharyngeal samples provided complete coverage of several phage genomes, and identified the oropharynx as a potential reservoir for enterobacteria phage T3. Additionally, phage-encoded platelet-binding factors associated with *Streptococcus mitis* virulence in the endocardium were detected for the first time in the oral cavity, providing a potential link between viral communities in the oropharynx and heart disease.

Materials and Methods

Subject recruitment for the oropharyngeal metagenomes and saliva PCR assay were approved by the San Diego State University Institutional Review Board (SDSU IRB 2121) and Environmental Health Services (BUA 06-02-062R). Subject recruitment for the saliva metagenomic study was approved by the Stanford University Administrative Panel on Human Subjects in Medical Research.

Oropharyngeal sampling

Oropharyngeal swab samples were collected in July 2007 from individuals with no symptoms of respiratory infection as assessed by a pre-screening questionnaire. A signed consent form was obtained from each subject prior to sample collection, as required by the San Diego State University Institutional Review Board (IRB). Study subjects ranged in age from 23 to 56 years old, and consisted of 8 females and 11 males. Samples were obtained by swabbing the area posterior and superior to the palatopharyngeal arch (lower border of the nasopharynx) on both the right and left sides with a sterile swab. Swabs were replaced into their original self-sealing specimen containers and transported to the laboratory.

Oropharyngeal sample processing and viral isolation

Upon arrival at the laboratory, 2 ml of SM buffer (50 mM Tris·HCl, pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$, 0.01% gelatin) was added to the oropharyngeal swabs. Samples were then treated with dithiothreitol to break up mucus as described in (17), and subsequently filtered through 0.8 µm polycarbonate filter (Sterlitech Corp.) and 0.45 µm Millex-HV filters. Filtrates from 19 individual samples were combined to form a pooled sample. The filtrate was brought to a density of 1.15 g ml^{-1} by addition of solid cesium chloride (CsCl). This sample was overlaid onto a CsCl step gradient to concentrate viral particles as detailed in (18). Ten microliters of each viral concentrate was vacuum filtered onto a 0.02 micron Anodisc (Millipore), stained with SYBR Gold (Invitrogen), and virus-like particles (VLPs) were visualized using epifluorescence microscopy.

Half of the viral concentrate was filtered with a .22 µm Millex-HV filter to remove cellular material. DNase I was added to the sample to a final concentration of 10 µg ml^{-1}, followed by incubation at 37°C for one hour to degrade free DNA. The other half of the viral concentrate was treated with chloroform but not filtered. 20 µl of chlorofrom was added per ml of viral concentrate, and the sample was then incubated at 4°C for 2 hours, and centrifuged at 2000 rpm for 15 minutes. The supernatant containing the viral concentrate was collected and treated with DNase I as described above. DNA was extracted from the filtered and chloroformed samples using a CTAB/Formamide protocol (19). Viral DNA was sequenced at the Joint Genome Institute using the 454 GS-FLX platform. The chloroformed metagenome contained 215,281 sequences with an average length of 206 base pairs and the filtered metagenome contained 245,025 sequences with an average length of 219 base pairs.

Saliva sampling for metagenomic study

For the metagenomic study, three subjects donated saliva samples at three time points over a 3-month time period from February 2008 to April 2008. All subjects had no pre-existing medical conditions and were determined to be periodontally healthy based on full baseline periodontal examinations performed prior to the study. A minimum of 3 ml of saliva was collected at each time point, and saliva was stored at -20°C until processing for metagenomic sequencing.

Initial processing of metagenomic sequences

Viral metagenomic sequences were deposited in NCBI under Genome Project numbers 43627 (oropharyngeal) and 43629 (saliva metagenomic sequence subset). The two oropharyngeal and nine saliva metagenomic libraries were de-replicated and then compared to the non-redundant database at NCBI (http://www.ncbi.nlm.nih.gov) using BLASTn and tBLASTx (20). Sequences were assigned taxonomy based on the most significant BLAST similarity with an e-value less than 10^{-5} and a minimum alignment length of 50 bp. All sequences classified as microbial were compared to the ACLAME prophage database using BLASTn and tBLASTx to distinguish prophage sequences from microbial genomic sequences (21). Sequences with best BLAST similarities to eukaryotic and microbial genomes were removed from the metagenomes prior to subsequent analyses.

Bioinformatics for oropharyngeal metagenomes

Sequences with best BLASTn similarities to viral genomes were mapped to the genomes of EBV, *E. Coli* phage T3, *P. acnes* phage PA6 and *S. mitis* phage SM1 using BLAT and visualized with the Integrated Genome Browser (22; 23). Reference sequences for viral genomes were obtained from NCBI (http://www.ncbi.nlm.nih.gov). Contiuous coverage of the SM1 genome at the amino acid level and pblA and pblB genes was calculated using a Perl script. Contigs were assembled using the 454 gsAssembler with a minimum overlap length of 35 bp and 98% identity and compared to the non-redundant database for taxonomic assignment. Phage genome annotations were obtained from (24) for phage T3, (25) for phage PA6, and (26) for phage SM1. Counts of significant similarities to T3 in other environmental metagenomes were obtained by comparing the T3 genome to the environmental samples database (env_nt) at NCBI using BLASTn.

Viral community composition was determined using GAAS, based on tBLASTx comparisons of all non-eukaryotic and non-microbial sequences to a database containing all complete viral genomes currently available at NCBI (27). GAAS parameters were at least 40% identity with minimum relative alignment length of 80% and e-value cutoff of 10^{-5}. Viral diversity was estimated using the PHACCs program, with input contig spectra generated by Circonspect (http://biome.sdsu.edu/circonspect), and average genome size estimated by GAAS (27; 28). Cross-contig spectra were generated using Circonspect and used for a Monte Carlo simulation as described in (29) to determine the percent of species shared and permuted between the two oropharyngeal metagenomes.

Bioinformatics and statistics for salivary metagenomes

Sequences with significant similarity (e-value<10^{-5}, >30% identity over at least 80% of the query length) to the pblA and pblB gene regions of phage SM1 were extracted from the salivary metagenomes. Continuous coverage of the pblA and pblB genes was calculated with a Perl script. Sequences extracted from each metagenome were compared in a bi-directional pairwise fashion using cd-hit-2d-est with a 90%

identity cutoff to determine the percentage of sequences shared (30). The similarity index between metagenomes was calculated as the number of pblA or pblB sequences in metagenome 1 shared with metagenome 2 plus the number of sequences in metagenome 2 shared with metagenome 1 divided by the total number of sequences. The dissimilarity matrix was constructed by subtracting all similarity values from 1, and then used as an input to multi-dimensional scaling in R (31). Coverage of pblA and pblB genes in each metagenome was assessed by dividing the gene into bins of 20, 50, or 100 base pairs and counting the number of sequences covering each bin. Coverage over all bins was compared using the XIPE program, which uses non-parametric statistical methods to compare two empirical distributions (32). Coverage dissimilarity was calculated as the proportion of bins identified as having different coverage by XIPE. Results were the same whether 20, 50, or 100 base pair bins were used. The correlation between coverage and sequence similarity was calculated using the cor.test procedure in R with the 'spearman' option (31).

Bacterial strains

The *S. mitis* SF100 and PS344 strains were provided courtesy of Dr. Paul Sullam and Dr. Ho Seong Seo. *S. mitis* SF100 contains the complete SM1 prophage including the pblA and pblB genes (26). *S. mitis* PS344 contains the prophage with pblA and pblB deleted (26). *S. mitis* strains were grown on blood agar (TSA with 5% sheep's blood added) or Todd Hewitt broth (THB) at 5% CO_2 at 37ºC.

PCR screening of saliva samples

Total DNA was extracted from 20 individual saliva samples as described in (33). Positive control DNA was prepared from an overnight culture of *S. mitis* SF100 grown in THB. DNA was extracted using the Nucleospin Tissue Kit (Macherey-Nagel Inc., Bethlehem, PA) with the addition of a gram positive lysis step as described in the manufacturer's instructions. Negative control DNA was similarly prepared from *S. mitis* PS344.

The PCR reaction mixture (50 µL total) contained target DNA, 1X Taq Buffer, 0.2 mM dNTPs, 1 µM of each primer, and 1 U Taq DNA polymerase. The forward primer (pblA1456F) was 5'-ACCGCAGAGGCAGCGAATGC-3', and the reverse primer (pblA2222F) was 5'-CCAGGCCATAGACGCAGCCG-3'. Primers were designed using primer BLAST at NCBI (34). The thermocycler conditions were: 5 min at 94°C; 30 cycles of 1 min at 94°C, 1 min at 58°C with a -0.5°C touchdown, 1min at 72°C; and 10 min at 72°C. PCR products were checked for size on a 1% agarose gel, and prepared for sequencing using the Accu-Prep PCR Purification Kit (Bioneer Corporation, South Korea). PCR products were sequenced at the SDSU Microchemical Core Facility using an ABI Prism 3100 Genetic Analyzer. Sequences were deposited in Genbank under accession numbers GU586484, GU586485, GU586486, GU586487, GU5864848, GU586489, and GU586490.

Sequences from PCR products were trimmed and aligned to each other and the reference sequence (Streptococcus phage SM1 pblA, GeneID: 1009419) using ClustalW (35). Nucleotide sequences were translated in all six frames using TranSeq and were aligned to the translated reference sequence to determine the correct translation (36). Translated PCR product sequences were then aligned to the SM1 reference sequence both with and without the presence of pblA homolog sequences. Homolog sequences were as follows: *S. pyogenes* MGAS315 SpyM3_1104 (GeneID: 1009419), *S. pneumoniae* 70585 SP70585_0072 (GeneID: 7683049), *S. agalactiae* Lambda Sa1 pblA (GeneID: 1013400), *S. agalactiae* Lambda Sa03 pblA (GeneID: 3686919), *S. pyogenes* M1GAS Spy_1448 (GeneID: 901501), *E. faecalis* V583 tape measure (GeneID: 1200878), *S. pyogenes* M1GAS SpyM3_1313 (GeneID: 1009628), *E. faecalis* V583 tail protein (GeneID: 1199267), and *S. pyogenes* phage 315.5 tail protein (GeneID: 1257924) (37-40). Alignments were visualized using JalView (41). A phylogenetic tree was created based on the aligned amino acid sequences using MrBayes 3.1 (42). Four independent Monte Carlo Markov chains were run for 500,000 generations using the mixed amino acid model option.

Phage induction assays

Phage inductions were performed using an adaptation of the protocol described in (26). Overnight cultures of *S. mitis* SF100 were grown in THB for 16 hours. Cultures were diluted 1:10 in fresh THB and incubated for 30 minutes. Cultures were treated with one of six treatments: 0.25 μg/mL mitomycin C, red wine diluted 1:100, white wine 1:10, cola diluted 1:10, nicotine 1:10 (2.4 mg), or soy sauce diluted 1:10. The nictoine treatment consisted of Johnson Creek Original Smoke Juice (Vapure, Inc.), which contains 24 mg/mL of nicotine. An untreated culture was used as a control. These treatments were selected based on overnight growth curves of *S. mitis* SF100 with a variety of treatments added. All cultures were incubated for 3 hours, and then filtered using an 0.45 µm Millex-HV filter to remove remaining bacterial cells.

Phage particles were enumerated using a flow cytometry procedure from (43). Samples were fixed with 0.5% glutaraldehyde for 30 minutes at 4ºC. Samples were then flash frozen in liquid nitrogen and stored at -80ºC until analysis. For analysis, samples were thawed at room temperature and diluted 1:100 in TE buffer. One unit of DNase was added to each diluted sample and allowed to incubate at room temperature for 15 minutes to eliminate free DNA in the sample. Fresh SYBR Green I (0.5X) was then added to each sample to stain for DNA and incubated at 80ºC for 10 minutes in the dark. After incubation, samples were allowed to cool in the dark at room temperature for 5 minutes. Internal standard beads (0.75 mm diameter YG fluorescent latex microspheres; Polysciences, Inc., Warrington, PA) were added to each sample at a concentration of 1 x 10^6 beads per sample. Samples were analyzed using a FACSAria flow cytometer (Becton Dickinson, San Jose, CA) using FACSDiva software. The cytometer threshold was set on green fluorescence while the machine flow rate analyzed ~1000 events per second. Contour plots were generated on a side scatter x-axis and a green fluorescence y-axis on bi-exponential scales. Two separate electronic gates were generated for enumeration of virus and beads. Samples were collected until 1 x 10^5 bead events were detected. Viral positive events between different induction conditions had a minimum of ~1600 viral counts and maximum of ~420,000 viral counts per 1 x 10^5 bead events. Three replicate experiments for each treatment were conducted and statistical significance was assessed using randomization tests as implemented in the R function permtest (31).

Results and Discussion

Metagenomic detection of oropharyngeal viruses

Metagenomic sequencing of oropharyngeal swabs detected both phage and eukaryotic viruses (Figure 1). Taxonomy was assigned to metagenomic sequences based on BLAST comparisons to the non-redundant database (e-value<10^{-5}). BLASTn analysis identified fifty-three sequences which were nearly identical (>98% identity at the nucleotide level) to Epstein-Barr virus (EBV). The majority of these sequences aligned to open reading frames in the EBV genome, including genes involved in viral replication and latency as well as virion structure (Figure 3.1A). No additional sequences were recruited to the EBV genome using amino acid level searches (tBLASTx). EBV primarily infects epithelial cells in the oropharynx (44; 45). EBV infection is generally controlled by the immune system in healthy individuals, but the virus remains latent in circulating B lymphocytes (45; 46). Viral reactivation can occur in seropositive-normal individuals, resulting in viral shedding in the oropharynx (47). While it is estimated that 90% of the healthy adult population is seropositive for EBV, reactivation only occurs in 10-20% of individuals with latent EBV infections (44; 45). The incomplete coverage of EBV in the oropharyngeal metagenome was likely a reflection of the low prevalence of individuals actively shedding virus in the pooled sample population, as the metagenome was a composite from all nineteen study subjects. Detection of EBV in a pooled sample indicates that metagenomic sequencing of oropharyngeal swabs has adequate sensitivity to serve as a rapid non-invasive screen for viruses in individuals.

The complete genome of *E. coli* phage T3 was recovered from oropharyngeal

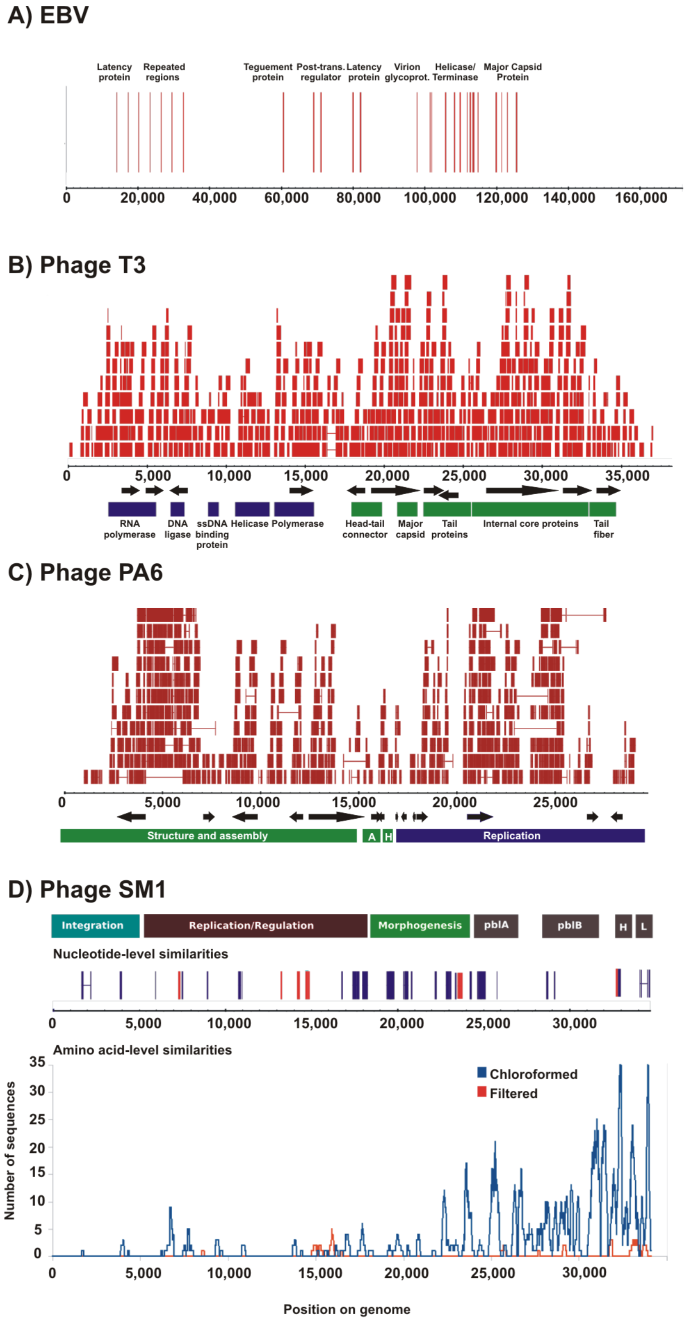


Figure 3.1: Coverage of viral genomes by oropharyngeal metagenomic sequences: Epstein-Barr virus (A), E. coli phage T3 (B), P. acnes phage P6 (C), and S. mitis phage SM1 (D). Similarities obtained from the chloroformed metagenome are in blue, and those from the filtered metagenome are shown in red. Nucleotide-level coverage (A,B,C, D top panel) was determined by alignment of metagenomic sequences to complete viral genome sequences using BLAT. Amino acid level coverage (D, bottom panel) was plotted using significant tBLASTx (e-value<10^{-5}) similarities to each genome. Contigs were assembled using the 454 gsAssembler and aligned to genomes using BLAT.

swabs (Figure 3.1B). Over 500 sequences were at least 98% identical to the T3 genome at the nucleotide level. These sequences provided approximately 3X coverage of T3, and could be combined into contigs as large as 4kb. Laboratory strains of phage T3 are widely used for experimental purposes, however, the origins of and natural reservoirs for T3 are largely unknown (48). A BLAST search of publicly available environmental metagenomes revealed a very low prevalence of sequences similar to T3, even in fecal samples which are considered to be a source of the phage (Table 2). Our results indicate that the oropharynx may be a previously undiscovered environmental reservoir for phage T3.

Table 3.1: Count of BLASTn similarities to E. coli phage T3 in environmental metagenomes. Similarities were determined by BLASTn (e-value<10-5).

Sample name	BLASTn identities (e-value<10^{-5})	Reference
Chloroformed oropharyngeal viruses	523	This study
Marine viruses	63	Angly et al. (2006)
Human gut microbiome	1	Kurokawa et al. (2007)
Coral viruses	1	Vega Thurber et al. (2009)
Freshwater viruses	35	Dinsdale et al. (2008)
Mosquito viruses	5	Dinsdale et al. (2008)

Metagenomic sequences provided high coverage of *Propionibacterium acnes* phage PA6 at the nucleotide level, and *Streptococcus mitis* phage SM1 at the amino acid level (Figure 3.1C and D). Contigs of up to 2 kb could be assembled and aligned to the PA6 genome, however, no contigs larger than 500 bp could be assembled which were significantly similar to SM1. Phage PA6 is a lytic phage whose host, *P. acnes*, is highly abundant in the oral cavity (25). Phage SM1 is a temperate phage previously isolated from *S. mitis* SF100, an endocarditis strain (26). SM1 carries two genes, pblA and pblB, which contribute to *S. mitis* virulence in the endocardium (26; 49; 50).

While *S. mitis* is a ubiquitous member of the normal oral flora, the presence of phage SM1 has never previously been reported in the mouth or oropharynx (1). The lack of long contigs and discontinuous coverage at the nucleotide level suggests that the SM1 nucleotide sequence was highly variable either between individuals, within individuals, or both. Temperate phage adopt the oligonucleotide usage patterns of their hosts, which can lead to sequence divergence at the nucleotide level if multiple, different hosts are present (51; 52). Since the oropharyngeal metagenomes were constructed from pooled samples from 19 individuals, it is likely that phage with varied hosts and host ranges were sampled.

Sample processing methods affect metagenomic composition

The composition of the oropharyngeal metagenomes differed depending on which sample preparation method was used (Figure 3.2A and B). Prior to DNA extraction, the pooled oropharyngeal swab sample was split, and each half was treated to reduce microbial contamination, either by the addition of chloroform or 0.22 micron filtration. The filtered metagenome contained a higher percentage of bacterial sequences, while the chloroformed metagenome was enriched in viral (including phage) sequences (Figure 3.2A). Filtering at 0.22 microns should trap bacterial cells while allowing viral particles to pass through (18). However, some viral particles will stick to the filter, especially larger viruses. EBV was the only eukaryotic virus detected in the oropharyngeal metagenomes, and it was present only in the chloroformed metagenome. EBV virions range in diameter from 120 to 220 nanometers and thus may not have passed through the filter (53). Additionally, some bacterial cells are likely to have escaped filtration. These cells would have been lysed during viral DNA extraction, releasing chromosomal DNA, and contaminating the viral metagenome (18). Chloroform treatment permeabilizes the membranes of bacterial cells, leading to cell death and the release of chromosomal DNA into the medium, where it can be digested with DNase I (18). This treatment also releases any intracellular viral particles, which may be fully assembled but have not yet induced host cell lysis. In general, viral capsids are resistant to chloroform, and remain intact until lysis during DNA extraction. The higher rate of viral recovery in the chloroformed metagenome suggests that chloroform treatment may be a more effective strategy for reducing microbial contamination and enriching for viruses in viral metagenomes.

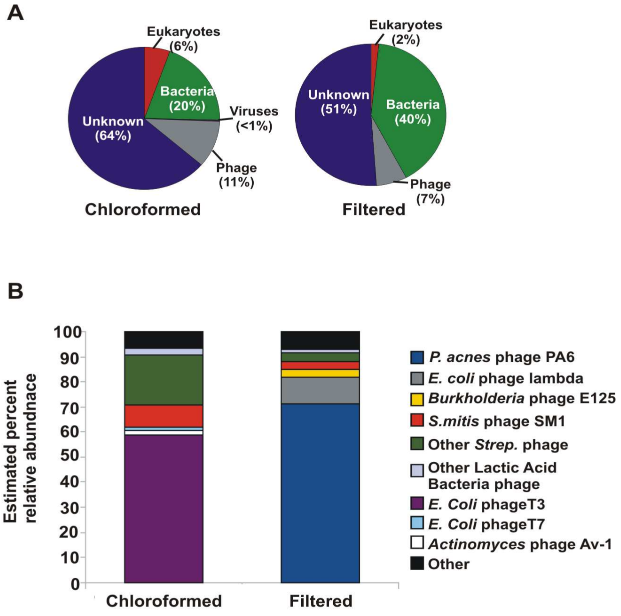


Figure 3.2: Taxonomic composition of the oropharyngeal metagenomes. (A) Composition of complete metagenomes, as determined by best tBLASTx similarities to the non-redundant database (e-value <10-5). (B) Composition of viral communities as determined by GAAS.

Phage communities in the two oropharyngeal metagenomes shared many species, but in different relative abundances (Figure 3.2B). Community composition was estimated using GAAS, which calculates relative abundances based on all significant BLAST similarities (27). *E. coli* phage T3 was the most abundant phage in the chloroformed sample, yet only comprised 1.6% of the community in the filtered sample. Similarly, *P. acnes* phage PA6 was the most abundant phage in the filtered sample, yet appeared in extremely low abundance (< 0.01%) in the chloroformed sample. *P. acnes* is ubiquitous in the healthy oral cavity, while gram negative bacteria such as *E. coli* are generally present in low abundance or not at all, as they are rapidly cleared in healthy people (4, 33, 34). Abedon demonstrated that phage with more abundant hosts tend to have shorter latent periods, i.e. a smaller lag time between adsorption and host lysis (54). Phage with less abundant hosts have longer latent periods, and will produce more progeny prior to lysis, generating a larger burst size (54). The addition of chloroform would cause the release of progeny phage from host cells, while filtration would remove host cells and their intracellular phage. The shift in phage T3 abundance between the filtered and chloroformed samples is likely the result of the release of intracellular T3 phage during chloroform treatment. This may also account for the increased abundance of T7 in the chloroformed sample (1.3% versus 0.1% in the filtered sample). The enrichment of *E. coli* phage lambda in the filtered metagenome (11.1% versus <0.01%) is seemingly in contrast to the long latent period hypothesis. However, lambda is a temperate phage, and several lambda sequences with flanking host sequences were detected in the metagenomes. This indicates that lambda was present as a prophage element integrated into the host genome, not as a free phage particle. Therefore, the higher abundance of lambda in the filtered metagenome was due to the higher level of bacterial DNA contamination.

Streptococcal phage were more abundant in the chloroformed sample than the filtered sample, comprising 33% and 7% of the viral communities respectively (Figure 3.2B). *S. mitis* phage SM1 was the most abundant Streptococcal phage in both metagenomes. Phage of other lactic acid bacteria (LAB) were also more abundant in the chloroformed metagenome (2.5% versus 1.0% in filtered). All of the LAB and Streptococcal phage detected were temperate phage. No flanking host sequences were detected adjacent to these phage sequences in the metagenomes, indicating the presence of free phage particles. Free phage would be enriched in the chloroformed sample due to the release of intracellular phage from host cells as described above. Streptococci are among the first bacteria to colonize the oral cavity, and remain in the mouth and oropharynx at high population densities throughout an individual's life (3; 5). Lactobacilli and other lactic acid bacteria are also common constituents of the normal oral flora, but are present at lower abundances, as reflected by the lower abundances of their phage in the oropharyngeal viral communities (3; 5).

Diversity of viruses in the oropharynx

Viral communities had the same predicted diversity, regardless of sample preparation method. Viral diversity was estimated using the PHACCs program as described in (28). The PHACCs method uses all metagenomic sequences, not just those with significant BLAST similarities (28). Viral communities in both the filtered and choloroformed samples were predicted to follow a power law distribution. The estimated richness of viral communities in both samples was 236 species, which was similar to estimates for the human respiratory tract, and low compared to the viral richness in marine environments (29; 17; 55). Estimates of microbial richness in the healthy oral cavity are similarly low; while over 700 microbial species have been identified, each individual is thought to only harbor 100-200 at any given time (1; 12). Despite the constant introduction of environmental microbes from food, water, and air, microbial and viral richness in the oropharynx is limited by several anatomical and biological mechanisms. Microbiota can be trapped in the mucosa prior to adherence, inhibited by chemicals in saliva such as lactoferrin, or cleared by the host immune system (3; 56). Additionally, normal flora prevent the adherence and growth of transient microbiota by producing bacteriocins and manipulating the pH of oral microenvironments (3; 56).

Viral communities in the filtered and chloroformed samples shared many genotypes, but at different relative abundances. Taxonomic data indicated that the majority of viruses which could be identified using BLAST appeared in both metagenomes, but in different proportions. To test whether this was true for all viral genotypes, not just those with significant BLAST similarities, cross-contigs were generated between the metagenomes, and a Monte Carlo simulation was conducted as described in (29). The simulation uses cross-contig spectra to estimate what proportion of genotypes are shared between communities, and what proportion of these shared genotypes are permuted, i.e. present in different abundances. The filtered and chloroformed viral communities were predicted to share more than 95% of genotypes with 30% permuted (Figure 3.3A). When each sample was compared to itself as a control, nearly all (>99%) of sequences were shared, however, less than 0.1% were permuted (Figure 3.3B and C). The simulation results corroborated the BLAST-based comparisons, demonstrating that while the filtered and chloroformed sample shared the same population of viruses, sample preparation methods altered their relative abundances.

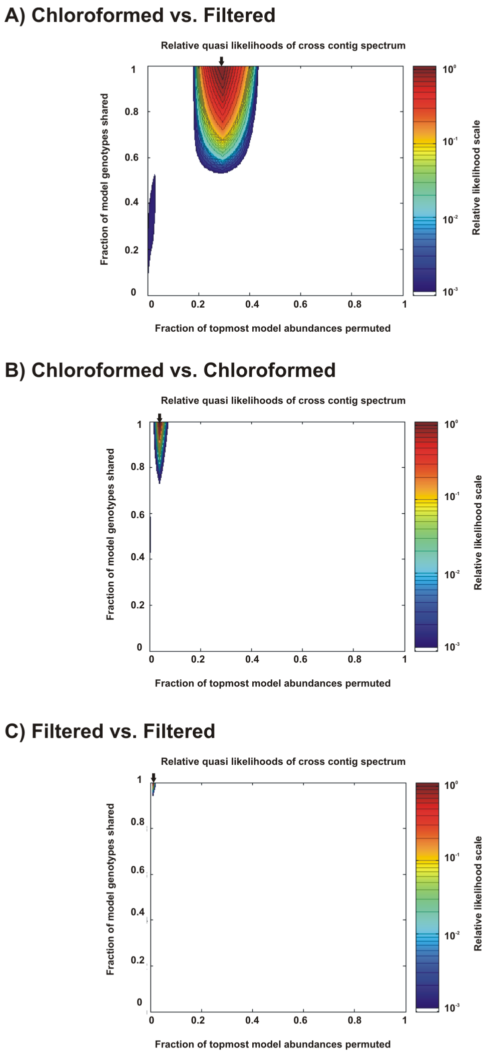


Figure 3.3: Monte Carlo analysis of cross-contig spectra for oropharyngeal metagenomes. The area of maximum likelihood is indicated by an arrow. (A) The filtered and chloroformed metagenomes were predicted to share more than 95% of genotypes with 30% of their relative abundances permuted. (B,C) Cross-contig spectra for each metagenome versus itself indicated that nearly all species were shared, while very few were permuted. It is expected that a metagenome versus itself would share 100% of sequences with 0% permuted, however, the small deviations seen here are the result of the sampling mechanism implemented in Circonspect.

Phage-encoded platelet binding factors in oropharyngeal and salivary metagenomes

Two genes of *S. mitis* phage SM1 that encode platelet-adhesion factors, pblA and pblB, were detected in the oropharyngeal metagenomes (Figure 3.1D; Figure 3.4).

While few sequences similar to pblA and pblB at the nucleotide level were identified using BLASTn, coverage of pblA, pblB, and holin and lysin genes was much higher than for the rest of the SM1 genome at the amino acid level (Figure 3.1D). PblA and pblB are integral phage tail proteins, and phage with pblA and pblB gene deletions have intact capsids but no tails (26; 50). PblA and pblB also mediate the attachment of *S. mitis* to platelets, which has been shown in a rabbit model to contribute significantly to the virulence of *S. mitis* in the endocardium (49; 50; 57). The interaction between *S. mitis* cells and platelets requires phage induction for maximal release of intracellular pblA and pblB, but the soluble proteins can bind to choline residues on the surface of host or non-host cells (58). Theoretically, pblA and pblB genes inserted into any phage capable of host cell permeabilization or lysis would be sufficient to mediate *S. mitis* adhesion to platelets. Siboo et al. suggested that pblA and pblB may be prime targets for horizontal gene transfer (HGT), as genes encoding phage tail proteins are especially labile regions, and can be highly variable even in phage with nearly identical genomes (26; 59; 60). Phage are major agents of HGT in many Streptococci, and have been shown to mediate inter-species genetic exchange, indicating that Streptococcal phage may have wide host ranges and undergo frequent recombination events (61-63). The high coverage of pblA and pblB genes in the oropharyngeal metagenomes suggests that in addition to phage SM1, there may be a population of recombinant phage which have acquired these genes through HGT. *S. mitis* and other Streptococci have been shown to enter the blood stream from the oral cavity following tooth extractions, so the dissemination of pblA and pblB genes in oral phage and microbes could potentially translate into an increased risk of endocarditis (2).

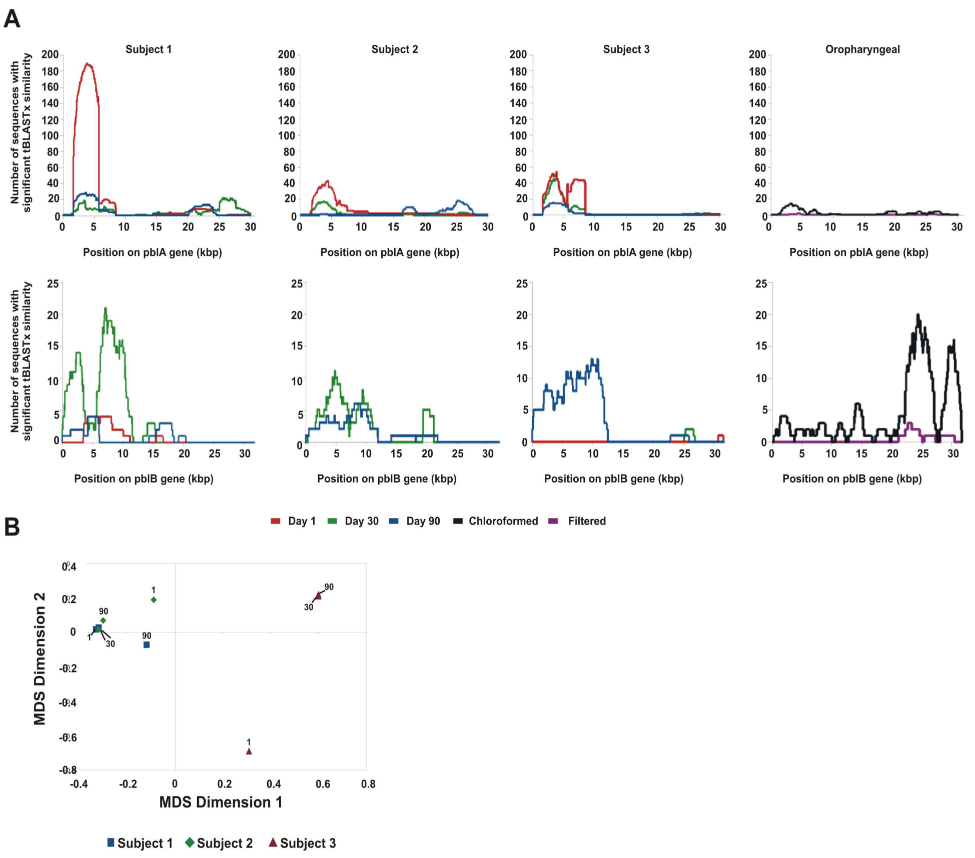


Figure 3.4: (A) Coverage of the pblA and pblB genes in salivary and oropharyngeal metagenomes. (B) Scatterplot of multi-dimensional scaling (MDS) coordinates for sequence dissimilarities between pblA and pblB sequences in salivary metagenomes. Points further apart represent less similar sequence sets. The time point for each saliva sample (1 day, 30 days, or 90 days) is indicated adjacent to each point.

Phage SM1-like pblA and pblB genes were also detected in salivary metagenomes from three individuals at three time points (Figure 3.4A). Subsequent to our analysis of the oropharyngeal metagenomes, nine pre-existing salivary metagenomes became available to us for screening for pblA and pblB genes. Sequences with significant tBLASTx similarities (e-value< 10^{-5}, identity >30%, query coverage >80%) to pblA were found in all individuals at all time points, while pblB was absent in Subject 2 at the 30 day time point. Aas et al. demonstrated that while some microbes preferentially colonize particular sites, *S. mitis* is ubiquitous, and can be detected throughout the oral cavity (1). The presence of phage SM1 pblA and pblB genes in both oropharyngeal and salivary metagenomes suggests that these genes, and most likely phage SM1 itself, have a similarly widespread distribution.

PblA sequences in the salivary metagenomes varied both between and within subjects (Figure 3.4B). The sequence comparison tool cd-hit-est-2d was used to assess the degree of variability of pblA sequences at the nucleotide level (30). PblB sequences were not analyzed as at some time points fewer than five sequences were identified. PblA sequences with 90% identity at the nucleotide level were considered to be congruent. A dissimilarity matrix was constructed from cd-hit-est-2d results and used as an input to multi-dimensional scaling (MDS) (Table 3.2). A scatterplot of MDS coordinates showed that pblA sequences differed between individuals and within individuals at different time points (Figure 3.4B). In all three individuals, sequences from closer time points appeared to be more similar than those from more distant time points (i.e. 1 vs 90 days). Sequences from subject 3 were extremely dissimilar to those from subjects 1 and 2 at all time points. To determine whether this divergence was driven by coverage differences, coverage of pblA genes was compared between metagenomes (Table 3.3). Coverage was not significantly correlated with sequence dissimilarities (Spearman's rho=0.26, p=0.13). Similar to the results in the oropharyngeal metagenomes, this suggests that pblA genes are variable between individuals, and in addition, within individuals over time. These nucleotide level changes may be indicative of adaptation of phage sequences to host oligonucleotide usage, reflecting the movement of phage SM1 genes into different and potentially novel hosts, either through host range expansion or lateral gene transfer (51; 52).

Table 3.2: Dissimilarity matrix for pblA sequences between salivary metagenomes. Dissimilarity was calculated at the proportion of unshared sequences between metagenomes as determined by cd-hit-2d-est.

	S1D1	S1D30	S1D90	S2D1	S2D30	S2D90	S3D1	S3D30	S3D90
S1D1	0	0.14	0.82	0.62	0.51	0.44	1	1	1
S1D30	0.14	0	0.72	0.46	0.55	0.58	1	1	1
S1D90	0.82	0.72	0	0.81	0.61	0.58	0.99	0.95	0.96
S2D1	0.62	0.46	0.81	0	0.50	0.29	0.98	0.80	0.77
S2D30	0.51	0.55	0.61	0.50	0	0.36	1	1	1
S2D90	0.44	0.58	0.58	0.29	0.36	0	1	0.96	0.98
S3D1	1	1	0.99	0.98	1	1	0	0.99	0.98
S3D30	1	1	0.95	0.80	1	0.96	0.99	0	0.34
S3D90	1	1	0.96	0.77	1	0.98	0.98	0.34	0

Table 3.3: Dissimilarity matrix for coverage of pblA genes between salivary metagenomes. Dissimilarity was calculated by dividing the pblA gene into 20 base pair bins and using the XIPE program to determine which bins were over-represented in each metagenome upon pairwise comparison with every other metagenome. The dissimilarity index is the proportion of non-identically distributed bins.

	S1D1	S1D30	S1D90	S2D1	S2D30	S2D90	S3D1	S3D30	S3D90
S1D1	0	0.07	0.13	0.10	0.23	0.10	0.30	0.10	0.13
S1D30	0.07	0	0.23	0	0.13	0	0.37	0	0.10
S1D90	0.13	0.23	0	0.27	0.20	0.20	0.10	0.20	0.13
S2D1	0.10	0	0.27	0	0.13	0	0.43	0	0
S2D30	0.23	0.13	0.20	0.13	0	0.13	0.50	0.03	0.10
S3D1	0.30	0.10	0.13	0	0.10	0	0.13	0	0

PCR detection of pblA in saliva samples

PblA gene fragments with high homology to phage SM1 pblA were detected in individual saliva samples from healthy individuals (Figures 3.5 and 3.6). Saliva samples were collected from 20 individuals and screened for the presence of an approximately 750 base pair region of pblA. This region, spanning nucleotides 1456 to 2222 of the pblA gene, was noticeably under-represented in both the oropharyngeal and salivary metagenomes (Figure 3.4). The pblA gene fragment was detected in 6 of the 20 individuals tested and was sequenced, along with positive control DNA from cultured *S. mitis* SF100. At the amino acid level, the positive control sequence was

7.1% divergent from the reference, similar to the saliva sequences, which ranged from 4.5% and 37.5% divergence (Figure 3.5). Phylogenetic analysis indicated that all saliva sequences and the positive control sequence were more closely related to the SM1 reference sequence than to any other pblA homolog (Figure 3.6). These results confirm the presence of SM1-like pblA genes in the healthy human oral cavity.

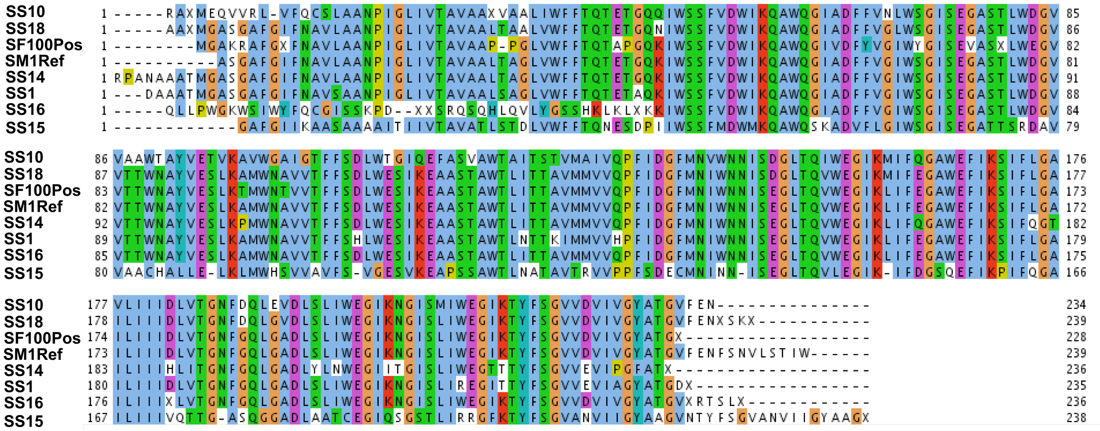

SS10 1 -----RAXMEQVVRL-VFQCSLAANPIGLIVTAVAAXVAALIWFFTQTETGQQIWSSFVDWIKQAWQGIADFFVNLWSGISEGASTLWDGV 85
SS18 1 -----AAXMGASGAFGIFNAVLAANPIGLIVTAVAALTAALVWFFTQTETGQNIWSSFVDWIKQAWQGIADFFVGLWSGISEGASTLWDGV 86
SF100Pos 1 --------MGAKRAFGXFNAVLAANPIGLIVTAVAAP-PGLVWFFTQTAPGQKIWSSFVDWIKQAWQGIVDFYVGIWYGISEVASXLWEGV 82
SM1Ref 1 ----------ASGAFGIFNAVLAANPIGLIVTAVAALTAGLVWFFTQTETGQKIWSSFVDWIKQAWQGIADFFVGIWSGISEGASTLWDGV 81
SS14 1 RPANAAATMGASGAFGIFNAVLAANPIGLIVTAVAALTAGLVWFFTQTETGQKIWSSFVDWIKQAWQGIADFFVGIWSGISEGASTLWDGV 91
SS1 1 ---DAAATMGASGAFGIFNAVSAANPIGLIVTAVAALSAGLVWFFTQTETAQKIWSSFVDWIKQAWQGIADFFVGIWSGISEGASTLWDGV 88
SS16 1 -----QLLPWGKWSIWYFQCGISSKPD--XXSRQSQHLQVLYGSSHKLKLXKKIWSSFVDWIKQAWQGIADFFVGIWSGISEGASTLWDGV 84
SS15 1 ------------GAFGIIKAASAAAAITIIVTAVATLSTDLVWFFTQNESDPIIWSSFMDWMKQAWQSKADVFLGIWSGISEGATTSRDAV 79
SS10 86 VAAWTAYVETVKAVWGAIGTFFSDLWTGIQEFASVAWTAITSTVMAIVQPFIDGFMNVWNNISDGLTQIWEGIKMIFQGAWEFIKSIFLGA 176
SS18 87 VTTWNAYVESLKAMWNAVVTFFSDLWESIKEAASTAWTLITTAVMMVVQPFIDGFMNIWNNISDGLTQVWEGIKMIFEGAWEFIKSIFLGA 177
SF100Pos 83 VTTWNAYVESLKTMWNTVVTFFSDLWESIKEAASTAWTLITTAVMMVVQPFIDGFMNIWNNISEGLTQVWEGIKLIFEGAWEFIKSIFLGA 173
SM1Ref 82 VTTWNAYVESLKAMWNAVVTFFSDLWESIKEAASTAWTLITTAVMMVVQPFIDGFMNIWNNISEGLTQVWEGIKLIFEGAWEFIKSIFLGA 172
SS14 92 VTTWNAYVESLKPMWNAVVTFFSDLWESIKEAASTAWTLITTAVMMVVQPFIDGFMNIWNNISEGLTQVWEGIKLIFQGAWEFIKSIFQGT 182
SS1 89 VTTWNAYVESLKAMWNAVVTFFSHLWESIKEAASTAWTLNTTKIMMVVHPFIDGFMNIWNNISEGLTQVWEGIKLIFEGAWEFIKSIFLGA 179
SS16 85 VTTWNAYVESLKAMWNAVVTFFSDLWESIKEAASTAWTLITTAVMMVVQPFIDGFMNIWNNISEGLTQVWEGIKLIFEGAWEFIKSIFLGA 175
SS15 80 VAACHALLE-LKLMWHSVVAVFS-VGESVKEAPSSAWTLNATAVTRVVPPFSDECMNINN-ISEGLTQVLEGIK-IFDGSQEFIKPIFQGA 166
SS10 177 VLIIIDLVTGNFDQLEVDLSLIWEGIKNGISMIWEGIKTYFSGVVDVIVGYATGVFEN--------------- 234
SS18 178 ILIIIDLVTGNFDQLGVDLSLIWEGIKNGISLIWEGIKTYFSGVVDVIVGYATGVFENXSKX----------- 239
SF100Pos 174 ILIIIDLVTGNFGQLGADLSLIWEGIKNGISLIWEGIKTYFSGVVDVIVGYATGX------------------ 228
SM1Ref 173 ILIIIDLVTGNFGQLGADLSLIWEGIKNGISLIWEGIKTYFSGVVDVIVGYATGVFENFSNVLSTIW------ 239
SS14 183 ILIIIHLITGNFGQLGADLYLNWEGIITGISLIWEGTTTYFSGVVEVIPGFATX------------------- 236
SS1 180 ILIIIDLVTGNFGQLGADLSLIWEGIKNGISLIREGITTYFSGVVEVIAGYATGDX----------------- 235
SS16 176 ILIIIXLVTGNFGQLGADLSLIWEGIKNGISLIWEGIKTYFSGVVDVIVGYATGVXRTSLX------------ 236
SS15 167 ILIIIVQTTG-ASQGGADLAATCEGIQSGSTLIRRGFKTYFSGVANVIIGYAAGVNTYFSGVANVIIGYAAGX 238

Figure 3.5: Amino acid alignment of saliva pblA sequences with the S. mitis SF100 positive control sequence and the SM1 reference sequence. The saliva and positive control sequences were translated in all 6 frames and aligned to determine the correct reading frame prior to alignment.

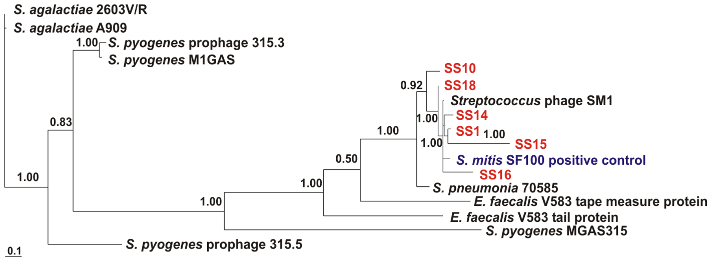


Figure 3.6: Phylogenetic relationships between pblA sequences from saliva samples and reference genomes. The Bayes values show the proportion of sampled trees in which the sequences to the right of the branch point clustered together. Saliva PCR sequences are shown in red and the positive control sequence from *S. mitis* SF100 is in blue.

Induction of phage SM1

Phage SM1 was induced by commonly ingested substances, such as nicotine and soy sauce (Figures 3.7 and 3.8). To determine the relative amounts of phage induced, we utilized a flow cytometry method to enumerate phage in each sample (43). Cultures of *S. mitis* SF100 were treated with red wine, white wine, soda, solubilized nicotine, soy sauce, or mitomycin C for phage induction and compared to an untreated control culture. Nicotine and soy sauce treatments produced significantly more phage particles ($p<0.05$) than the non-induced control, while red wine, white wine, and soda had no significant effect on phage production. Phage induction provides a vehicle for virulence genes to travel between bacterial species. Acquisition of toxin genes by Group A and Group C Streptococcus has been shown to occur by lysogenization following prophage induction, and it is likely that pblA and pblB genes could disseminate in the same manner (63). Mitchell et al. demonstrated that phage SM1 induction even at low levels is sufficient to facilitate *S. mitis* binding to platelets (57). Induction of phage by food or beverages in individuals with severe periodontal disease could lead to an increased endocarditis risk, as they are highly prone to acquiring bacteremia from routine activities such as toothbrushing (2; 64).

Additional considerations

The oropharyngeal viral community described here consisted almost exclusively of phage. With the exception of EBV, no eukaryotic viruses were detected. In healthy individuals, the absence of eukaryotic viruses may be characteristic of the non-diseased state. However, it is also possible that enveloped viruses were not efficiently isolated by the CsCl density gradient method due to their anomalous density (18).

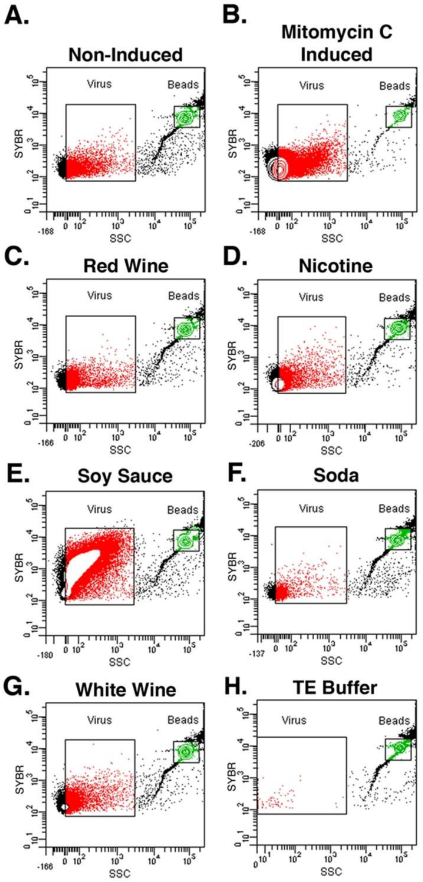


Figure 3.7: Representative flow cytometry data of viral induction. Samples were not induced (panel A) or induced under different conditions (panels B-G). Contour plots are displayed on bi-exponential scales with Side Scatter (SSC) on the x-axis and SYBR Green I (SYBR) on the y-axis. Panel H is beads alone in TE buffer to determine background signals and generate electronic gate for virus counts.

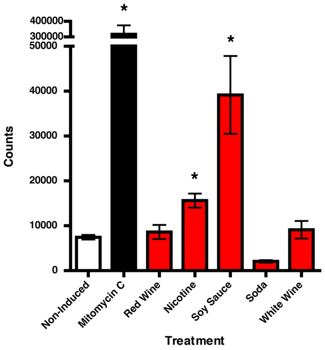


Figure 3.8: Phage induction assay. The mean number of phage events counted using flow cytometry (±SEM) is on the y-axis (n=3). White wine, nicotine, soda, and soy sauce were diluted 1:10, while red wine was diluted 1:100. Asterisks indicate a statistically significant increase in phage count ($p<0.05$).

A caveat to this study was the use of Multiple Displacement Amplification (MDA) with phi29 polymerase prior to 454 sequencing. While MDA generally does not bias the representation of individual genomes in metagenomic samples, small circular and long linear genomes may be disproportionately amplified (65; 66). In any case, all metagenomes were amplified using the same reaction conditions, allowing for valid comparisons between samples even if bias were introduced.

Conclusions

Metagenomics is a powerful tool for both characterization of environmental viral communities and discovery. In this study, we set out to evaluate the use of oropharyngeal swabs for viral detection, and unexpectedly discovered phage-encoded virulence genes in oropharyngeal viral communities. Detection of the pblA and pblB genes in saliva as well as oropharyngeal samples suggests that they are widely disseminated both in the oral cavity, and in the human population at large. Within the metagenomes, pblA and pblB sequences varied significantly at the nucleotide level between individuals and within individuals, suggesting HGT between phage SM1 and other resident phage genomes. HGT and the host range expansion of phage SM1 may be facilitated in the oral cavity by commonly ingested substances, as shown by our phage induction assay. Several studies have established a link between endocarditis and oral hygiene, demonstrating that Streptococci readily enter the bloodstream during tooth extractions, and following toothbrushing in individuals with periodontal disease (2; 67; 64). Future studies will be needed to determine the endocarditis risk associated with the presence of phage SM1 and/or pblA and pblB genes in the oral cavity.

References

1. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the Normal

Bacterial Flora of the Oral Cavity. J. Clin. Microbiol. 2005 Nov;43(11):5721-5732.

2. Bahrani-Mougeot FK, Paster BJ, Coleman S, Ashar J, Barbuto S, Lockhart PB. Diverse and Novel Oral Bacterial Species in Blood following Dental Procedures. J. Clin. Microbiol. 2008 Jun;46(6):2129-2132.

3. Hull MW, Chow AW. Indigenous microflora and innate immunity of the head and neck. Infect. Dis. Clin. North Am 2007 Jun;21(2):265-282, v.

4. Hentges DJ. The Anaerobic Microflora of the Human Body. Clinical Infectious Diseases 1993 Jun;16:S175-S180.

5. Jenkinson HF, Lamont RJ. Oral microbial communities in sickness and in health. Trends Microbiol 2005 Dec;13(12):589-595.

6. Moutsopoulos N, Greenwell-Wild T, Wahl S. Differential Mucosal Susceptibility in HIV-1 Transmission and Infection. Advances in Dental Research 2006 Apr;19(1):52-56.

7. Moutsopoulos NM, Nares S, Nikitakis N, Rangel Z, Wen J, Munson P, Sauk J, Wahl SM. Tonsil Epithelial Factors May Influence Oropharyngeal Human Immunodeficiency Virus Transmission. Am J Pathol 2007 Aug;171(2):571-579.

8. Shillitoe EJ. The role of viruses in squamous cell carcinoma of the oropharyngeal mucosa. Oral Oncology Apr;45(4-5):351-355.

9. Szkaradkiewicz A, Kruk-Zagajewska A, Wal M, Jopek A, Wierzbicka M, Kuch A. Epstein-Barr virus and human papillomavirus infections and oropharyngeal squamous cell carcinomas. Clinical and Experimental Medicine 2002 Nov;2(3):137-141.

10. Diaz PI, Chalmers NI, Rickard AH, Kong C, Milburn CL, Palmer RJ, Kolenbrander PE. Molecular Characterization of Subject-Specific Oral Microflora during Initial Colonization of Enamel. Appl. Environ. Microbiol. 2006 Apr;72(4):2837-2848.

11. Nasidze I, Li J, Quinque D, Tang K, Stoneking M. Global diversity in the human salivary microbiome. Genome Research 2009;19(4):636-643.

12. Paster BJ, Olsen I, Aas JA, Dewhirst FE. The breadth of bacterial diversity in the human periodontal pocket and other oral sites. Periodontology 2000 2006;42(1):80-87.

13. Kazor CE, Mitchell PM, Lee AM, Stokes LN, Loesche WJ, Dewhirst FE, Paster

BJ. Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. J. Clin. Microbiol 2003 Feb;41(2):558-563.

14. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proceedings of the National Academy of Sciences of the United States of America 2005;102(36):12891-12896.

15. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 2005 Nov;39(5):729-736.

16. Nakamura S, Yang C, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, Maeda N, Kawai J, Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T. Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. PLoS ONE 2009 Jan;4(1):e4219.

17. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS ONE 2009 Oct;4(10):e7370.

18. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc 2009;4(4):470-483.

19. Sambrook J. Molecular Cloning: A Laboratory Manual, Third Edition. 3rd ed. Cold Spring Harbor Laboratory Press; 2001.

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990 Oct;215(3):403-410.

21. Leplae R, Hebrant A, Wodak SJ, Toussaint A. ACLAME: A CLAssification of Mobile genetic Elements. Nucleic Acids Res 2004 Jan;32(Database issue):D45-D49.

22. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Research 2002 Apr;12(4):656-664.

23. Nicol JW, Helt GA, Blanchard SG, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 2009 Oct;25(20):2730-2731.

24. Pajunen MI, Elizondo MR, Skurnik M, Kieleczawa J, Molineux IJ. Complete

Nucleotide Sequence and Likely Recombinatorial Origin of Bacteriophage T3. Journal of Molecular Biology 2002 Jun;319(5):1115-1132.

25. Farrar MD, Howson KM, Bojar RA, West D, Towler JC, Parry J, Pelton K, Holland KT. Genome Sequence and Analysis of a Propionibacterium acnes Bacteriophage. J Bacteriol 2007 Jun;189(11):4161-4167.

26. Siboo IR, Bensing BA, Sullam PM. Genomic organization and molecular characterization of SM1, a temperate bacteriophage of Streptococcus mitis. J. Bacteriol 2003 Dec;185(23):6968-6975.

27. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol 2009 Dec;5(12):e1000593.

28. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 2005;6:41.

29. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The Marine Viromes of Four Oceanic Regions. PLoS Biol 2006 Nov;4(11):e368.

30. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006 Jul;22(13):1658-1659.

31. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: 2008.

32. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative metagenomics. BMC Bioinformatics 2006;7:162.

33. Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. Evaluation of saliva as a source of human DNA for population and association studies. Anal. Biochem 2006 Jun;353(2):272-277.

34. NCBI. Primer-BLAST [Internet]. 2008.Available from: http://www.ncbi.nlm.nih.gov/tools/primer-blast

35. MA L, G B, NP B, R C, PA M, H M, F V, IM W, A W, R L, JD T, TJ G, DG H. ClustalW and ClustalX version 2.0. Bioinformatics 2007 Sep;:btm404.

36. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000 Jun;16(6):276-277.

37. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, Lai HS, Lin SP, Qian Y, Jia HG, Najar FZ, Ren Q, Zhu H, Song L, White J, Yuan X, Clifton SW, Roe BA, McLaughlin R. Complete genome sequence of an M1 strain of Streptococcus pyogenes. Proceedings of the National Academy of Sciences of the United States of America 2001 Apr;98(8):4658-4663.

38. Paulsen IT, Banerjei L, Myers GSA, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF, Tettelin H, Dodson RJ, Umayam L, Brinkac L, Beanan M, Daugherty S, DeBoy RT, Durkin S, Kolonay J, Madupu R, Nelson W, Vamathevan J, Tran B, Upton J, Hansen T, Shetty J, Khouri H, Utterback T, Radune D, Ketchum KA, Dougherty BA, Fraser CM. Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. Science 2003 Mar;299(5615):2071-2074.

39. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proceedings of the National Academy of Sciences of the United States of America 2002;99(19):12391-12396.

40. Ramirez M, Severina E, Tomasz A. A High Incidence of Prophage Carriage among Natural Isolates of Streptococcus pneumoniae. J. Bacteriol. 1999 Jun;181(12):3618-3625.

41. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009 May;25(9):1189-1191.

42. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001 Aug;17(8):754-755.

43. Brussaard CPD. Enumeration of bacteriophages using flow cytometry. Methods Mol. Biol 2009;501:97-111.

44. Pender MP. Infection of autoreactive B lymphocytes with EBV, causing chronic autoimmune diseases. Trends in Immunology 2003 Nov;24(11):584-588.

45. Vetsika E, Callan M. Infectious mononucleosis and Epstein-Barr virus. Expert Rev Mol Med 2004 Nov;6(23):1-16.

46. Fields BN, Knipe DM, Howley PM, Chanock RM, Monath TP, Melnick JL, Roizman B, Straus SE. Fields Virology. 3rd ed. Lippincott Williams & Wilkins; 1996.

47. Ling PD, Lednicky JA, Keitel WA, Poston DG, White ZS, Peng R, Liu Z, Mehta SK, Pierson DL, Rooney CM, Vilchez RA, Smith EO, Butel JS. The dynamics of herpesvirus and polyomavirus reactivation and shedding in healthy adults: a 14-month longitudinal study. J. Infect. Dis 2003 May;187(10):1571-1580.

48. Abedon ST. The Murky Origin of Snow White and Her T-Even Dwarfs. Genetics 2000 Jun;155(2):481-486.

49. Bensing BA, Rubens CE, Sullam PM. Genetic loci of Streptococcus mitis that mediate binding to human platelets. Infect. Immun 2001 Mar;69(3):1373-1380.

50. Bensing BA, Siboo IR, Sullam PM. Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. Infect. Immun 2001 Oct;69(10):6186-6192.

51. Blaisdell BE, Campbell AM, Karlin S. Similarities and dissimilarities of phage genomes. Proc. Natl. Acad. Sci. U.S.A 1996 Jun;93(12):5854-5859.

52. Pride D, Wassenaar T, Ghose C, Blaser M. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 2006;7(1):8.

53. Buchen-Osmond C. ICTVdB-The Universal Virus Database. Columbia University, New York, USA: 2006.

54. Abedon ST. Selection for Bacteriophage Latent Period Length by Bacterial Density: A Theoretical Examination. Microbial Ecology 1989 Oct;18(2):79-88.

55. Marhaver KL, Edwards RA, Rohwer F. Viral communities associated with healthy and bleaching corals. Environ Microbiol 2008 Sep;10(9):2277-2286.

56. Mobbs KJ, van Saene HK, Sunderland D, Davies PD. Oropharyngeal Gram-negative Bacillary Carriage*. Chest 1999 Jun;115(6):1570-1575.

57. Mitchell J, Siboo IR, Takamatsu D, Chambers HF, Sullam PM. Mechanism of cell surface expression of the Streptococcus mitis platelet binding proteins PblA and PblB. Mol. Microbiol 2007 May;64(3):844-857.

58. Mitchell J, Sullam PM. Streptococcus mitis phage-encoded adhesins mediate attachment to {alpha}2-8-linked sialic acid residues on platelet membrane gangliosides. Infect. Immun 2009 Aug;77(8):3485-3490.

59. Angly F, Youle M, Nosrat B, Srinagesh S, Rodriguez-Brito B, McNairnie P, Deyanat-Yazdi G, Breitbart M, Rohwer F. Genomic analysis of multiple Roseophage SIO1 strains. Environmental Microbiology 2009;11(11):2863-2873.

60. Lucchini S, Desiere F, Brüssow H. Comparative Genomics of Streptococcus thermophilus Phage Species Supports a Modular Evolution Theory. J Virol 1999 Oct;73(10):8647-8656.

61. Davies MR, Tran TN, McMillan DJ, Gardiner DL, Currie BJ, Sriprakash KS. Inter-species genetic movement may blur the epidemiology of streptococcal diseases in endemic regions. Microbes Infect 2005 Jul;7(9-10):1128-1138.

62. Holden MTG, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, Jourdan T, Bason NC, Holroyd NE, Mungall K, Quail MA, Sanders M, Simmonds M, Willey D, Brooks K, Aanensen DM, Spratt BG, Jolley KA, Maiden MCJ, Kehoe M, Chanter N, Bentley SD, Robinson C, Maskell DJ, Parkhill J, Waller AS. Genomic Evidence for the Evolution of Streptococcus equi: Host Restriction, Increased Virulence, and Genetic Exchange with Human Pathogens. PLoS Pathog 2009 Mar;5(3):e1000346.

63. Vojtek I, Pirzada ZA, Henriques-Normark B, Mastny M, Janapatla RP, Charpentier E. Lysogenic Transfer of Group A Streptococcus Superantigen Gene among Streptococci. The Journal of Infectious Diseases 2008 Jan;197(2):225-234.

64. Lockhart PB, Brennan MT, Sasser HC, Fox PC, Paster BJ, Bahrani-Mougeot FK. Bacteremia Associated With Toothbrushing and Dental Extraction. Circulation 2008;117(24):3118-3125.

65. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 2002 Apr;99(8):5261-5266.

66. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC

Genomics [date unknown];7:216-216.

67. Lockhart PB, Brennan MT, Thornhill M, Michalowicz BS, Noll J, Bahrani-Mougeot FK, Sasser HC. Poor oral hygiene as a risk factor for infective endocarditis-related bacteremia. J Am Dent Assoc 2009 Oct;140(10):1238-1244.

Acknowledgements

We thank the Joint Genome Institute for sequencing the oropharyngeal metagenomes, as well as Paul Sullam and Ho Seong Seo for generously providing *S. mitis* strains SF100 and PS344, and Kelly Doran, Tom Gibson, Marlene DeMers, and Mike Gurney for their microbiology assistance. We would also like to acknowledge Gary C. Armitage and Peter Loomer for their assistance in the saliva metagenomic project. We are also grateful to Linda Wegley for helpful discussions. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344.

Chapter 3 has been accepted for publication in PNAS as: Willner, D, Furlan, M, Schmieder, R, Grasis, J, Pride, DT, Relman, D, Angly, F, McDole, T, Mariella, RP, Rohwer, F, and Haynes, M. Metagenomic detection of phage-encoded platelet binding factors in the human oropharynx. The dissertation author was the primary investigator and author of this material.

CHAPTER 4: COMPARATIVE METAGENOMICS IN HUMAN DISEASE

The human respiratory tract is constantly exposed to a wide variety of viruses, microbes and inorganic particulates from environmental air, water and food. Physical characteristics of inhaled particles and airway mucosal immunity determine which viruses and microbes will persist in the airways. Here we present the first metagenomic study of DNA viral communities in the airways of diseased and non-diseased individuals. We obtained sequences from sputum DNA viral communities in 5 individuals with cystic fibrosis (CF) and 5 individuals without the disease. Overall, diversity of viruses in the airways was low, with an average richness of 175 distinct viral genotypes. The majority of viral diversity was uncharacterized. CF phage communities were highly similar to each other, whereas Non-CF individuals had more distinct phage communities, which may reflect organisms in inhaled air. CF eukaryotic viral communities were dominated by a few viruses, including human herpesviruses and retroviruses. Functional metagenomics showed that all Non-CF viromes were similar, and that CF viromes were enriched in aromatic amino acid metabolism. The CF metagenomes occupied two different metabolic states, probably reflecting different disease states. There was one outlying CF virome which was characterized by an over-representation of Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase, an enzyme involved in the bacterial stringent response. Unique environments like the CF airway can drive functional adaptations, leading to shifts in metabolic profiles. These results have important clinical implications for CF, indicating that therapeutic measures may be more effective if used to change the respiratory environment, as opposed to shifting the taxonomic composition of resident microbiota.

Introduction

Each day the human respiratory tract comes into contact with billions of airborne particles, including viruses, microbes and allergens (1). Particle size and the local airway host immune response determine which inhaled viruses and particles will adhere to epithelial surfaces and persist in the airways (1, 2). The lungs and lower respiratory tract have generally been considered sterile in the absence of respiratory disease although very little is known about the microbiota of the upper and lower airways of non-diseased individuals. Microbes and viruses, including phage, have been implicated in chronic pulmonary diseases, such as chronic obstructive pulmonary disease (COPD), asthma, and cystic fibrosis (CF) (3-8). However, most of this work has been performed using standard microbial cultures and PCR-based studies, which provide an incomplete picture of the airway microbiota and little opportunity for viral discovery compared to metagenomic techniques.

Metagenomics is a powerful tool for viral and microbial community characterization since nucleic acids are isolated directly from environmental samples and sequenced, requiring no culturing, cloning, or *a priori* knowledge of what viruses may be present. Viruses are the most numerous and diverse biological entities on Earth, and metagenomics has been used extensively to describe viral communities in marine ecosystems (9-12). The first metagenomic studies of the human microbiome were of viruses in blood, feces, and the lungs, and went far to describe previously unexplored viral communities (13-17). Recent metagenomic studies of the human microbiome have largely focused on microbial populations, predominantly in the gut and the surface of the skin (18-21).

Cystic fibrosis is an autosomal recessive genetic disease caused by a mutation in the cystic fibrosis transmembrane conductance regulator protein (CFTR), a gated ion channel (22, 23). CF affects paranasal sinuses as well as the lower respiratory, hepatobiliary, pancreatic and lower gastro-intestinal tracts (23). The current median age of survival for individuals with CF is approximately 38 years. Over 80% of CF mortalities are attributable to respiratory failure from chronic bacterial infections of the lungs, most commonly caused by *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Burkholderia cepacia (4, 24)*. Individuals with CF have impaired mucociliary clearance (MCC) which results in airway mucus plugging (2)(25)(2). This creates hypoxic microenvironments, forcing invasive microbial species to adapt (2). This unique airway environment is believed to increase viral replication and susceptibility to viral infections in individuals with CF (8, 22). Expectorated sputum provides a sample of airway secretions from the proximal airways. Sputum also contains material from the entire respiratory tract including airway mucus, cellular debris, DNA, and degraded proteins as well as microbes, their associated phage, and eukaryotic viruses (26, 27).

Here we report the first metagenomic study of airway DNA viral communities using sputum samples from both cystic fibrosis and Non-cystic fibrosis (Non-CF) individuals, including the spouse of an individual with CF and an individual with mild asthma. Viral communities from Non-CF volunteers were characterized and compared to viromes of individuals with cystic fibrosis to determine if there is a core respiratory tract virome in non-diseased individuals. Metabolic profiles inferred from metagenomic sequences were distinctly different between Non-CF and CF viromes. Our results indicate that regardless of the presence or absence of shared taxa, a core set of metabolic functions defines the non-diseased and diseased respiratory tract DNA viromes.

Results and Discussion

Phage taxonomy reflects airway pathology

In all metagenomes, the majority of sequences (>90%) were unknown when compared to the non-redundant database using BLASTn (Table S1), which is comparable to the percentage of unknown sequences in other environmental viromes (9, 12, 28). CF viromes had more tBLASTx similarities to phage genomes overall than Non-CF viromes, and were similar to a wider range of phage (Figure 4.1, Table S2). The tBLASTx analysis identified a core set of 19 phage genomes which had similarities to sequences in all metagenomes (Table S3). An additional 12 genomes had significant similarities to viromes from all CF individuals but none of the Non-CF individuals. This suggests a core set of phage characteristic of the human respiratory tract, and an additional core group in CF individuals. A few phage genomes appeared to dominated the Non-CF2, Non-CF3, and Non-CF5 viromes when tBLASTx similarities to phage genomes were plotted against the Phage Proteomic Tree (Figure 4.1). Over 90% of tBLASTx hits to phage in Non-CF2 were to *Streptococcus* phage Cp-1, and 80% of tBLASTx similaritites in Non-CF3 were attributable to two phage, *Haemophilus influenza* phage HP-1 and *Brucella melitensis* 16M BrucI prophage. The large relative abundance of these phage may reflect their prevalence in inhaled air, since environmental air has been shown to contain diverse bacterial communities (29).

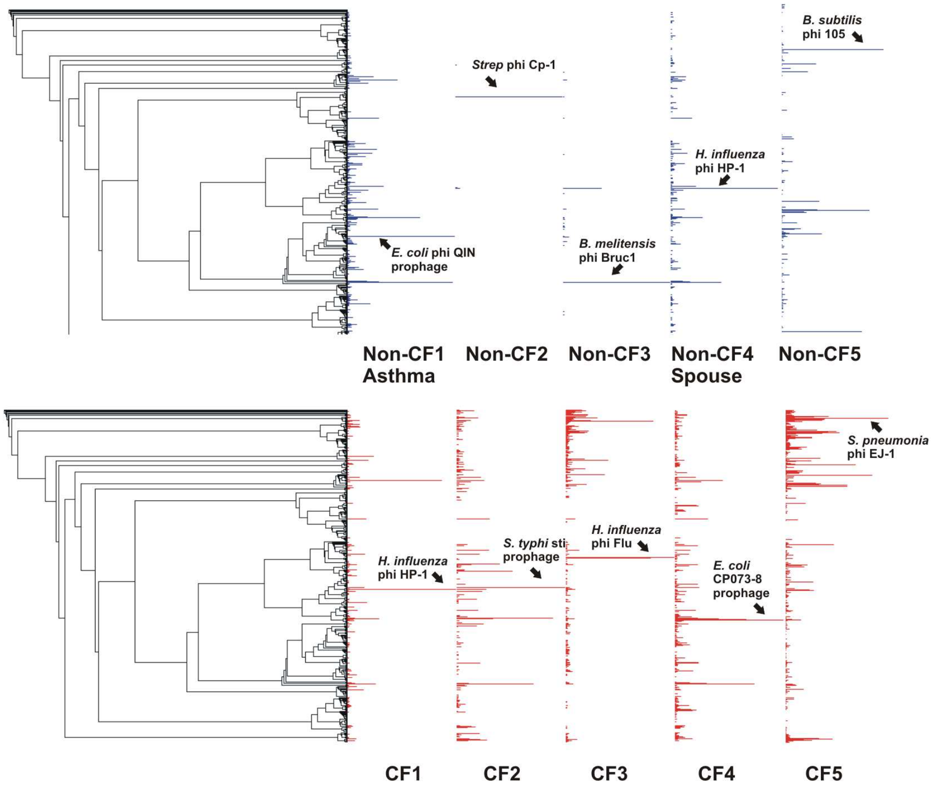


Figure 4.1: Mapping of best tBLASTx hits to the phage proteomic tree by percentage for Non-CF (A) and CF (B) viromes. The phage genome with the highest percentage of hits (normalized to the length of the genome) is labeled for each virome.

The phage profiles of Non-CF1Asthma and Non-CF4Spouse were more similar to those of CF individuals than to other Non-CF individuals. This likeness was confirmed by PCA (Figure 4.2). Non-CF1Asthma and Non-CF4Spouse had values for the first and second principal components which were nearly identical to those of the CF metagenomes. The other Non-CF metagenomes had more random distribution of phage genotypes and did not appear to cluster on the PCA graph. More specifically, Non-CF2, NonCF3, and Non-CF5 all had positive values for the first principal component (0.40, 0.42, and 0.27 respectively) while all other metagenomes had negative values. This was driven by a large positive loading of the first principal component by the *Streptococcus* phage Cp-1, which segregated Non-CF2, and negative loadings on the set of phage genomes shared by Non-CF1Asthma, Non-CF4Spouse and the CF metagenomes. Additionally, the second principal component was positively loaded by the *Brucella melitensis 16M* phi Bruc1 prophage genome which was nearly absent in Non-CF2, giving a negative value of the second principal component for Non-CF2.

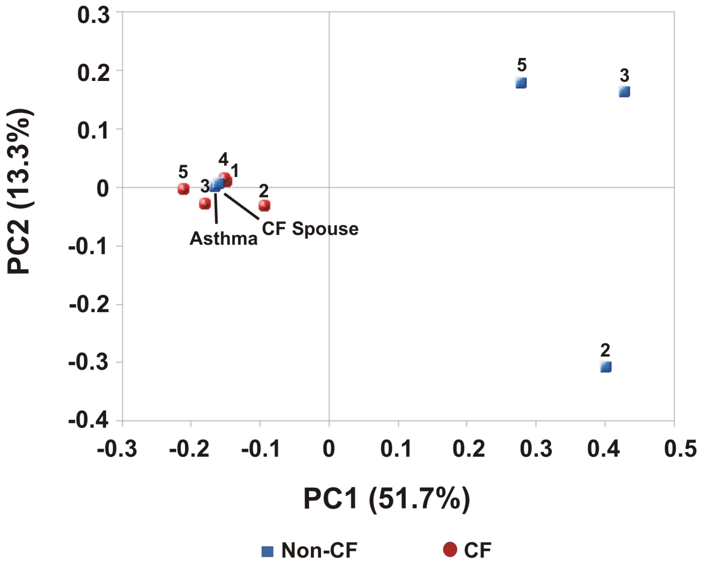


Figure 4.2: Principal components analysis (PCA) of respiratory tract viromes based on phage taxonomic composition. Non-CF metagenomes are shown in blue and CF metagenomes are shown in red. Inputs to PCA were normalized percentages of best tBLASTx hits to completely sequenced phage genomes. Non-CF1Asthma and NonCF4Spouse cluster with the CF metagenomes.

These results indicate that the sputum phage community in Non-CF individuals appears to represent a random, transient sampling of the exterior environment. In CF individuals, phage communities are driven by airway pathology, and correspond to a shared internal respiratory environment. The phage community in the Non-CF4Spouse virome reflects a continuous sampling of CF-associated phage via a shared external environment. Common phage taxonomy in CF individuals and Non-CF1Asthma occurs because of shared respiratory pathology (i.e., similar internal environments). Both CF and asthma are conditions marked by impaired mucociliary clearance (MCC) (2, 25, 30). MCC is slowed in asthma, leading to increased retention of microbes and hence their phage (30). In CF, mucus is extremely viscous and stagnant, forming obstructive plugs, and creating hypoxic microenvironments that serve as scaffolds for bacterial biofilm formation (2, 25). Therefore, in both asthma and CF, phage communities are derived from microbes which persist in the airways for longer periods of time than in healthy individuals.

Inferred host ranges for respiratory tract phage

The putative microbial host range of respiratory tract phage reflected a few dominant but distinct phage in Non-CF2, Non-CF3, and Non-CF5 (Figure 4.3).Host ranges of Non-CF1Asthma and Non-CF4Spouse were highly similar to those of the CF phage communities, but were under-represented in *Streptococcus* and *Staphylococcus* phage. The higher abundance of *Staphylococcus* phage in CF is consistent with the increased induction of *Staphylococcus* prophage by antibiotics in CF individuals, as shown by previous studies (31). *P. aeruginosa* was cultured from the sputum of all CF participants, yet *Pseudomonas* phage were not abundant in the metagenomes. *Pseudomonas* phage may be of novel types not closely related to those in the database, making them undetectable by tBLASTx. Even if known phage are present, infections of *Pseudomonas* in CF may be unsuccessful, since phage may not be able to penetrate the biofilm to access susceptible microbial hosts (32). Alternatively, *P. aeruginosa* may not be as abundant in the CF airway as indicated by culturing, an idea supported by 16S rDNA and Terminal Restriction Fragment Polymorphism (T-RFLP) analysis of bacteria in CF sputum and bronchoalveolar lavage fluid (33-35)). T-RFLP uses fluorescently labeled 5' PCR primers coupled with restriction digests to allow for rapid profiling of unknown microbial communities, providing a less biased picture of microbial diversity than culture-based studies (35).

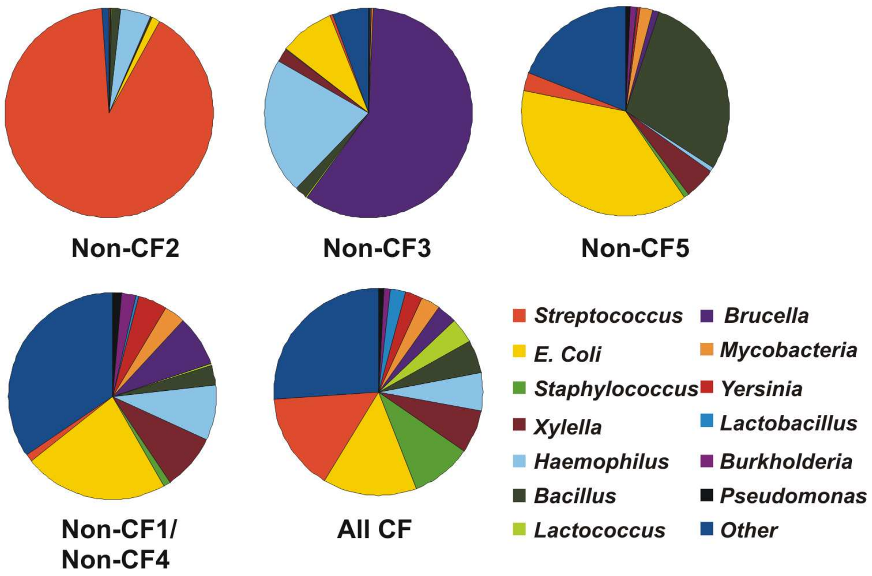


Figure 4.3: Putative host range for phage communities in respiratory tract viromes. Host range was inferred from normalized best tBLASTx hits to phage genomes. Host ranges for CF viromes and for Non-CF1Asthma and NonCF4Spouse were not statistically significantly different as determined by XIPE and were combined.

Diversity of respiratory tract viruses

There were approximately 175 unique species of DNA viruses in respiratory tract viral communities (Table 4.1). There were no significant differences in the estimated number of species between CF and Non-CF viromes. Diversity estimates were based on sequence assemblies and PHACCs, so all metagenomic sequences were used, not just those with BLAST similarities to viral databases (36). The estimated number of DNA viral species has been reported to be as low as 1440 in hot springs, and as high as 129,000 in the open ocean (9, 37). In comparison with other environmental viromes, the respiratory tract viromes had low species richness. Similarly, Rogers et al. (34) found low diversity of Bacteria in CF sputum using T-RFLP analysis. Low species richness probably results from physical and biological barriers to microbial and viral persistence, including both MCC as well as innate and adaptive immunity (2, 38). Richness may be further depressed in CF individuals because of antibiotic therapies and the metabolic adaptations required for microbial and viral survival in the unique microenvironment of the CF airway (26, 27).

Table 4.1: Diversity estimates for human respiratory tract DNA viromes. Repeated sets of 10,000 random sequences were retrieved from each metagenome and assembled to obtain contig spectra. Diversitymodeling based on contig spectra was performed with PHACCs, using a logarithmic model and an average genome size of 50 kb.

Sample	Species Richness	Evenness	Shannon Index
Non-CF1Asthma	164	0.89	4.52
Non-CF2	156	0.95	4.81
Non-CF3	113	0.94	4.45
Non-CF4Spouse	187	0.94	4.92
Non-CF5	594	0.86	5.46
Non-CF Mean	**243**	**0.92**	**4.83**
CF1	69	0.85	3.85
CF2	154	0.86	4.34
CF3	104	0.80	4.32
CF4	121	0.92	4.42
CF5	75	0.84	3.91
CF Mean	**105**	**0.85**	**4.17**
Overall Mean	**174**	**0.89**	**4.5**

Cross-BLASTn analysis showed that CF viromes shared more sequences with each other than Non-CF viromes. Sequences from each metagenome were compared pairwise to all other metagenomes using BLASTn to identify shared sequences as explained in Methods (39). The majority of the common CF sequences were not found in any Non-CF metagneomes. Sequential BLAST analysis identified 31,413 sequences common to all CF viromes, and 12,824 of these did not appear in any of the Non-CF viromes. Non-CF viromes shared 11,995 sequences, and 330 could not be found in any CF virome. Both the larger group of shared and unique sequences in CF metagenomes suggests that CF viral communities are more similar than Non-CF communities.

Taxonomy of eukaryotic viruses

Eukaryotic DNA viral communities in CF individuals were dominated by a few viral genomes which were highly variable in their abundances. Non-CF individuals shared numerous eukaryotic viruses with more even abundances, suggestive of a core virome (Figure 4.4A). All CF metagenomes had similarities (>1%) to Reticuloendotheliosis virus (Figure S1) and other retro-transcribing viruses (Figure 4.4B). We confirmed bioinformatically that similarities to retroviruses were not actually similarities to the human genome, therefore, we assume that retroviruses must have been present in the metagenomes as DNA intermediates. indicating that retroviruses may establish persistent infections in the airways, and could be useful therapeutic vectors for CF as previously suggested (40). CF viromes also shared several human herpesviruses (HHV) including Epstein-Barr virus (HHV-4), HHV-6B, and HHV-8P. Infection with Epstein-Barr virus in adolescent CF patients has been linked to exacerbations and poor outcomes, and has also been observed in adults (41).

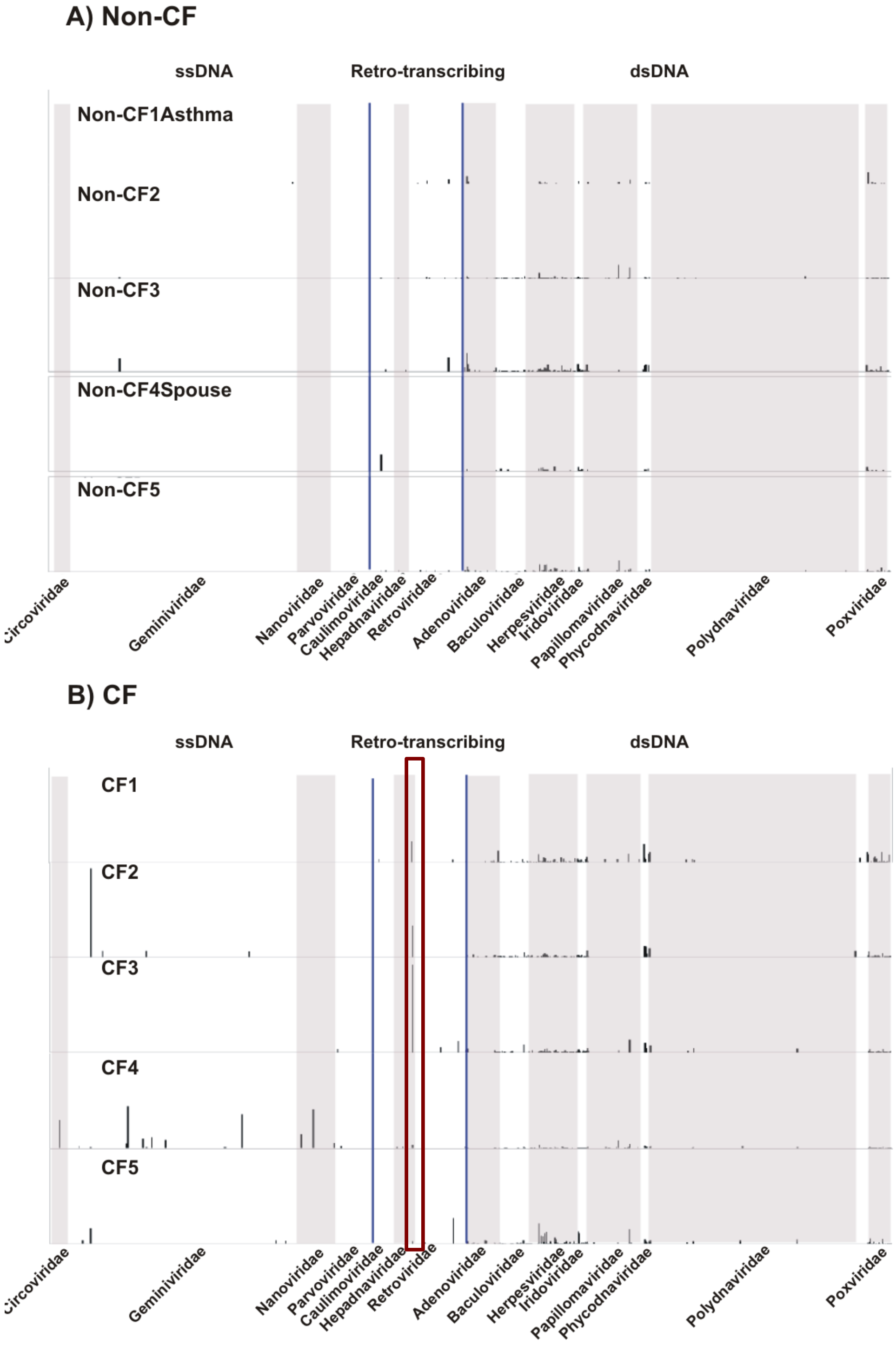


Figure 4.4: Distribution of normalized best tBLASTx hits to DNA and Retro-transcribing eukaryotic viruses in Non-CF(A) and CF(B) individuals. Reticuloendotheliosis virus is indicated by the red rectangle in (B).

Metabolic profiles of respiratory tract viruses

Non-CF individuals shared a common viral metabolic profile which was distinctly different from that of CF individuals (Figure 4.5). Functional annotations were assigned to metagenomic sequences by tBLASTx comparison to the non-redundant SEED database at the highest subsystem level, which consists of 25 classifications (Figure 4.6A). The percentage of known sequences (i.e., sequences with significant similarity to the database) was much higher than reported in the literature for other viral metagenomes (Figure S3) (28).

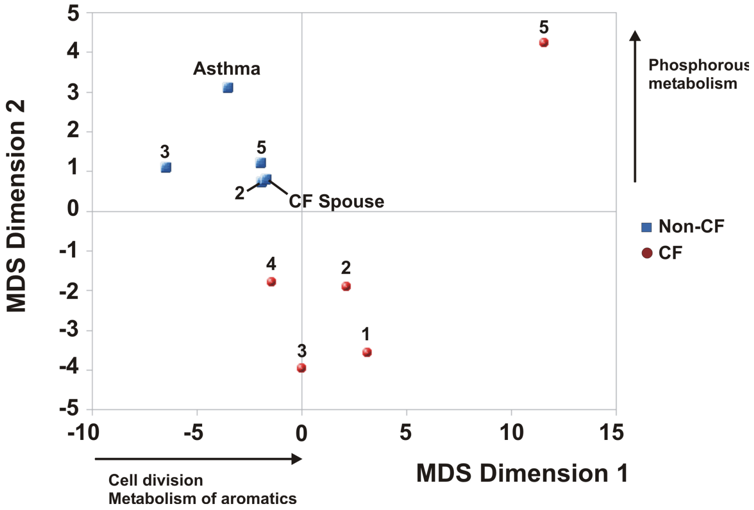


Figure 4.5: Non-metric multidimensional (NM-MDS) scaling of top-level SEED metabolic subsystems. All Non-CF metagenomes are shown in blue. CF1-5 are shown in red. The inputs to NM-MDS were the number of hits to subsystems in the highest level of the SEED hierarchy.

Metabolic functions encoded by viruses are determined by the environment, and functional genes carried by phage largely mirror those of their hosts (28). The CF airway has distinct regions characterized by hypoxia and low pH, and airway secretions are enriched in amino acids, DNA, phospholipids and other cellular debris (4, 26). The specific adaptations required for survival in this environment are reflected by the metabolic profiles of CF viromes. Non-CF1Asthma and Non-CF4Spouse shared phage taxonomy with CF viromes, but did not share metabolic profiles because they have a Non-CF airway environment. These results are similar to findings in the human gut, where microbiomes were determined to share a set of core metabolic genes even when different microbial taxa were present, and aberrant physiological states (i.e., obesity) lead to definitive changes in the metabolic consortium (21). As indicated by CF5, there may be more than one disease state which defines metabolism in CF, reflecting differences in pathology, disease development and/or treatment regimes.

All of the CF metagenomes (including CF5) were over-represented in functions related to the metabolism of aromatic compounds (Figure 4.6A). At the second hierarchical subsystem level, CF1-4 were over-represented in anaerobic degradation of aromatics, while CF5 had more genes related to peripheral catabolism pathways, most of which were aerobic (Figure 4.6B). Non-CF metagenomes were enriched for metabolism of central intermediates via aerobic mechanisms. CF sputum is derived from hypoxic microenvironments which require persistent microbes to acquire anaerobic adaptations (26). Aromatic amino acids have been implicated both as preferred carbon sources and also regulators of quinolone signaling and biofilm formation for *Pseudomonas aeruginosa* in CF sputum (26, 27).

The presence of anaerobic aromatic catabolism genes in phage may represent lateral gene transfer with well-adapted hosts (46). Alternatively, phage may be degrading aromatics in order to reduce biofilm formation and the exopolysaccharide layer, allowing access to susceptible Bacterial hosts.

CF5 was dramatically over-represented in phosphorous metabolism and virulence pathways (Figure 4.6A). Over 75% of tBLASTx similarities to the phosphorous metabolism subsystem were to the gene encoding Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase. This enzyme catalyzes the removal of a phosphate group from guanosine pentaphosphate (pppGpp) to generate guanosine tetraphosphate (ppGpp) (47). Both pppGpp and ppGpp are part of the canonical bacterial stringent response which is enacted to slow growth rates during nutrient stress (48). They have also been linked to bacterial virulence, antibiotic resistance, biofilm formation, quorum sensing, and phage induction in a variety of bacteria including *Pseudomonas aeruginosa* (47, 48). For many bacteria ppGpp is a more potent effector molecule than pppGpp, suggesting a need for increased levels of Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase (47, 49).

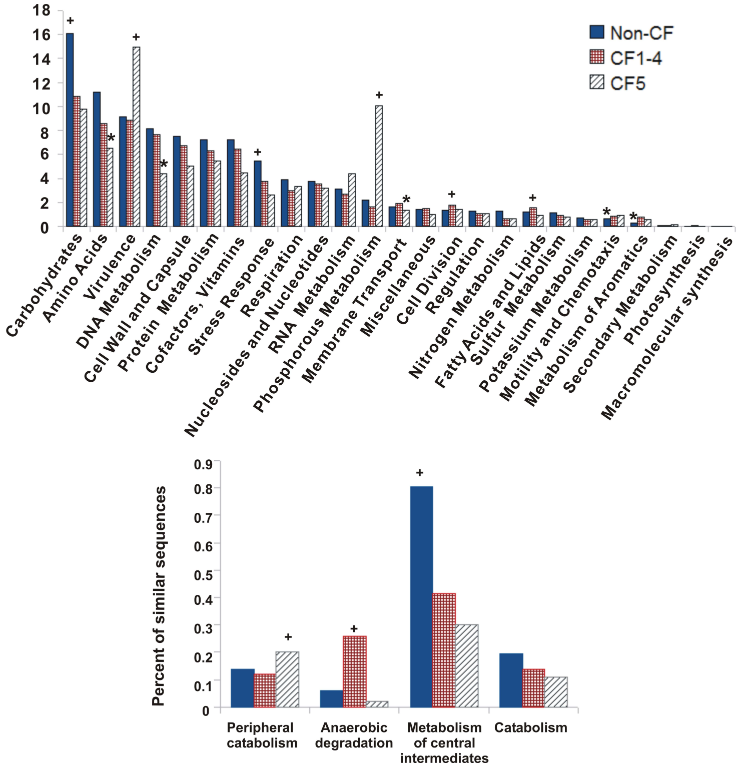


Figure 4.6: (A) Distribution of metabolic subsystems in respiratory tract viromes. (B) Second-level subsystems from the SEED hierarchy for aromatic metabolism. Subsystems over-represented in a group are marked with a (+) while those that are under-represented are marked with an asterisk (*).

Additional considerations for human microbiome studies

We used sputum samples as a proxy for the human respiratory tract, much as fecal samples have been used as a proxy for the human gut (18-20). Expectorated sputum has been routinely exploited as a rapid, inexpensive, non-invasive method to sample the lung and lower respiratory tract, and sputum samples can achieve sensitivity and accuracy comparable to bronchoalveolar lavage for detection of respiratory infections (50). T-RFLP analysis of bacterial communities demonstrated that sputum is not substantially contaminated by saliva and bacterial flora of the oral cavity (34). However, the degree to which sputum represents the upper and lower respiratory tract is unknown, especially in healthy individuals. Microbial communities in fecal samples have been shown to differ significantly from those in intestinal mucosal samples, based on 16S rDNA analysis, and similarly, sputum samples may contain different communities than the lung or lower respiratory tract (19).

In this study, human genomic DNA contamination was detected bioinformatically and removed. Previously, we sequenced control viromes from CF sputum which were not DNase I treated. These metagenomes contained over 90% of sequences from human genomic DNA as determined by BLASTn analysis (data not shown). This human DNA comes from neutrophils present in the airway, either through the active dissemination of neutrophil extracellular traps (NETs) or by the release of cellular contents during cell death (51). Using the protocol described above, the percent of human DNA detected ranged from 10% to 34% (Table 2). This was markedly lower than in the control metagenomes, and was comparable to the percentage of human DNA (24% and 36%) obtained by Allander et al. (16) for viral isolation from pooled nasal aspirate samples. As studies of the human microbiome move away from characterization of microbes using 16S rDNA and towards complete metagenomic analysis of both microbial and viral communities, human genomic DNA contamination becomes unavoidable.

After all contaminating sequences were removed, there were still at least 130,000 sequences comprising over 30 Mbp in all metagenomes. To verify the presence of viruses in the metagenomes, we assembled two metagenomes and compared contigs to the non-redudant database using BLASTn. There were 23 contigs assembled from Non-CF2 which had BLASTn matches to *Streptococcus* phage Cp-1 (E-value < 10^{-5}), with an alignment length greater than 50 bp, and greater than 85% identity (Figure S4). The assembly of the CF3 metagenome yielded high coverage and significant BLASTn hits to the genome of *H. influenza* prophage Mu (Figure S5).

Here, we isolated DNA viruses from sputum, including both phage and eukaryotic viruses. The majority of respiratory infections (>75%) have been attributed to RNA viruses such as rhinoviruses, coronaviruses, and paramyxoviruses, so many previous studies have focused on the characterization of RNA viruses in the respiratory tract (38). CF is predominantly a microbial disease, and phage are known to exert important top-down controls on microbial communities (52). However, little work has been done to describe phage communities and DNA viruses associated with CF or with the airways in general (4, 38). Over 98% of all completely sequenced phage have DNA genomes, therefore to assess phage diversity, taxonomy, and function, it was necessary to isolate viral DNA (53). Future studies of the respiratory tract virome should be expanded to include characterization of RNA viral communities.

A caveat to this study was the use of Multiple Displacement Amplification (MDA) with phi29 polymerase to amplify viral DNA prior to pyrosequencing. MDA generally provides an even representation of genomes except at the ends, however, certain genomes (small and circular or large and linear) may be preferentially amplified (54, 55). To avoid random biases introduced by initial reaction conditions, we performed five separate amplifications which were then combined. All of the metagenomes used here were collected, processed and amplified in an identical manner, so any biases would have been introduced equally in all samples.

Conclusions

Metagenomic analysis of the human respiratory tract DNA virome illustrated that airway viral communities in the diseased and non-diseased states are defined by metabolism and not by taxonomy. The non-diseased airway virome contains a set of shared core metabolic functions, which deviate strongly in the face of chronic disease. These deviations are driven by dramatic environmental changes in the airways, induced by the nature of cystic fibrosis, such as the introduction of hypoxic microenvironments and novel carbon sources (26, 27). In cases where phage taxonomy was shared between Non-CF and CF individuals, metabolic functions still remained distinct. The converse was also true, that is, even when Non-CF viromes differed in phage and eukaryotic viral constituents, they maintained typical Non-CF metabolic profiles. The presence of two alternative metabolic states in CF reflects the heterogenous nature of disease. Though CF is generally considered to be well-characterized, there is still inherent individual variation. The need for alternative therapies for CF is increasing, as microbial antibiotic resistance becomes widespread. The results of this study suggest that CF therapeutics might be better aimed at changing the environment of the airways rather than targeting dominant taxa.

Methods

Ethics statement

Subject recruitment and sample collection were approved by the San Diego State University Institutional Review Board (SDSU IRB 2121) and Environmental Health Services (BUA 06-02-062R). Written consent forms were obtained from all study subjects.

Study population

The five individuals with CF who volunteered for this study were patients at the Cystic Fibrosis Foundation accredited Adult cystic fibrosis Clinic at the University of California San Diego Medical Center. Patients were eligible if they could be classified as clinically stable (i.e., in a non-exacerbated state and free from systemic antibiotic therapy for at least thirty days), and had no reportable cold or flu-like symptoms in the previous thirty days. All volunteers with CF were screened for signs and symptoms of a upper respiratory infection for the thirty days prior to the study. All CF subjects were required to have a well documented diagnosis with either two known mutations in the cystic fibrosis Transmembrane Regulator (CFTR) or an abnormally high sweat chloride test. In addition, all CF patients had *Pseudomonas aeruginosa* present in their sputum, as determined by culturing in the clinic's microbiology lab. The five CF individuals randomly selected for the study consisted of two males and three females. The age range was from 20 to 35 years and all patients had severe airway obstruction as assessed by standard spirometry (FEV1<50% of predicted).

Four Non-CF volunteers were recruited from the campus of San Diego State University, and were subject to the same exclusion criteria for upper respiratory infection. One of these Non-CF individuals had mild asthma controlled by medication. A final Non-CF volunteer was the spouse of a CF patient and was recruited from the greater San Diego area. The five Non-CF individuals consisted of four females and one male, with an age range of 24 to 50 years.

Sample collection

Sputum samples of approximately 10 ml were obtained from CF patients at the Adult cystic fibrosis Clinic by expectoration into a sterile cup, as directed by clinic staff. Since sputum expectoration is difficult in general for Non-CF individuals, all Non-CF subjects were first required to do an oral rinse with water to prevent excessive salivary contamination and then take five deep breaths to loosen lung secretions. Subjects were then instructed to cough deeply into a sterile cup. The deep breathing and coughing procedures were repeated until at least 1 ml of sputum was obtained.

Metagenomic library preparation

All sputum samples were diluted with an equal volume of Suspension Medium

(SM) buffer (1M NaCl, 10 mM $MgSO_4$, 50 mM Tris-HCl pH 7.4). To aid in mucus dissolution, samples were incubated with 10 ml of 6.5 mM dithiothreitol (Acros Organics: Morris Plains, New Jersey) for 30 minutes at 37^o C. The treated sputum was homogenized using a PowerGen 125 mechanical homogenizer (Fisher Scientific: Hampton, New Hampshire) until it was uniform in color and there was no visible particulate debris. Homogenized samples were filtered through a 0.8 micron black polycarbonate filter (GE Water & Process Technologies: Trevose, Pennsylvania) followed by a 0.45 micron MILLEX®HV filter (Millipore: Carrigtwohill, Colorado) to remove eukaryotic and microbial cells. Viruses in the 0.45 micron filtrate were purified and concentrated using a cesium chloride (CsCl) gradient to remove free DNA and any remaining cellular material (56). After collection of viral concentrates from the CsCl gradient, the presence of virus-like particles (VLPs) and the absence of microbial contamination were verified by epifluorescence microscopy using SYBR® Gold (Invitrogen: Eugene, Oregon) as described in (56). Sputum samples from healthy subjects contained approximately 10^7 VLPs per ml, while the samples from CF patients contained approximately 10^9 VLPs per ml. A sample epifluoresence micrograph is shown in Figure S6. Chloroform was added to the viral concentrates to rupture the membranes of any remaining cells. Following a one hour incubation and centrifugation, choloroform was removed by pipetting. To degrade any remaining free DNA prior to viral DNA extraction, samples were treated with 2 units per µl of Dnase I (Sigma-Aldrich: St. Louis, MO) at 37˚C for 1 hour. Viral DNA was isolated using CTAB/phenol:choloroform extractions and amplified using multiple displacement amplification with Phi29 polymerase (56). Viral DNA was sequenced at 454 Life

Sciences (Branford, CT) using the GSFLX pyrosequencing platform to produce ten total viral metagenomic libraries. The ten viral metagenomes are accessible from NCBI (www.ncbi.nlm.gov) under the genome project ID 39545.

Initial bioinformatic processing of metagenomes

All metagenomes were compared to the Human Genome build 36.3 (http://www.ncbi.nlm.mih.gov) using BLASTn to determine how effective the combination of cesium chloride density gradient centrifugation and DNase I treatment was for removing human genomic contamination from the viral preps (39). Sequences with 80% identity over 80% of their length to human sequences were considered contaminating human genomic DNA and were removed prior to further bioinformatic analyses.

Following removal of human sequences, dinucleotide relative abundance analysis was used as a secondary screen to detect human DNA contamination, which manifests as an overall depression of CG dinucleotides (57, 58). In all of the decontaminated metagenomes, the relative abundance odds ratios for CG dinucleotides were between 0.83 and 1.09, within the normal range as defined by Karlin, indicating successful removal of human DNA (Table 2) (57, 58). All viromes were AT rich (in comparison to microbial metagenomes) as expected, with GC content between 40-43%, just below the average of approximately 45% previously reported for viral metagenomes (58). The human genomic DNA decontaminated metagenomic libraries were named according to the subject group they were derived from (Non-CF or CF) and were numbered 1 through 5 in each group. Viromes derived from the individual with asthma and the CF spouse were designated as Non-CF1Asthma and Non-CF4Spouse.

Diversity estimation

To estimate viral diversity and community structure within metagenomes, contig spectra were generated using the free software Circonspect (http://sourceforge.net/projects/circonspect/). Average contig spectra were calculated using assemblies of 10,000 randomly selected sequences with enough repetitions to achieve 2X coverage of each metagenome. The assembly parameters were 98% minimal match and 35 base pair overlap. Sequences less than 100 base pairs were discarded and all other sequences were trimmed to 100 base pairs prior to assembly to obtain identical sequence size in the repeated assemblies. Average contig spectra were used as inputs to Phage Communities from Contig Spectra (PHACCS) tool (http:biome.sdsu.edu/phaccs), which estimates diversity using rank-abun(36). Diversity estimates were based on the best-fit model, in this case the logarithmic model.

Sequential BLAST analysis

Metagenomic libraries were compared to each other using BLASTn to find shared sequences between all Non-CF viromes and all CF viromes. One metagenome from each set (Non-CF or CF) was chosen randomly and compared to a second randomly selected metagenome. Common sequences (E-value<10^{-5} and a minimum of 98% similarity over at least 35 base pairs) were identified and then used as a database for BLASTn versus a third metagenome. This was repeated for the fourth and fifth metagenomes. The entire process was repeated using a different random ordering of metagenomes. Sequential BLASTn analysis resulted in two datasets, one containing sequences common to all Non-CF metagenomes and the other with sequences common to CF metagenomes. The common Non-CF sequences were then compared using BLASTn to all CF metagenomes to determine which sequences were not present in any CF library (i.e., unique to Non-CF individuals). This was also performed in reverse, to find unique CF sequences.

Comparison to phage and viral genome databases

Metagenomic libraries were compared to two boutique databases, the first containing 510 complete phage genomes (http://phage.sdsu.edu/phage) and the second, 3,074 complete eukaryotic viral genomes (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) using tBLASTx with an E-value cutoff of 10^{-5} (39). Counts of best tBLASTx similarities to each genome were normalized for genome size by weighting the number of significant similarities by the total number of base pairs in the database divided by the size of the genome in base pairs. Similarity counts were also normalized for the number of sequences per metagenome, to allow direct comparisons between metagenomes. Normalized best tBLASTx similarities to the phage database were plotted against the Phage Proteomic Tree version 4 (http://phage.sdsu.edu/~rob/PhageTree/v4) using Bio-Metamapper (53, 56). Similarities to dsDNA, ssDNA, and retro-transcribing eukaryotic viruses were plotted according to NCBI taxonomy (http://www.ncbi.nlm.nih.gov/genome).

Similarities to RNA viruses were not included because they were artifactual, since only DNA was sequenced in this study. Significant similarities to RNA viruses comprised less than 1% of all tBLASTx similarities.

Assessment of metabolic potential

The metabolic potential of each virome was assessed by BLASTx (e-value < 10^{-5}) comparison to the SEED database using the MG-RAST service(59,60). MG-RAST assigns sequences to three hierarchical levels of metabolic subsystems, which consist of groups of genes that comprise a metabolic function or pathway (61). The non-parametric statisical program XIPE was used to detect significant differences between metabolic profiles of viral metagenomes at a 95% confidence level (62). XIPE identifies the specific subsystems driving the differences between metagenomes, and in which metagenome the function was are over-represented.

Complete metagenomic assembly

Complete assembly of the Non-CF2 and CF3 metagenomes was performed using PHRAP as a quality check to confirm sucessful isolation of viral genomes (63). These two metagenomes were assembled because they had high coverage of phage genomes as indicated by tBLASTx. There were 9,508 contigs ranging in size from 40 to 14,982 bp for Non-CF2, and 8,163 contigs from 212 to 7,748 base pairs for CF3. Contigs were compared to the non-redundant nucleotide database maintained at NCBI (http://www.ncbi.nlm.nih.gov) using BLASTn to assign taxonomy.

Statistical analyses

All statistical analyses, with the exception of XIPE, were performed using the software package R (www.r-project.org) (64). Principal components analysis (PCA) with the R function *prcomp* was used to examine overall taxonomic similarities between metagenomes (65). The first two principal components were used to generate 2D scatter plots. Non-metric multidimensional scaling (NM-MDS) with the R function *isoMDS* was used to determine relationships between metagenomes based on metabolic profiles. The analysis was performed with NM-MDS instead of PCA for metabolic potential because all metagenomes had at least one hit to each of the 25 subsystems (i.e., there were no zero values). Similar to PCA, NM-MDS does not require *a priori* classification of the data and plotting the MDS coordinates shows natural grouping patterns. Clusters observed in PCA and NM-MDS scatterplots were confirmed statistically using k-means clustering. To determine the optimal number of clusters, within-group sums of squares were calculated for partitions involving between 1 and 9 clusters (65)63). Cluster membership was determined by using the R function *kmeans* with the optimal number of clusters.

References

1. Heyder J (2004) Deposition of Inhaled Particles in the Human Respiratory Tract and Consequences for Regional Targeting in Respiratory Drug Delivery. *Proc Am Thorac Soc* 1:315-320.

2. Knowles MR, Boucher RC (2002) Mucus clearance as a primary innate defense mechanism for mammalian airways. *J Clin Invest.* 109:571–577.

3. Corne JM, Marshall C, Smith S, Schreiber J, Sanderson G, Holgate ST, Johnston SL (2002) Frequency, severity, and duration of rhinovirus infections in asthmatic

and non-asthmatic individuals: a longitudinal cohort study. *Lancet* 359:831-834.

4. Harrison F (2007) Microbial ecology of the cystic fibrosis lung. *Microbiology (Reading, Engl.)* 153:917-923.

5. McManus TE, Marley AM, Baxter N, Christie SN, O'Neill HJ, Elborn JS, Coyle PV, Kidney JC (2008) Respiratory viral infection in exacerbations of COPD. *Respiratory Medicine* 102:1575-1580.

6. Beringer PM, Appleman MD (2000) Unusual respiratory bacterial flora in cystic fibrosis: microbiologic and clinical features. *Curr Opin Pulm Med* 6:545-550.

7. Miller RV, Rubero VJ (1984) Mucoid conversion by phages of Pseudomonas aeruginosa strains from patients with cystic fibrosis. *J Clin Microbiol.* 19:717–719.

8. van Ewijk BE, van der Zalm MM, Wolfs TF, van der Ent CK (2005) Viral respiratory infections in cystic fibrosis. *Journal of Cystic Fibrosis* 4:31-36.

9. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4:e368.

10. Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl. Environ. Microbiol* 73:7629-7641.

11. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A* 99:14250-14255.

12. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452:340-343.

13. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003) Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol* 185:6220-6223.

14. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* 4:e3.

15. Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *BioTechniques* 39:729-736.

16. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B (2005) Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc. Natl. Acad. Sci. U.S.A* 102:12891-12896.

17. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, Maeda N, Kawai J, Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T(2009) Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. *PLoS ONE* 4:e4219.

18. Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3:e2836.

19. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312:1355-1359.

20. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027-1031.

21. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. *Nature* 457:480-484.

22. Livraghi A, Randell SH (2007) Cystic fibrosis and other respiratory diseases of impaired mucus clearance. *Toxicol Pathol* 35:116-129.

23. Kulczycki LL, Kostuch M, Bellanti JA (2003) A clinical perspective of cystic fibrosis and new genetic findings: relationship of CFTR mutations to genotype-phenotype manifestations. *Am. J. Med. Genet. A* 116A:262-267.

24. Cystic Fibrosis Foundation Patient Registry/ 2007 Annual Data Report to the Center Directors (2008) (Cystic Fibrosis Foundation, Bethesda, MD).

25. Randell SH, Boucher RC (2006) Effective mucus clearance is essential for respiratory health. *Am. J. Respir. Cell Mol. Biol* 35:20-28.

26. Palmer KL, Mashburn LM, Singh PK, Whiteley M (2005) Cystic fibrosis sputum supports growth and cues key aspects of Pseudomonas aeruginosa physiology. *J. Bacteriol* 187:5267-5277.

27. Palmer KL, Aye LM, Whiteley M (2007) Nutritional cues control Pseudomonas aeruginosa multicellular behavior in cystic fibrosis sputum. *J. Bacteriol* 189:8079-8087.

28. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F(2008) Functional metagenomic profiling of nine biomes. *Nature* 452:629-632.

29. Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, Yap J, Yao F, Suan ST, Ing SK, Haynes M, Rohwer F, Wei CL, Tan P, Bristow J, Rubin EM, Ruan Y (2008) The Airborne Metagenome in an Indoor Urban Environment. *PLoS ONE.* 3:e1862.

30. Rogers DF (2004) Airway mucus hypersecretion in asthma: an undervalued pathology? *Curr Opin Pharmacol* 4:241-250.

31. Goerke C, Wirtz C, Flückiger U, Wolz C (2006) Extensive phage dynamics in Staphylococcus aureus contributes to adaptation to the human host during infection. *Mol. Microbiol* 61:1673-1685.

32. Azeredo J, Sutherland IW (2008) The use of phages for the removal of infectious biofilms. *Curr Pharm Biotechnol* 9:261-266.

33. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Bruce KD (2004) Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16s ribosomal DNA terminal restriction fragment length polymorphism profiling. *J. Clin. Microbiol* 42:5176-5183.

34. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Kehagia V, Connett GJ, Bruce KD (2006) Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with cystic fibrosis. *J. Clin. Microbiol* 44:2601-2604.

35. Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol* 63:4516-4522.

36. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F (2005) PHACCS, an online tool for

estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6:41.

37. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from yellowstone hot springs. *Appl. Environ. Microbiol* 74:4164-4174.

38. See H, Wark P (2008) Innate immune response to viral infection of the lungs. *Paediatr Respir Rev* 9:243-250.

39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol* 215:403-410.

40. Goldman MJ, Lee PS, Yang JS, Wilson JM (1997) Lentiviral vectors for gene therapy of cystic fibrosis. *Hum. Gene Ther* 8:2261-2268.

41. Winnie GB, Cowan RG (1992) Association of Epstein-Barr virus infection and pulmonary exacerbations in patients with cystic fibrosis. *Pediatr. Infect. Dis. J* 11:722-726.

42. Vadivukarasi T, Girish KR, Usha R (2007) Sequence and recombination analyses of the geminivirus replication initiator protein. *J. Biosci* 32:17-29.

43. Klein F, Kotb WFMA, Petersen I (2008) Incidence of human papilloma virus in lung cancer. *Lung Cancer*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19019488 [Accessed May 27, 2009].

44. Lohavanichbutr P, Houck J, Fan W, Yueh B, Mendez E, Futran N, Doody DR, Upton MP, Farwell DG, Schwartz SM, Zhao LP, Chen C (2009) Genomewide gene expression profiles of HPV-positive and HPV-negative oropharyngeal cancer: potential implications for treatment choices. *Arch. Otolaryngol. Head Neck Surg* 135:180-188.

45. Zawadzka-Głos L, Jakubowska A, Chmielik M, Bielicka A, Brzewski M (2003) Lower airway papillomatosis in children. *Int. J. Pediatr. Otorhinolaryngol* 67:1117-1121.

46. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299-304.

47. Jain V, Kumar M, Chatterji D (2006) ppGpp: stringent response and survival. *J. Microbiol* 44:1-10.

48. Potrykus K, Cashel M (2008) (p)ppGpp: still magical? *Annu. Rev. Microbiol* 62:35-51.

49. Raskin DM, Judson N, Mekalanos JJ (2007) Regulation of the stringent response is the essential function of the conserved bacterial G protein CgtA in Vibrio cholerae. *Proc. Natl. Acad. Sci. U.S.A* 104:4636-4641.

50. Xiang X, Qiu D, Chan KP, Chan SH, Hegele RG, Tan WC (2002) Comparison of three methods for respiratory virus detection between induced sputum and nasopharyngeal aspirate specimens in acute asthma. *J. Virol. Methods* 101:127-133.

51. Wartha F, Beiter K, Normark S, Henriques-Normark B (2007) Neutrophil extracellular traps: casting the NET over pathogenesis. *Curr. Opin. Microbiol* 10:52-56.

52. Fuhrman JA, Schwalbach M (2003) Viral influence on aquatic bacterial communities. *Biol. Bull* 204:192-195.

53. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol* 184:4529-4535.

54. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A* 99:5261-5266.

55. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7:216.

56. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470-483.

57. Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* 11:540-546.

58. Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19302541 [Accessed May 27, 2009].

59. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.

60. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.

61. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691-5702.

62. Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7:162.

63. Green P *PHRAP* Available at: http://bozeman.mbt.washington.edu/phrap.docs/phrap.html.

64. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).

65. Everitt BS, Hothorn T (2006) *A Handbook of Statistical Analyses Using R* (Chapman & Hall/CRC). 1st Ed.

Acknowledgements

We thank Rob Edwards, Liz Dinsdale, and Paul Quinton for helpful discussions and critical readings of the manuscript. We are grateful to Roche/454 Life Sciences for sequencing our metagenomes.

Chapter 4 has been published in full as: Willner D, Furlan M, Haynes M, Schmieder R, Angly F, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. (2009) Metagenomic analysis of respiratory tract viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*. 4(10): e7370. The dissertation author was the primary investigator and author of this material. Supporting information is presented in the Appendix to this chapter.

Appendix

This appendix contains supporting information for Willner D, Furlan M, Haynes M, Schmieder R, Angly F, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. (2009) Metagenomic analysis of respiratory tract viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*. 4(10): e7370.

Table S1. Taxonomic designations of metagenomic sequences based on BLASTn (e-value<10−5) comparison to the non-redundant database at NCBI. Sequences which had no significant similarities were assigned as “unknown”, while those with significant similarities were considered to be “known”.

Library Name	**Known, %**	**Viral, %**	**Bacterial, %**	**Eukaryotic, %**
Non-CF1Asthma	4.76	0.30	99.46	0.01
Non-CF2	2.80	25.11	72.88	0.01
Non-CF3	7.89	0.05	99.64	0.01
Non-CF4Spouse	4.12	0.10	99.42	0.01
Non-CF5	4.20	0.06	98.63	0.01
Non-CF Average	4.75	5.12	94.01	0.01
CF1	8.44	0.02	99.73	0.01
CF2	6.40	0.11	99.57	0.01
CF3	7.61	6.24	93.65	<0.01
CF4	3.91	0.09	99.15	0.02
CF5	9.12	1.28	98.67	0.01

Table S2. Results of comparison of metagenomes to the database of 510 fully sequenced phage genomes using tBLASTx (e-value<10−5). The number of unique genomes refers to how many phage genomes had a significant BLAST similarity in only one of the five Non-CF or CF metagenomes.

	Number of hits (% of metagenome)	Number of genomes hit (% of database)	Number of unique genomes (% of genomes hit)
Non-CF1Asthma	5754 (2.38%)	220 (43.14%)	20 (9.09%)
Non-CF2	2020 (0.98%)	72 (14.12%)	6 (8.33%)
Non-CF3	2906 (1.29%)	73 (14.31%)	5 (6.85%)
Non-CF4Spouse	4851 (1.97%)	241 (47.25%)	31 (12.86%)
Non-CF5	782 (0.25%)	109 (21.37%)	15 (13.76%)
CF1	6888 (4.27%)	234 (45.88%)	4 (1.71%)
CF2	1192 (0.55%)	164 (32.16%)	3 (1.83%)
CF3	22365 (14.42%)	321 (62.94%)	13 (4.05%)
CF4	4198 (1.77%)	298 (58.43%)	16 (5.37%)
CF5	20264 (10.16%)	366 (71.76%)	34 (9.29%)

Table S3. Relative abundances of the 19 phage genomes which appear in all human respiratory tract viromes based on tBLASTx similarities (e-value<10−5). Relative abundances were calculated as the normalized number of similarities to each phage divided by the total number of similarities to phage for each metagenome.

	Non-CF1 Asthma	Non-CF2	Non-CF3	Non-CF4 Spouse	Non-CF5	CF1	CF2	CF3	CF4	CF5
***Aeromonas hydrophila* phi Aeh1**	0.05	0.01	0.05	0.06	0.04	0.02	0.05	0.01	0.03	0.04
***Aeromonas* phi 31**	0.80	0.01	0.03	0.38	0.12	0.19	0.12	0.06	0.22	0.02
***Bacillus cereus* phi phBC6A51**	0.02	0.22	0.09	0.15	0.07	0.08	0.12	0.13	0.28	0.05
***Bacillus subtilis* phi 105**	0.20	0.16	0.03	0.12	13.43	0.20	0.09	0.04	0.04	0.02
***Bacillus subtilis* phi SPBc2**	0.19	0.01	0.01	0.32	0.03	0.34	0.75	0.05	0.30	0.02
***Brucella melitensis* 16M phi Bruc1 prophage**	8.96	0.05	59.40	7.04	1.00	3.65	5.54	0.54	5.24	0.48
***Escherichia coli* phi CP073-4 prophage**	0.70	0.06	0.10	0.58	0.15	0.18	1.43	0.29	0.41	0.06
***Escherichia coli* phi CP4-6 prophage**	6.13	0.09	0.50	4.40	4.72	4.13	2.61	0.65	2.85	0.68
***Escherichia coli* phi QIN prophage**	9.11	0.05	0.20	0.04	2.54	0.25	0.18	0.01	0.47	0.06

Table S4. Results of comparison of metagenomes to the database of 3074 fully sequenced eukaryotic viral genomes using tBLASTx (e-value<10−5). The number of unique genomes refers to how many viral genomes had a significant BLAST similarity in only one of the five Non-CF or CF metagenomes.

	Number of hits (% of metagenome)	Number of genomes hit (% of database)	Number of unique genomes (% of genomes hit)
Non-CF1Asthma	4935 (2.04%)	113 (4.44%)	7 (6.19%)
Non-CF2	4881 (2.38%)	107 (4.20%)	9 (8.41%)
Non-CF3	3125 (1.38%)	101 (3.97%)	4 (3.96%)
Non-CF4Spouse	65 (0.03%)	29 (1.14%)	1 (3.45%)
Non-CF5	6122 (1.99%)	125 (4.91%)	12 (9.60%)
CF1	1984 (1.23%)	99 (3.89%)	24 (24.2%)
CF2	1895 (0.87%)	79 (3.10%)	7 (7.59%)
CF3	1687 (1.09%)	87 (3.42%)	5 (5.75%)
CF4	3253 (1.37%)	110 (4.32%)	26 (23.64%)

Table S5. Relative abundances of the 20 eukaryotic DNA viral genomes which appear in all human respiratory tract viromes based on tBLASTx similarities (e-value<10−5). Relative abundances were calculated as the normalized number of similarities to each virus divided by the total number of similarities to eukaryotic DNA viruses for each metagenome.

	Non-CF1 Asthma	Non-CF2	Non-CF3	Non-CF4 Spouse	Non-CF5	CF1	CF2	CF3	CF4	CF5
Acanthamoeba polyphaga mimivirus	0.91	0.14	2.83	0.38	0.92	1.17	2.47	0.33	0.19	0.26
Aedes taeniorhynchus iridescent virus	0.31	0.03	0.08	0.29	0.04	0.16	0.15	0.02	0.03	0.03
Amsacta moorei entomopoxvirus 'L'	1.01	0.15	2.61	1.44	0.79	1.77	2.06	0.27	0.15	0.18
Bovine adenovirus A	0.12	0.06	0.84	0.01	0.18	0.29	0.67	0.32	0.12	0.30
Bovine herpesvirus 5	0.60	0.04	0.98	1.22	0.51	0.29	0.25	0.11	0.04	1.18
Cercopithecine herpesvirus 1	0.50	0.04	0.49	1.42	0.39	0.19	0.11	0.05	0.02	1.04
Cercopithecine herpesvirus 16	0.92	0.09	1.38	0.36	2.73	1.08	1.27	0.29	0.10	1.60
Cercopithecine herpesvirus 2	0.65	0.04	1.05	0.37	2.76	0.92	0.87	0.30	0.04	1.73
Cercopithecine herpesvirus 9	0.07	0.02	2.73	0.45	0.09	0.16	0.18	0.03	0.00	0.02
Chlorella virus ATCV-1	3.40	0.51	7.69	0.58	1.81	4.06	4.31	0.90	0.61	0.61
Chlorella virus FR483	0.71	0.19	2.33	0.35	0.59	0.96	3.97	0.27	0.32	0.31
Ectocarpus siliculosus virus 1	0.83	0.12	2.72	0.33	0.49	0.63	1.07	0.34	0.24	0.14
Frog virus 3	0.04	0.09	0.59	0.53	0.08	0.51	0.70	0.01	0.02	0.16
Human herpesvirus 1	0.29	0.56	0.50	1.83	1.21	0.39	0.19	0.08	0.18	0.90
Human herpesvirus 2	0.42	0.01	0.22	0.00	0.92	0.64	0.66	0.22	0.05	0.60

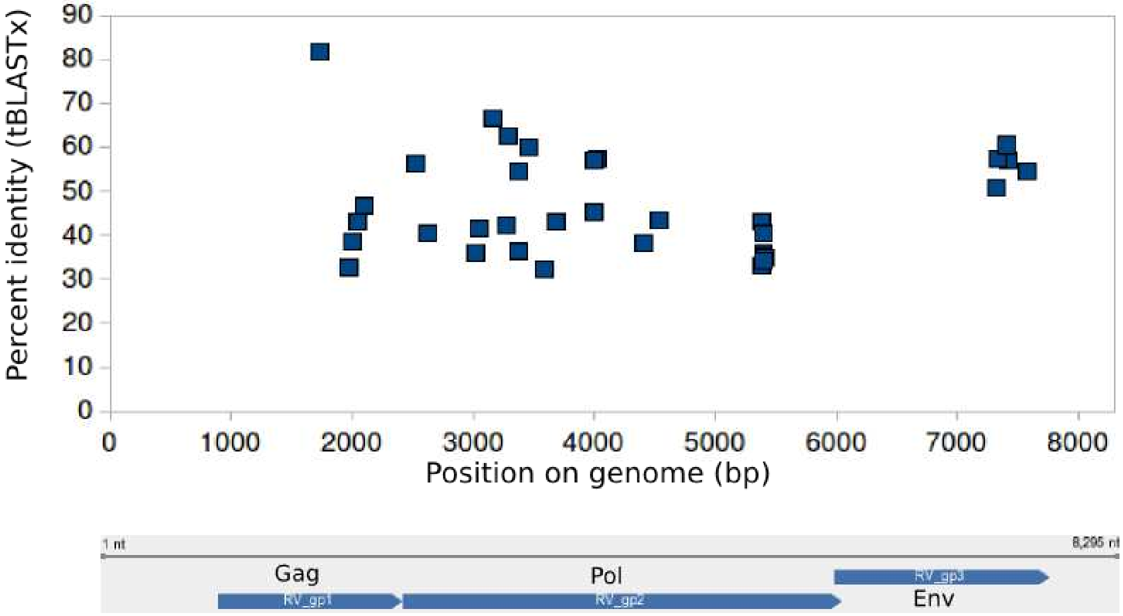


Figure S1. Combined coverage of Retiucloendotheliosis virus across all CF metagenomes as determined by tBLASTx. The graphic of the 8295 kb Reticuloendotheliosis genome is from NCBI.

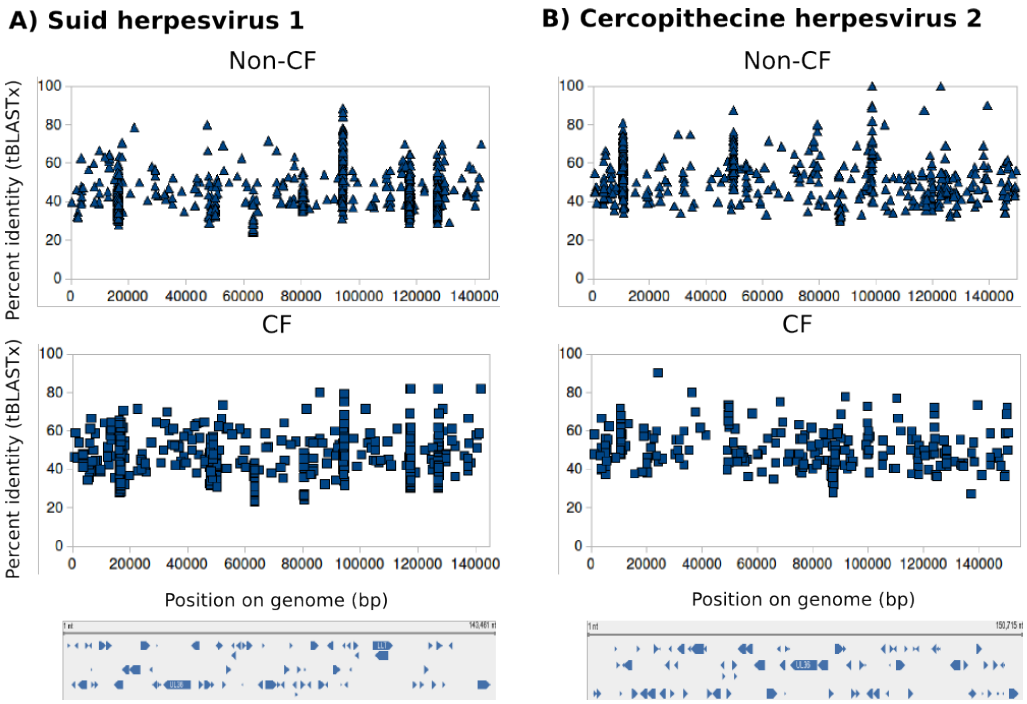


Figure S2. Combined coverage of Suid herpesvirus 1 and Cercopithecine herpesvirus 2 across all Non-CF and CF metagenomes as determined by tBLASTx. The graphics of the two reference genomes are from NCBI (http://www.ncbi.nlm.nih.gov).

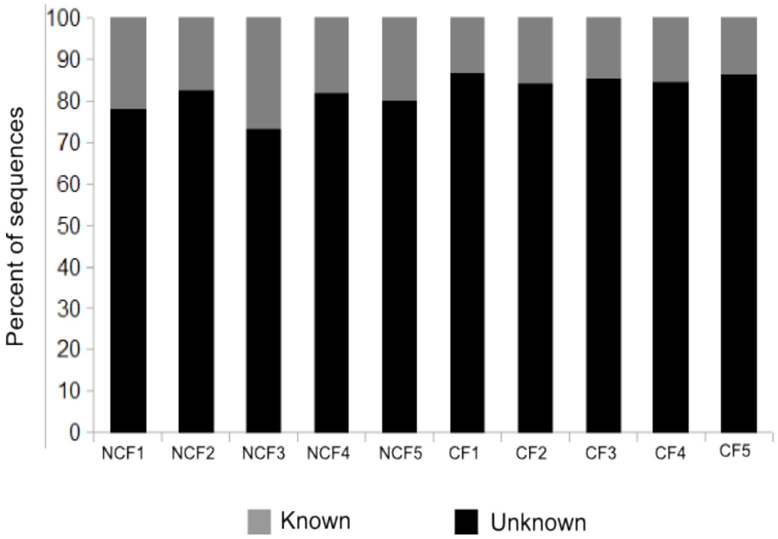


Figure S3. Percentage of metagenomic sequences with known and unknown metabolic functions as determined by BLASTx to the SEED database. A sequence was considered as known if it had a significant (e-value<10−5) hit to a gene in a metabolic pathway.

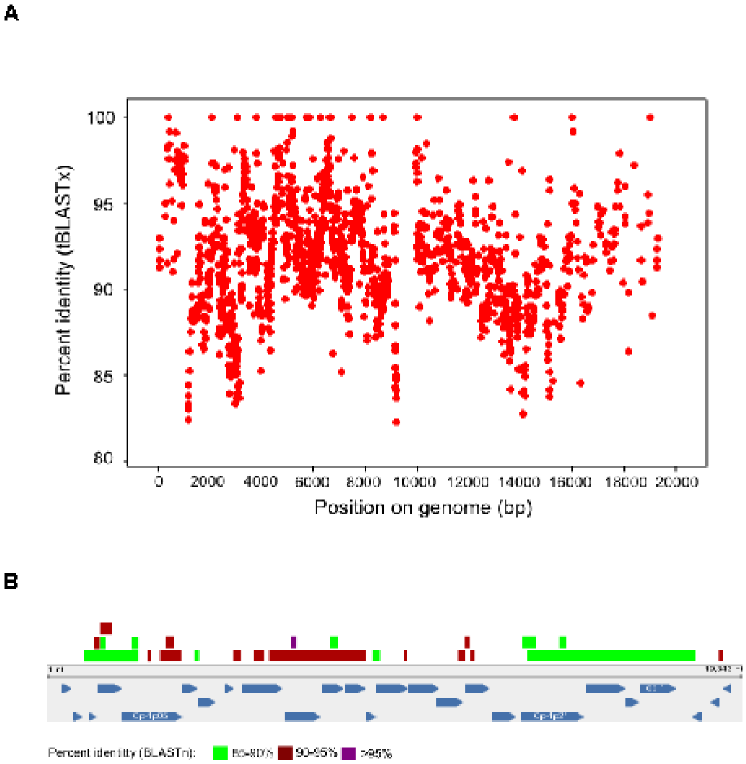


Figure S4. Coverage of the Streptococcus pneumonia phage Cp-1 genome in metagenome Non-CF2 by raw metagenomic sequences as determined by tBLASTx (A) and by assembled contigs as determined by BLASTn (B). The graphic of the 19,343 kb phage Cp-1 genome is from NCBI.

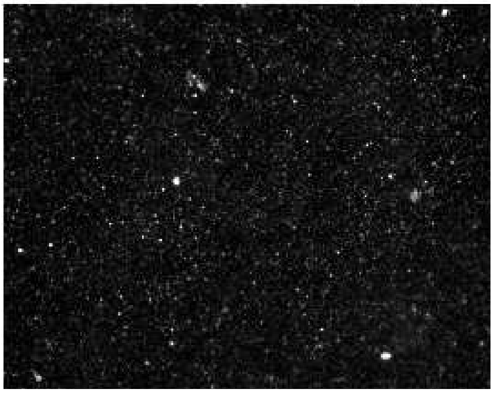


Figure S5. Virus-like Particles (VLPs) from the sputum sample of a CF patient. The VLPs were visualized by capture on a 0.02 µm Anodisc filter, SYBR Gold staining, and viewing under an epifluorescence microscope. The viruses appear as tiny bright pinpricks of light.

CHAPTER 5: COMPARATIVE METAGENOMICS IN ANIMAL DISEASE

Comparative metagenomic methods can also be used to compare viral communities healthy and diseased animals. Metagenomic analysis can provide insight into viruses which may be the cause of illness in morbid animals, as well as the shifts in viral communities that are the hallmark of the disease state. This chapter presents a case study of viral metagenomics in healthy and morbid hybrid striped bass from an aquaculture system.

Introduction

Aquaculture generates over 40% of the world's fish supply, and over 200 species of finfish and shellfish are currently farmed (1; 2). The industry is rapidly growing, and farmed fish are being raised at higher population densities, increasing the potential for the rapid spread of pathogens (1-3). Disease outbreaks caused by bacteria and viruses are recurrent in the intensive rearing of fish and can cause extensive production losses, and constrain further industrial development (3; 4). Conventional approaches to both the prevention and treatment of disease, such as the use of disinfectants or antibiotics, have had only limited success (3). Vaccines have been used for prophylaxis of some bacterial diseases, and DNA vaccines for a few viral diseases, including lympyhocystis disease virus, are currently under development (5; 6). Phage therapy and probiotics have been proposed as alternative treatments for some bacterial infections, however these methods have yet to be implemented in aquaculture systems (7-14).

Fish are continuously exposed to microbes and viruses present in water, sediments, equipment, and food, and on the surface of farming equipment (2). The microbial and viral communities of fish skin, gills, and intestinal tracts of are influenced by these environmental contacts. Many disease-causing organisms are ubiquitous in aquaculture and are benign under normal conditions, becoming pathogenic when the environment is disturbed (e.g. by overcrowding, changes in diet, or temperature fluctuations) and fish become stressed (15). Other pathogens constitutively cause disease, and will lead to immediate outbreaks when introduced into aquaculture systems (15).

Over 80% of farmed hybrid striped bass (HSB) is produced in the United States, and the largest producer is Kent Sea Tech in California (2). While HSB are generally considered to be resistant to most diseases, there are reports in the literature of both viral and bacterial infections in farmed bass including Mycobacteriosis, Vibriosis, Photobacteriosis, *Streptococcus iniae* infection and infectious pancreatic necrosis virus (2; 16-18). Farmed HSB are likely to harbor many other potential pathogens, as they are often exposed to wild bass species and water inflows (15). The cultured microbial flora associated with healthy HSB in both wild and aquaculture systems have been previously described, however, the corresponding viral communities remain largely uncharacterized (19-21). Determining the composition of viral communities associated with healthy fish, and understanding how these communities change under stress and disease conditions is essential to the development of new preventative and therapeutic regimes.

Methods and Materials

Sample preparation

In this study, a refined protocol consisting of homogenization followed by filtration and density-gradient centrifugation was used to isolate viruses from both fish mucus and intestines. Adult hybrid striped bass (*Morone chrysops* × *Morone saxatilis*) were collected in April 2006 from a 5x2 m open-air aquaculture pond located in the Kent Sea Tech Corporation aquaculture farm near the Salton Sea in California. Aquaculture farm veterinarians classified fish as healthy or morbid by visual inspection and dissection following sacrifice. The five morbid fish used presented with lesions at the surface of the skin and intestinal tracts free of faecal matter, while the five healthy fish had no lesions and normal gut contents. Mucus (slime) at the surface of the skin was aseptically scraped from the caudal fin to the gills of each fish with a sterile razorblade and resuspended into a buffered storage medium (SM). Gut contents were collected aseptically by flushing the intestines with SM buffer.

Samples were pooled by fish health status and body site (i.e. healthy gut, morbid gut, healthy mucus, morbid mucus) and homogenized by sonication. Samples were centrifuged at low speed for 10 minutes at 4°C. The supernatant was filtered (0.45 µm and 0.2 µm) to separate microbes (attached onto the filter) from the viruses (filtrate). The viral particles in the filtrate were purified using cesium chloride density-gradient centrifugation as described in (22). Viral DNA was extracted using formamide and cetyltrimethylammonium bromide extraction (23). Viral DNA was amplified using Phi29 polymerase to generate 10 µg of each sample for sequencing. The metagenomes used in this case study were sequenced using GS20 technology, and characteristics of the libraries are presented in Table 5.1.

Table 5.1: Characteristics of the fish-associated metagenomes.

	Raw number of sequences	Non-eukaryotic sequences	Percent non-eukaryotic	Average length (bp)	Percent known
Healthy Gut	47139	1628	3.45	111.8	96.58
Morbid Gut	53750	4065	7.56	114.8	89.2
Healthy Slime	61476	13923	22.65	98.5	89.2
Morbid Slime	60111	22312	37.12	98.3	82.11

Initial processing of metagenomic sequences

Dinucleotide relative abundance analysis was used to detect eukaryotic DNA contamination in all four fish metagenomes. Dinucleotide relative abundance odds ratios were used as inputs to principal component analysis, along with pre-computed odds ratios for a set of 45 viral metagenomes from a variety of environments as described in (24). In scatterplots of the first three principal components, the gut metagenomes clustered together away from the majority of the other viromes, while the mucus metagenomes clustered more centrally (Figure 5.1, left panel). The dinucleotide signatures of the gut metagenomes were most similar to those of two viromes derived from human sputum samples which were known to contain upwards of 90% human DNA (24). Both gut metagenomes had markedly low CG odds ratios (0.32 for Healthy Gut and 0.46 for Morbid Gut) which were comparable to those in the sputum samples. All metagenomes were then compared to the non-redundant database using BLASTn to specifically identify host (fish) and other eukaryotic sequences. After all contaminating eukaryotic sequences were removed, the mucus metagenomes clustered more tightly together, and the gut metagenomes showed an increase in CG odds ratios (1.05 and 0.67 respectively for healthy and morbid), and no longer clustered with the contaminated sputum metagenomes (Figure 5.1, right panel).

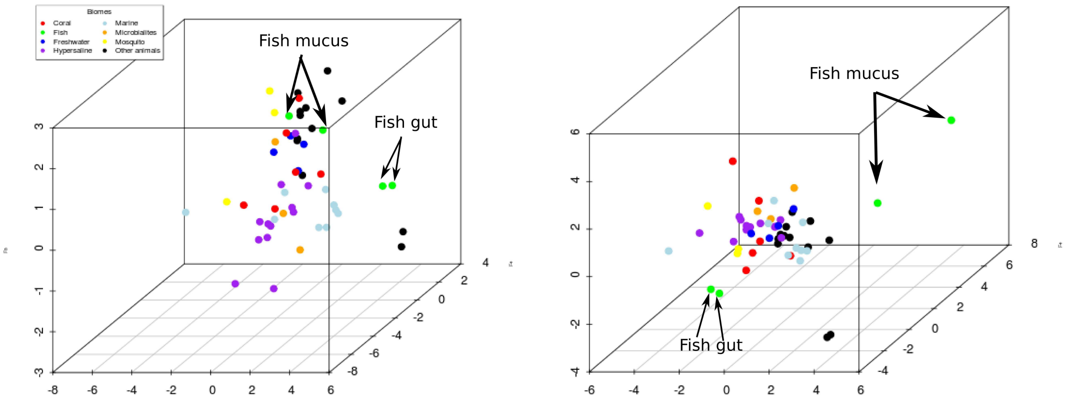


Figure 5.1 Three dimensional scatterplots of the first three prinicipal components from dinucleotide relative abundance analysis of viral metagenomes. Fish metagenomes are indicated in green. The left panel shows the fish metagenomes prior to removal of contaminating eukaryotic sequences, while the right panel shows the metagenomes following decontamination.

Bioinformatics

The fish-associated metagenomes were compared to two boutique databases (e-value < 10^{-5}) for taxonomic assignment (25). The first consisted of 3,074 completely sequenced eukaryotic viruses available from NCBI (http://www.ncbi.nih.nlm.gov), and the second contained 510 complete phage genomes (http://phage.sdsu.edu/phage). Metagenomic sequences from the aquaculture pond water at Kent Sea Tech, which was sampled on the same day as the healthy and morbid fish, were also compared to these databases. This pond metagenome is described in more detail in (26).

The GAAS metagenomic tool was used to determine the taxonomy of eukaryotic viral communities in fish skin mucus and the aquaculture pond (27). The two gut libraries had less than 10 tBLASTx similarities to eukaryotic viruses, and were thus excluded from further analyses. Phage communities in the fish-associated metagenomes were characterized by comparison to the Phage Proteomic Tree and visualization using Bio-Metamapper, and environmental distance between communities was determined using UniFRAC (28-30). Counts of best tBLASTx similarities were normalized to target genome size to avoid bias towards larger genomes as previously discussed. Functional annotations were assigned to sequences using the MG-RAST service, and metabolic profiles were compared using the non-parametric statistical tool XIPE (31; 32).

PHACCs was used to estimate viral diversity and community structure for the four fish-associated metagenomes using a logarithmic model (33). Average contig spectra were calculated from assemblies of 1000 randomly selected sequences with enough repetitions to achieve 5X coverage of each metagenome using Circonspect (http://biome.sdsu.edu/Circonspect). Average genome length for each metagenome was estimated using GAAS (27). The four fish-associated metagenomes were compared pairwise to each other using bi-directional BLASTn to identify the percentage of sequences shared.

Results

Eukaryotic viral communities

GAAS analysis indicated that papillomaviruses were the most abundant viral family in healthy mucus (Figure 5.2A). Papillomaviruses were present in a much lower abundance (3%) in morbid mucus and were not detected in the aquaculture pond water at all. Figure 2.2B shows where individual BLAST similarities matched to genomes of three human papillomaviruses. BLAST similarities were detected to 64 papillomaviruses overall. Healthy mucus similarities to papillomaviruses were concentrated in the E1, E2, and L2 genomic regions as well as in the upstream regulatory region (URR), which are conserved in all papillomaviruses (34). No similarities were observed in other genomic regions, which suggest the presence of a novel papillomavirus.

The eukrayotic viral community in morbid skin mucus was marked by poxviruses, iridoviruses, and herpesviruses. Poxviruses comprised almost 10% of the viral community in morbid mucus, but were present only in very low abundance in the aquaculture pond (<1%) and were undetectable in healthy mucus (Figure 5.2A). Iridoviruses and herpesviruses were more than twice as abundant in morbid mucus than in healthy mucus. All iridovirus similarities in the morbid mucus virome were to Infectious Spleen and Kidney Necrosis Virus (ISKNV), while there were no similarities to ISKNV in healthy mucus. Both iridoviruses detected in healthy mucus were also found in the aquaculture pond water, but ISKNV was not. Cyprinid herpesvirus 3 (Koi herpesvirus), which primarily infects carp, comprised the majority of herpesvirus similarities in morbid mucus (35). This virus was also present in healthy mucus, but at a much lower abundance, as were other herpesviruses. Analysis of eukaryotic viruses in the fish mucus metagenomes illustrates the two uses of viral metagenomics described above, as the results suggest that the etiologic agent in morbid fish may have been an iridovirus (viral discovery and diagnostics), and also that the viral community shifts in general in the disease state (microbial ecology).

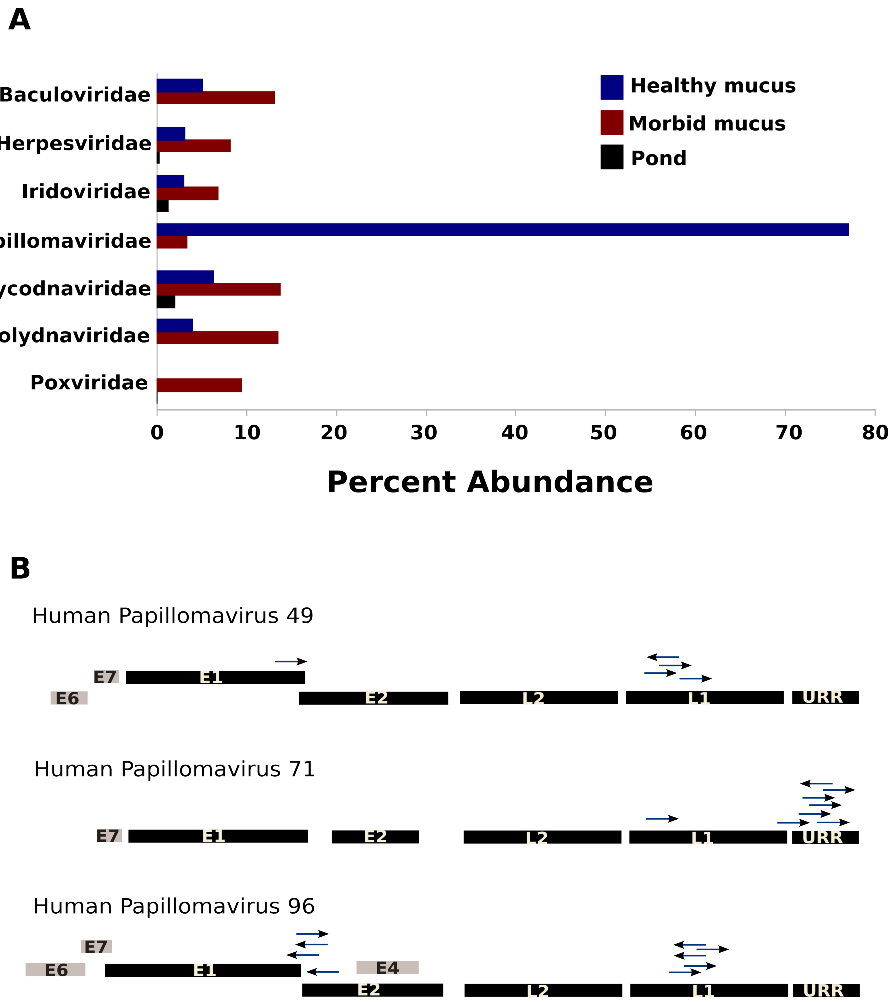


Figure 5.2 (A) Relative abundances of selected viral groups in fish mucus metagenomes as estimated by GAAS. GAAS estimates are based on tBLASTx similiarities (e-value < 0.00001, minimum percent similarity = 30%). (B) Coverage of papillomavirus genomes in healthy mucus metagenome. Arrows indicated metagenomic sequences with >90% similarity at the nucleotide level.

Phage communities

Phage communities were more characteristic of body site (gut versus skin mucus) than the disease state of fish. The two gut metagenomes had phage communities which were most similar to each other (Figure 5.3). Phage communities in the mucus metagenomes were most closely related to each other, and most distant from gut communities. Phage communities from the gut and the aquaculture pond were dominated by ssDNA microphage. Microphage have previously been found in high abundance in other aquatic systems, including the ocean and microbialites (36; 37). In addition, the healthy gut metagenome had similarities to four *Bacillus* phage and two *Salmonella* phage, while the morbid gut had none. This is indicative of feeding habits, as *Bacillus* are the bacteria most often associated with fish feed, and *Salmonella* species have been found to contaminate aquaculture feeds worldwide (Nedoluha and Westhoff, 1995; Nesse *et al.*, 2003)(20; 38).

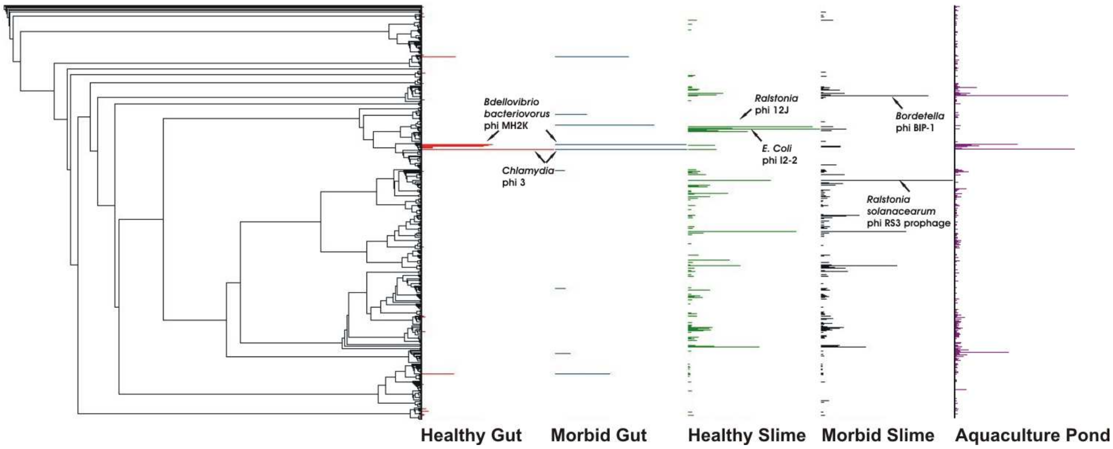


Figure 5.3 Phage proteomic tree displaying normalized relative abundances of similarities to phage genomes. Fish-associated and pond metagenomes were compared to a database of 510 fully sequenced phage genomes using tBLASTx (e-value < 10^{-5}).

There were many similarities to ssDNA microphage in the mucus metagenomes, however they were in lesser abundance than in the gut. Skin mucus phage communities were marked by high abundances of *Ralstonia* phage, including phage of *R. picketii* and prophage of *R. solanacearum* (Figure 5.3). *R. solanacearum* prophage were also highly abundant in the aquaculture pond, but were absent in gut metagenomes. The metagenomic approach rapidly delineated taxonomic differences in fish-associated phage communities. Other methods such as culturing and electron microscopy may have missed these differences, as they only allow for the study of a limited subset of the total community.

Metabolic profiles

There were no significant differences between the metabolic profiles of the fish mucus metagenomes, however, both differed significantly from the pond metagenome. The four fish viromes were compared to the SEED database using the MG-RAST service to determine their metabolic potential (39). MG-RAST assigns sequences to metabolic subsystems or pathways at three hierarchical level using BLASTX homologies (32; 39). Using an E-value cutoff of 10^{-5}, the two gut metagenomes both had less than 20 similarities to genes involved in metabolic pathways, and were excluded from further analysis. Figure 5.4 shows the percentage of sequences in each skin mucus metagenome assigned to each metabolic subsystem. As compared to the pond, the two fish mucus viromes were depleted in functions related to DNA metabolism and nucleotide/nucleoside metabolism. Mucus metagenomes were also significantly enriched in functions related to membrane transport, and specifically, in branched-chain amino acid and heavy metal transporters. These functions reflect the unique environment of fish skin mucus, which is enriched in amino acids and serves as a barrier to trap heavy metals before they can reach internal body sites (40; 41).

Diversity of fish-associated metagenomes

Skin mucus metagenomes were more diverse than gut metagenomes regardless of fish health status. The Shannon Index was higher for both mucus metagenomes than for gut metagenomes, indicating higher overall diversity. Specifically, the richness was

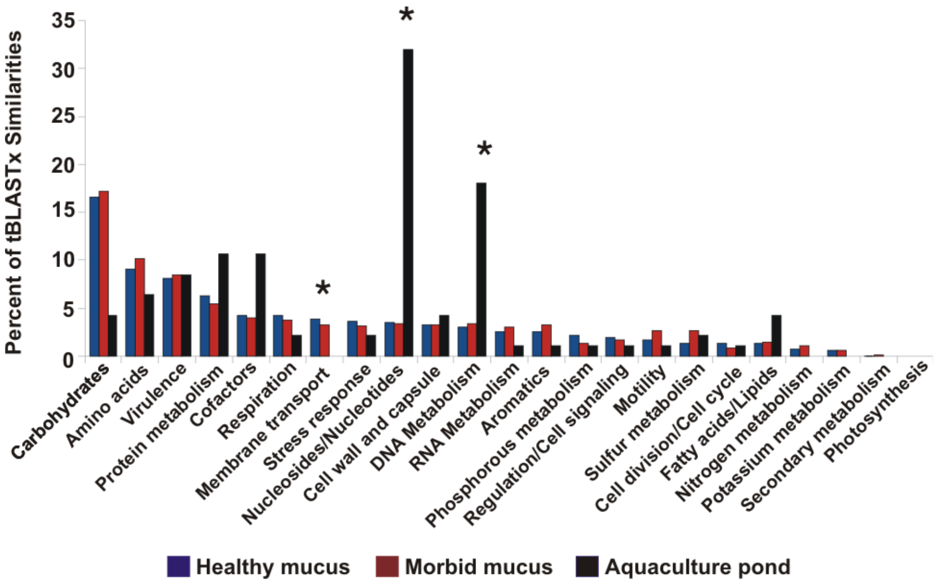


Figure 5.4 Metabolic profile of fish mucus and aquaculture pond metagenomes. Functional annotations were obtained by comparison to the SEED database using BLASTx (e-value < 10^{-5}). Asterisks indicate significant differences as determined by the non-parametric statistical program XIPE.

approximately an order of magnitude higher in mucus metagenomes, which were estimated to have 273 and 109 species for healthy and morbid respectively. The estimated number of viral species in mucus metagenomes was comparable to that in human sputum, while the richness in gut metagenomes (~30 species) was much lower than reported values for other environmental genomes (42).

Both gut metagenomes shared over 60% of sequences with each other, yet neither shared more than 12% of sequences with either mucus metagenome. Mucus metagenomes had fewer common sequences, with shared sequences comprising 19% of the morbid mucus library and 30% of the healthy mucus library. Less than 1% of sequences in the mucus metagenomes could be found in either gut metagenome. Metagenomes derived from the same environment consistently shared more sequences than those from environments with the same health status, suggesting that in general body site (gut vs. mucus) drives viral diversity.

Discussion

Metagenomic analysis of viruses associated with healthy and morbid hybrid striped bass revealed that eukaryotic viral communities were dramatically different in the mucus of healthy and morbid fish. In healthy slime, eukaryotic viral communities were dominated by papillomaviruses. Papillomaviruses have been identified as nearly ubiquitous commensals (i.e. at sub-clinical levels) on the healthy skin of humans and various other mammals (43; 44). Recently, two novel papillomaviruses were discovered in skin lesions of sea turtles, extending the known host range, as previously papillomaviruses had only been characterized in mammals and avians (45). The case study suggests that the host range may extend even further, and that there may be an as yet uncharacterized papillomavirus infecting fish. Additionally, since similarities to papillomaviruses were only found in the healthy mucus, this virus is likely to be a commensalist which is part of the normal skin flora, and may be replaced by more pathogenic organisms in diseased fish. Morbid skin mucus contained high abundances of poxviruses, herpesvirus, and the iridovirus ISKNV. ISKNV has previously been implicated in disease outbreaks in Chinese aquaculture, and has been shown experimentally to infect a variety of fish species including largemouth bass (46). Here, it is uncertain whether ISKNV or a related iridovirus was the etiologic agent of disease, however, these results suggest that further investigation is warranted (e.g., PCR). Previously, it was demonstrated the prevalence of herpesviruses in corals increases dramatically in response to stress, as it does in humans (47). Herpesviruses may be part of the normal flora of fish skin mucus, however, their abundance may increase in response to disease. These results suggested that the etiologic agent in the morbid HSB may have been viral, or that eukaryotic viral communities shifted as a whole in response to the disease state, and natural flora were superseded by opportunistic species.

While eukaryotic viral communities were indicative of the disease state, phage communities were characteristic of body sites. Phage communities from the gut consisted largely of ssDNA microphage. The prevalence of microphage in the gut of HSB most likely results from the continuous ingestion of pond water by fish, as intestinal microbiota are thought be a selected subset of those in water and food (48). Skin mucus phage communities were dominated by *Ralstonia* phage. *R. solanacearum* is a plant pathogen which is ubiquitous in aquatic environments including rivers, canals, and agricultural drainage waters (49-51). Additionally, *R. solanacearum* has been to shown to be viable in aquaculture conditions and filtered water, where is can persist for long periods of time in a viable but non-cultureable state (49; 52). *R. solanacearum* prophage in fish skin mucus were probably acquired from the aquaculture pond, where hosts are likely to be abundant. *R. picketii* is a human pathogen which often contaminates medical and laboratory solutions, and causes nosocomial infections (53). *R. picketii* has also been identified in the respiratory mucus of patients with cystic fibrosis, which suggests that it could also colonize fish mucus (53). Phage of *R. picketii* were found only in healthy fish skin mucus, indicating that these phage and their hosts may be part of the normal mucosal flora in aquacultured HSB.

Few differences were detected in phage communities derived from the same body site (gut vs. skin mucus) regardless of disease status. Similar results have been previously reported for the microbial flora of HSB. Bacterial communities on the skin of healthy HSB were found to be more like communities in the water column than intestinal flora (20). Skin mucus is the first line of defense against potential pathogens, and serves as a barrier between surrounding water and fish tissue including skin. Phages in mucus may represent a transient sampling of the exterior environment, rather than a selective community as in the gut. Our estimates of viral richness corroborate this idea, as skin mucus communities were estimated to have nearly ten times as many unique viral species as the gut. Additionally, the lack of sequence similarities between gut and mucus metagenomes from fish with the same health status suggests that the composition of fish-associated viral communities is driven by environmental differences (i.e. internal versus external) rather than disease. Nedoluha and Westhoff (1995) previously reported similar results for bacteria in farmed HSB. Bacterial richness was much lower in the gut than on the skin and gills, suggesting that the intestinal environment is more selective than external body surfaces (19). This is supported by the work of Sugita et al. (1996), which showed that the natural intestinal flora of cultured fish produce antimicrobial compounds to prevent colonization by ingested microbes (54).

Viral metabolic functions in skin mucus reflected the unique nature of the mucosal surface, but were indistinguishable between healthy and morbid fish. Metabolic profiles of phage communities in healthy and morbid skin mucus were distinctly different from that of the aquaculture pond. These results correspond to the previous work of Dinsdale et al. (2008a), which demonstrated that phage communities in different environments have distinct and characteristic metabolic profiles. Metabolic profiles in fish skin mucus were specifically enriched in branched-chain amino acid and heavy metal transporters, reflecting the unique microenvironment of the mucus layer. Fish skin mucus has biochemical and physical properties similar to mucus in other animals, including humans, and plays a critical role in innate immunity (40; 41). Fish mucus contains many antimicrobial agents including lysozyme, proteases, complement proteins,and antibacterial peptides which can destroy microbes (55). Upon lysis, microbial cell contents are released into mucus, which can be used by advantageous microbes as carbon sources. Pathogenic strains of *Vibrio* isolated from diseased fish were shown to have the ability to use mucus as a sole carbon source (56). *Pseudomonas aeruginosa* in the respiratory mucus of human cystic fibrosis patients utilize carbon sources in mucus for growth, especially branched chain and aromatic amino acids (57). Microbes in fish mucus may be specifically using branched chain amino acids and small peptides as carbon sources, increasing the need for transport proteins. Since metabolic functions encoded by phage mirror those of their microbial hosts, DNA sequence corresponding to these transporters appears in viral metagenomes (26). In both mucus metagenomes, there were also many similarities to transporters of heavy metals, including zinc, nickel, manganese, and molybdenum (Figure 5.4). Mucus secretion increases when fish are exposed to metals, which is thought to be a defense mechanism, since metal ions will be trapped in mucus and sloughed off before they can reach interior tissues (41; 58). This may lead to a high concentration of heavy metal ions in mucus, and microbes would need efficient efflux systems, including membrane transporters, in order to survive.

Gut viromes were excluded from metabolic profile analysis due to a dearth of similarities to target database. Isolated viral fractions from CsCl density gradient centrifugation were not DNAse treated prior to viral DNA extraction, resulting in a significant number of contaminating eukaryotic DNA sequences in the metagenomes. Gut viromes were especially sensitive to contamination by host DNA and there were a relatively small number of reads remaining in these viromes once contaminating sequences were removed. The addition of a DNAse step would have allowed for more viral sequences in all metagenomes, and thus a larger total information gain during data analysis steps.

References

1. Naylor RL, Goldburg RJ, Primavera JH, Kautsky N, Beveridge MC, Clay J, Folke C, Lubchenco J, Mooney H, Troell M. Effect of aquaculture on world fish supplies. Nature 2000 Jun;405(6790):1017-1024.

2. USDA ,APHIS, Veterinary Service. Assessing infectious disease emergence potential in the U.S. aquaculture industry: phase 1, U.S. aquaculture industry profile. 2007;

3. Bondad-Reantaso MG, Subasinghe RP, Arthur JR, Ogawa K, Chinabut S, Adlard R, Tan Z, Shariff M. Disease and health management in Asian aquaculture. Vet. Parasitol 2005 Sep;132(3-4):249-272.

4. Murray AG, Peeler EJ. A framework for understanding the potential for emerging diseases in aquaculture. Prev. Vet. Med 2005 Feb;67(2-3):223-235.

5. Biering E, Villoing S, Sommerset I, Christie KE. Update on viral vaccines for fish. Dev Biol (Basel) 2005;121:97-113.

6. Lorenzen N, LaPatra SE. DNA vaccines for aquacultured fish. Rev. - Off. Int. Epizoot 2005 Apr;24(1):201-213.

7. Balcázar JL, de Blas I, Ruiz-Zarzuela I, Cunningham D, Vendrell D, Múzquiz JL.

The role of probiotics in aquaculture. Vet. Microbiol 2006 May;114(3-4):173-186.

8. Brunt J, Austin B. Use of a probiotic to control lactococcosis and streptococcosis in rainbow trout, Oncorhynchus mykiss (Walbaum). J. Fish Dis 2005 Dec;28(12):693-701.

9. Das S, Ward LR, Burke C. Prospects of using marine actinobacteria as probiotics in aquaculture. Appl. Microbiol. Biotechnol 2008 Dec;81(3):419-429.

10. Irianto A, Austin B. Use of dead probiotic cells to control furunculosis in rainbow trout, Oncorhynchus mykiss (Walbaum). J. Fish Dis 2003 Jan;(1):59-62.

11. Nakai T, Sugimoto R, Park KH, Matsuoka S, Mori K, Nishioka T, Maruyama K. Protective effects of bacteriophage on experimental Lactococcus garvieae infection in yellowtail. Dis. Aquat. Org 1999 Jun;37(1):33-41.

12. Park SC, Nakai T. Bacteriophage control of Pseudomonas plecoglossicida infection in ayu Plecoglossus altivelis. Dis. Aquat. Org 2003 Jan;53(1):33-39.

13. Park SC, Shimamura I, Fukunaga M, Mori KI, Nakai T. Isolation of bacteriophages specific to a fish pathogen, Pseudomonas plecoglossicida, as a candidate for disease control. Appl. Environ. Microbiol 2000 Apr;66(4):1416-1422.

14. Pieters N, Brunt J, Austin B, Lyndon AR. Efficacy of in-feed probiotics against Aeromonas bestiarum and Ichthyophthirius multifiliis skin infections in rainbow trout (Oncorhynchus mykiss, Walbaum). J. Appl. Microbiol 2008 Sep;105(3):723-732.

15. Meyer FP. Aquaculture disease and health management. J. Anim. Sci 1991 Oct;69(10):4201-4208.

16. Wechsler SJ, McAllister PE, Hetrick FM. Neutralizing activity against infectious pancreatic necrosis virus in striped bass, Morone saxatilis, from the Chesapeake Bay. J. Wildl. Dis 1987 Jan;23(1):154-155.

17. Ostland VE, Stannard JA, Creek JJ, Hedrick RP, Ferguson HW, Carlberg JM, Westerman ME. Aquatic Francisella-like bacterium associated with mortality of intensively cultured hybrid striped bass Morone chrysops x M. saxatilis. Dis. Aquat. Org 2006 Oct;72(2):135-145.

18. Bowser PR, Wooster GA, Chen C, Mo RS. Polymicrobic infection of hybrid striped bass (Morone chrysops x Morone saxatilis) with three bacterial pathogens: a case report. J. Fish Dis 2004 Feb;27(2):123-127.

19. Nedoluha PC, Westhoff D[. Microbiological Analysis of Striped Bass (Marone saxatilis) Grown in Flow-Through Tanks. Journal of Food Protection 1995 Dec;58:1363-1368.

20. Nedoluha PC, Westhoff D. Microbiology of striped bass grown in three aquaculture systems. Food Microbiology 1997 Jun;14(3):255-264.

21. Nedoluha PC, Westhoff D[. Microbiological Analysis of Striped Bass (Marone saxatilis) Grown in a Recirculating System. Journal of Food Protection 1997 Aug;60:948-953.

22. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc 2009;4(4):470-483.

23. Sambrook J. Molecular Cloning: A Laboratory Manual, Third Edition. 3rd ed. Cold Spring Harbor Laboratory Press; 2001.

24. Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes [Internet]. Environ. Microbiol 2009 Mar;[cited 2010 Mar 29] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19302541

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990 Oct;215(3):403-410.

26. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. Functional metagenomic profiling of nine biomes. Nature 2008 Apr;452(7187):629-632.

27. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol 2009 Dec;5(12):e1000593.

28. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol 2005 Dec;71(12):8228-8235.

29. Lozupone C, Hamady M, Knight R. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 2006;7:371.

30. Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol 2002 Aug;184(16):4529-4535.

31. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative metagenomics. BMC Bioinformatics 2006;7:162.

32. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33(17):5691-5702.

33. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 2005;6:41.

34. García-Vallvé S, Alonso A, Bravo IG. Papillomaviruses: different genes have different histories. Trends Microbiol 2005 Nov;13(11):514-521.

35. Waltzek TB, Kelley GO, Stone DM, Way K, Hanson L, Fukuda H, Hirono I, Aoki T, Davison AJ, Hedrick RP. Koi herpesvirus represents a third cyprinid herpesvirus (CyHV-3) in the family Herpesviridae. J. Gen. Virol 2005 Jun;86(Pt 6):1659-1667.

36. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 2008 Mar;452(7185):340-343.

37. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The Marine Viromes of Four Oceanic Regions. PLoS Biol 2006 Nov;4(11):e368.

38. Nesse LL, Nordby K, Heir E, Bergsjoe B, Vardund T, Nygaard H, Holstad G. Molecular analyses of Salmonella enterica isolates from fish feed factories and fish feed ingredients. Appl. Environ. Microbiol 2003 Feb;69(2):1075-1081.

39. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional

analysis of metagenomes. BMC Bioinformatics 2008;9:386.

40. Bols NC, Brubacher JL, Ganassin RC, Lee LE. Ecotoxicology and innate immunity in fish. Dev. Comp. Immunol 2001 Dec;25(8-9):853-873.

41. Shephard KL. Functions for fish mucus. Reviews in Fish Biology and Fisheries 1994 Dec;4(4):401-429.

42. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS ONE 2009 Oct;4(10):e7370.

43. Antonsson A, Hansson BG. Healthy skin of many animal species harbors papillomaviruses which are closely related to their human counterparts. J. Virol 2002 Dec;76(24):12537-12542.

44. Antonsson A, Forslund O, Ekberg H, Sterner G, Hansson BG. The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. J. Virol 2000 Dec;74(24):11636-11641.

45. Herbst LH, Lenz J, Van Doorslaer K, Chen Z, Stacy BA, Wellehan JFX, Manire CA, Burk RD. Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. Virology 2009 Jan;383(1):131-135.

46. He JG, Deng M, Weng SP, Li Z, Zhou SY, Long QX, Wang XZ, Chan SM. Complete genome analysis of the mandarin fish infectious spleen and kidney necrosis iridovirus. Virology 2001 Dec;291(1):126-139.

47. Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C, Edwards RA, Haynes M, Angly FE, Wegley L, Rohwer FL. Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proc. Natl. Acad. Sci. U.S.A 2008 Nov;105(47):18413-18418.

48. Cahill MM. Bacterial flora of fishes: A review. Microbial Ecology 1990 Jan;19(1):21-41.

49. Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF. Analysis of bacteria contaminating ultrapure water in industrial systems. Appl. Environ. Microbiol 2002 Apr;68(4):1548-1555.

50. Tomlinson D, Elphinstone J, Soliman M, Hanafy M, Shoala T, Abd El-Fatah H, Agag S, Kamal M, Abd El-Aliem M, Fawzi F, Stead D, Janse J. Recovery of Ralstonia solanacearum from canal water in traditional potato-growing areas of

Egypt but not from designated Pest-Free Areas (PFAs). European Journal of Plant Pathology 2009 Dec;125(4):589-601.

51. van Elsas JD, Kastelein P, de Vries PM, van Overbeek LS. Effects of ecological factors on the survival and physiology of Ralstonia solanacearum bv. 2 in irrigation water. Can. J. Microbiol 2001 Sep;47(9):842-854.

52. Alvarez B, Lopez MM, Biosca EG. Survival strategies and pathogenicity of Ralstonia solanacearum phylotype II subjected to prolonged starvation in environmental water microcosms. Microbiology 2008 Nov;154(11):3590-3598.

53. Coenye T, Vandamme P, LiPuma JJ. Infection by Ralstonia species in cystic fibrosis patients: identification of R. pickettii and R. mannitolilytica by polymerase chain reaction. Emerging Infect. Dis 2002 Jul;8(7):692-696.

54. Sugita H, Shibuya K, Shimooka H, Deguchi Y. Antibacterial abilities of intestinal bacteria in freshwater cultured fish. Aquaculture 1996 Oct;145(1-4):195-203.

55. Ellis AE. Innate host defense mechanisms of fish against viruses and bacteria. Dev. Comp. Immunol 2001 Dec;25(8-9):827-839.

56. Bordas MA, Balebona MC, Zorrilla I, Borrego JJ, Moriñigo MA. Kinetics of adhesion of selected fish-pathogenic Vibrio strains of skin mucus of gilt-head sea bream (Sparus aurata L.). Appl. Environ. Microbiol 1996 Oct;62(10):3650-3654.

57. Palmer KL, Mashburn LM, Singh PK, Whiteley M. Cystic Fibrosis Sputum Supports Growth and Cues Key Aspects of Pseudomonas aeruginosa Physiology. J. Bacteriol. 2005 Aug;187(15):5267-5277.

58. Coello WF, Khan MA. Protection against heavy metal toxicity by mucus and scales in fish. Arch. Environ. Contam. Toxicol 1996 Mar;30(3):319-326.

Acknowledgements

Chapter 5, in full, has been accepted for publication as part of the chapter entitled: Viral metagenomics: from fish slime to the world. Willner, D, Desnues, C, and Rohwer, F. Chapter in “Metagenomics and its applications in agriculture, biomedicine, and environmental studies,” Nova Scientific Publishers. The dissertation author was the primary investigator and author of this material.

CHAPTER 6: CONCLUSIONS

Viral metagenomics provides a methodology to characterize viral ecology in host-associated systems, and also to explore how the disease state changes viral communities in both humans and animals. This dissertation has provided a review of methods to generate and analyze viral metagenomes, as well as three experimental studies employing these methods. Chapter 6 summarizes the results presented in the dissertation and offers suggestions for future work in host-associated viral metagenomics.

Contamination in host-associated metagenomes

Separation of viruses and viral DNA from host genomic DNA is a special consideration in host-associated metagenomics. While molecular methods such as DNAse treatment and density gradient centrifugation can reduce the levels of contaminating DNA prior to sequencing, bioinformatic methods are generally required to filter out remaining host sequences (1, 2). The metagenomic signatures method described in Chapter 2 provides a rapid screen to determine if metagenomes are contaminated with eukaryotic DNA . BLAST-based searches for contaminating sequences can be time consuming and computationally intensive, so this pre-screening is useful to prevent unnecessary computation. The dinucleotide signature method has been incorporated in a freely available R package (http://sourceforge.net/projects/dinucleotidesig) which allows users to evaluate their own metagenomes. The R package has also been included as part of a metagenome pre-processing pipeline for quality control (Persephone) currently used at San Diego State University.

Tapping into the undiscovered

Viral metagenomics provides a method to assess previously unknown viral diversity, and can reveal both novel and unexpected viruses. Traditional methods to characterize viral communities are often biased, relying on culturing and/or a priori knowledge of what viruses may be present to generate PCR-based and other probes. Other methods such as Pulsed-Field Gel Electrophoresis (PFGE) are time-consuming and relatively uninformative, providing no information about taxonomy or metabolic functions. Studies of oropharyngeal viruses using these methods have led to the conclusion that the oropharynx is not heavily populated or impacted by phage (3, 4). However, the results presented in Chapter 3 of this dissertation imply that the advent of a more thorough and exhaustive technique (i.e. metagenomics) was necessary to examine phage in the oral cavity. Using viral metagenomics, phage-encoded platelet-binding factors were discovered in the healthy human oropharynx, a completely unexpected result. PblA and pblB genes were previously found in virulent strains of S. mitis isolated from the blood of endocarditis patients, but never in the oral cavity, and never in healthy individuals (5-7). There are many other reports in the literature which describe the success of metagenomics where other methods have failed. For example, Palacios et al. determined that a novel arenavirus was the cause of mortality in three transplant recipients using metagenomics after culturing, PCR, and microarray analyses were inconclusive (8).

Comparative viral metagenomics in health and disease

Comparative studies of viral communities in healthy and diseased individuals can shed light on the unique nature of the disease state. In the CF study presented in Chapter 4, phage in CF and Non-CF airways had unique metabolic profiles regardless of which taxa were present. The CF phage metabolic profile reflected host adaptations to the unique environment of the CF airway. The vast repertoire of functional genes shared by CF phage in all five individuals suggested that in addition to reflecting host functional profiles, phage may be the driving force behind microbial diversification and the expansion of host metabolisms. In contrast, viral metabolic profiles in healthy and morbid fish mucus (Chapter 5) were not significantly different, but taxonomic profiles of eukaryotic viruses changed dramatically in the disease state. Additionally, the composition of phage communities reflected the body site of the fish from which the sample was taken, and seemed to be uncorrelated with health status. Without non-diseased groups for comparison, the unique characteristics of the CF virome and the fish-associated viral communities would not have been apparent. A comparative metagenomic study of oropharyngeal viruses in individuals with and without endocarditis (as a follow-up to Chapter 3) could provide similar insight into how phage-encoded virulence factors in the oral cavity influence the disease.

Future directions

The viral metagenomic studies presented in this dissertation provide a starting point for a wealth of future studies, both metagenomic and otherwise. Metagenomic studies such as those presented in Chapters 3, 4, and 5 have traditionally been limited by the cost and time required for sequencing, as well as computational power for data analysis. Rapid innovations in sequencing techonologies and the development of bioinformatic tools and and platforms have made larger-scale metagenomic studies feasible. With current 454 technology, one sequencing plate generates on average 800,000 sequences with an average length of 450 base pairs. Current 454 technology includes multiplexing of samples via the attachment of unique oligonucleotide tags to sequences in each sample group (9, 10). This means that many samples can be sequenced in parallel in one sequencing run on one plate, which dramatically reduces sequencing costs. Parallel implementations of BLAST searches (e.g. MPI-BLAST) allow for bioinformatic analysis of large samples in reasonbale time frames (11). New, more rapid methods and algorithms for homology searches including Hiden Markov Models and k-mer word searches have also recently been developed (e.g. (12, 13) and http://edwards.sdsu.edu/rtmg). With these innovations, new studies to explore human and associated viruses are now possible, such as time series and case-control and cohort studies. Pooled studies are also no longer be required to analyze larger sample groups such the 19 individuals in the oropharyngeal virus study. As an extension of the study comparing the airway virome of CF and Non-CF individuals, a time series examining viral communities in CF individuals prior to exacerbation, during exacerbation and following antibiotic treatment has been initiatied. In additon, metagenomic analysis of viruses in CF lung tissue by lobe in two individuals is currently being conducted, aided by the use of multiplexed sequencing reads.

As discussed in Chapter 1, findings of metagenomic studies results must be confirmed, clarified and/or extended by other methods. In this way, metagenomics is a starting point for a whole host of other experiments and analyses. For example, the study of oropharyngeal viruses indicated the presence of pblA and pblB genes in the oral cavity. Whether the pblA and pblB proteins can effectively bind oral microbes is yet to be determined. The results of the CF virome study indicate the importance of phage-encoded functional genes in microbial adaptations to the CF airway. However, the nature of phage-host interactions remains largely unknown. Culture-based and microscopy studies are currently underway to assess the dynamics of phage infections in the CF lung. In the fish mucus viromes (Chapter 5), metagenomic data showed significant similarities to conserved regions of the papillomavirus genome, suggesting the presence of a novel papillomavirus in the skin mucus of healthy fish. To truly validate the presence of this virus, a PCR-based assay would be necessary. Future studies such as these will show the true power of viral metagenomics as an agent for discovery and scientific exploration.

References

1. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc. Natl. Acad. Sci. U.S.A* 98:11609-11614.

2. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470-483.

3. Bachrach G, Leizerovici-Zigmond M, Zlotkin A, Naor R, Steinberg D (2003) Bacteriophage isolation from human saliva. *Letters in Applied Microbiology* 36:50-53.

4. Hitch G, Pratten J, Taylor P (2004) Isolation of bacteriophages from the oral cavity. *Letters in Applied Microbiology* 39:215-219.

5. Mitchell J, Sullam PM (2009) Streptococcus mitis phage-encoded adhesins mediate attachment to {alpha}2-8-linked sialic acid residues on platelet membrane

gangliosides. *Infect. Immun* 77:3485-3490.

6. Siboo IR, Bensing BA, Sullam PM (2003) Genomic organization and molecular characterization of SM1, a temperate bacteriophage of Streptococcus mitis. *J. Bacteriol* 185:6968-6975.

7. Bensing BA, Rubens CE, Sullam PM (2001) Genetic loci of Streptococcus mitis that mediate binding to human platelets. *Infect. Immun* 69:1373-1380.

8. Palacios G et al. (2008) A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. *N Engl J Med* 358:991-998.

9. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* 5:235-237.

10. Andersson AF et al. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3:e2836.

11. Archuleta J, Feng W, Tilevich E (2007) A pluggable framework for parallel pairwise sequence search. *Conf Proc IEEE Eng Med Biol Soc* 2007:127-130.

12. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23:205-211.

13. Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS (2009) Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics* 10:316.