UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO STATE UNIVERSITY

EXPLORING THE GLOBAL VIROME AND DECIPHERING THE ROLE OF PHAGES IN CYSTIC FIBROSIS

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

in

Biology

by

Ana Georgina Cobián Güemes

Committee in charge:

University of California San Diego

Professor Douglass Conrad
Professor Justin Meyer
Professor Joseph Pogliano

San Diego State University

Professor Forest Rohwer, Chair
Professor Robert Edwards

2019

The Dissertation of Ana Georgina Cobián Güemes 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

2019

Dedication

To Adrian, Pilar and Jorge for always supporting me through this unique journey.

Epigraph

“No temas a la perfección, nunca la alcanzarás”

Salvador Dali

Table of Contents

List of Figures

List of Tables

List of Equations

Acknowledgments

Thanks to my mentor Dr. Forest Rohwer for giving me the opportunity to explore the amazing world of phages. Thanks for teaching me how to ask the really interesting questions and work really hard to answer them. Thanks for pushing me to the limits of knowledge.

Thanks to Dr. Rob Edwards who encouraged me to come to San Diego and has always being very generous with this advice, computers and time. Thanks to Dr. Doug Conrad for caring about the microbes and sharing his clinical knowledge and resources to get us one step closer to apply genomics in the clinic. Thanks to Dr. Justin Meyer for teaching a bioinformatician to do phage plaque essays and guiding me through the phage mutations space. Thanks to Dr. Joe Pogliano and Dr. Elio Schaechter for the microbiology class and providing sharp and original questions for my research.

Dr. Anca Segall is an inspiration and I thank her for guiding me through my phage experiments and allowing me to work in her lab. (I hope one day I can identify an integrase by eye, like she does). Dr Linda Wegley-Kelley, I couldn't have published this dissertation witouth your guidance and support, thank you.

The Rohwer lab Cystic Fibrosis team was an inspiration and I thank Dr. Yan Wei Lim, Mike Furlan, Dr. Rob Quinn and Dr. Katrine Whiteson for welcoming me and showing me very usefull methods and tools. Thanks to Dr. Cynthia Silveira for bringing her viral world knowledge to the CF rapid response team and making the "not so rapid response" weekends enjoyable. Thanks to Gina Spidel and Eugenia Kronen for their hard work in getting all the

reagents and permits that we need to do science; and to Patti Swinford and Medora Bratlien for helping with the PhD related matters.

Thanks to my Rohwer lab comrades that made this five years more fun: Yan Wei Lim, Lauren Paul, Savannah Sanchez, Emma George, Adam Barno, Kevin Green, Sean Benler, Ty Roach, Mark Little, Marisa Rojas, Brandon Reyes, Ines Glatier D'Auriac, Lance Boling, Jose Reina, Esther Rubio, Aaron Hartman, Helena Villela, Doug Naliboff, Brandie White, Jason Baer and Zach Quilan. Thanks to the wise advice from Dr. Linda Wegley-Kelley and Dr. Juris Grasis that helped me navigate grad school.

Thanks to the Edwards lab for bioinformatics help and friendship: Kate McNair, Daniel Cuevas and Adrian Cantu. Thanks to the awesome Heather Maughman and Merry Youle for helping me improve my writing skills. Thanks to Dr. Conrad's team at the UCSD CF clinic for making it possible to have samples to work with: Jenna Mielke, Nelly and Rohaum. Thanks to the phage wisperers Shr-Hau Hung and Greg Peters for making any phage related protocol work. Thanks to the phage hunters team: Tram Le, Jessica Octavio, Lorena Dominguez, Helena Villela, Lili Han.

Thanks to Dr. Liz Dinsdale for her advice, and support in getting our sequencing runs to work. Thanks to the everyone in the "Phage Journal club", the "Bioinformatics Breakfast" and the "Biomath Group" for their delightfull scientific conversations.

Adrian Cantu, thanks for all your programming support, science discussions, love, care and patience. Thanks for coming with me to this journey and making my life amazing.

Mom and dad, thanks for giving me all the tools to achieve this goal. Thanks for you love and support. Thanks to all my family for showing their support for my PhD. Thanks for calling, visiting and making sure everything was OK. A special thanks to my grandpa for encouraging me to go abroad for the PhD and for asking the question of "Who counted them?", in this case I did counted the phages. Thanks to our friends in San Diego, specially to Peter and Sally Salamon.

Thanks to the funding agencies that supported me during my PhD: CONACyT, UCMEXUS and the Spruance Foundation.

Chapter 1, in full, is published in Annual Review of Virology. Ana Georgina Cobián Güemes, Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton and Forest Rohwer; 2016. The dissertation author was the primiary investigator and author of this paper.

Chapter 2, in part, is in preparation for submission. Ana Georgina Cobián Güemes, Vito Adrian Cantú, Jose Carlos Reina Cabello and Forest Rohwer. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is published in mBIO. Ana Georgina Cobián Güemes, Yan Wei Lim, Robert A. Quinn, Douglas J. Conrad, Sean Benler, Heather Maughan, Rob Edwards, Thomas Brerrin, Vito Adrian Cantú, Daniel Cuevas, Rohaum Hamidi, Pieter Dorrestein and Forest Rohwer; 2019. The dissertation author was the primiary investigator and author of this paper.

Chapter 4 is in preparation for submission. Ana Georgina Cobián Güemes, Cynthia B. Silveira, Mark Little, Vito Adrian Cantú, Sean Benler, Jose Carlos Reina Cabello, Rob

Edwards, Douglas Conrad and Forest Rohwer; 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is in preparation for submission. Ana Georgina Cobián Güemes, Tram Le, Maria Isabel Rojas, Lorena Dominguez, Jessica Claire Octavio, Lance Boling, Helena Villela, Shr-Hau Hung, Lili Han, Vito Adrian Cantú, Rob Edwards, Anca Segal, Douglas Conrad and Forest Rohwer; 2019. The dissertation author was the primary investigator and author of this paper.

VITA

2019	Doctor of Phylosophy in Biology, University of California San Diego and San Diego State University
2013 – 2014	Research Assistant in Bioinformatics, San Diego State University, CA, USA
2012 – 2013	Bioinformatics analyst, Research Center for Food and Development (CIAD), Sonora, Mexico
2012	Master of Sciences in Biochemistry, National Autonomous University of Mexico
2010	Bachelor of Science in Genomics, National Autonomous University of Mexico

PUBLICATIONS

Cobián Güemes, A. G., Vito Adrian Cantú, Jose Carlos Reina Cabello and Forest Rohwer "Fragment Recreuitment Assembly Purification" (in preparation)

Cobián Güemes, A. G., Forest Rohwer, Rob Cole, Sandra Morales, Saima Aslam and Susan M. Lehman "Population dynamics of a phage therapy product during treatment of a *Pseudomonas aeruginosa* lung infection" (in preparation)

Cobián Güemes, A. G., Tram Le, Maria Isabel Rojas, Lorena Dominguez, Jessica Claire Octavio, Lance Boling, Helena Villela, Shr-Hau Hung, Lili Han, Vito Adrian Cantú, Rob Edwards, Anca Segal, Douglas Conrad and Forest Rohwer "Compounding Achromophages for therapeutic applications." (in preparation)

Cobián Güemes, A. G., Cynthia B. Silveira, Jose Carlos Reina Cabello, Mark Little, Adrian Cantú, Sean Benler, Rob Edwards, Douglas Conrad and Forest Rohwer. "Mobile genetic elements in Cystic Fibrosis" (in preparation)

Cobián Güemes, A. G., Yan Wei Lim, Robert A. Quinn, Douglas J. Conrad, Sean Benler, Heather Maughan, Rob Edwards, Thomas Brettin, Vito Adrian Cantú, Daniel Cuevas, Rohaum Hamidi, Pieter Dorrestein, and Forest Rohwer. "Cystic Fibrosis Rapid Response: Translating Multi-omics Data into Clinically Relevant Information." *MBio* 10, no. 2 (2019).

Cobián Güemes, A. G., Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton, and Forest Rohwer. "Viruses as Winners in the Game of Life." *Annual Review of Virology* 3, no. 1 (2016).

Benler, Sean, **Cobián Güemes, A. G.**, Katelyn Mcnair, Shr-Hau Hung, Kyle Levi, Rob Edwards, and Forest Rohwer. "A Diversity-generating Retroelement Encoded by a Globally Ubiquitous Bacteroides Phage." *Microbiome* 6, no. 1 (2018).

Galtier D'Auriac, Ines, Robert A. Quinn, Heather Maughan, Louis-Felix Nothias, Mark Little, Clifford A. Kapono, **Cobián Güemes, A. G.**, Brandon T. Reyes, Kevin Green, Steven D. Quistad, Matthieu Leray, Jennifer E. Smith, Pieter C. Dorrestein, Forest Rohwer, Dimitri D. Deheyn, and Aaron C. Hartmann. "Before Platelets: The Production of Platelet-activating Factor during Growth and Stress in a Basal Marine Organism." *Proceedings of the Royal Society B: Biological Sciences* 285, no. 1884 (2018).

Knowles, Ben, Barbara Bailey, Lance Boling, Mya Breitbart, **Cobián Güemes, A. G.**, Javier Del Campo, Rob Edwards, Ben Felts, Juris Grasis, Andreas F. Haas, Parag Katira, Linda Wegley Kelly, Antoni Luque, Jim Nulton, Lauren Paul, Gregory Peters, Nate Robinett, Stuart Sandin, Anca Segall, Cynthia Silveira, Merry Youle, and Forest Rohwer. "Variability and Host Density Independence in Inductions-based Estimates of Environmental Lysogeny." *Nature Microbiology* 2, no. 7 (2017).

Knowles, B., C. B. Silveira, B. A. Bailey, K. Barott, V. A. Cantu, **Cobián Güemes, A. G.**, F. H. Coutinho, et al. "Lytic to Temperate Switching of Viral Communities." *Nature* 531, no. 7595 (2016): 466–70.

Gomez-Jimenez, S., L. Noriega-Orozco, R. R. Sotelo-Mundo, V. A. Cantu, **Cobián Güemes, A. G.**, R. G. Cota-Verdugo, L. A. Gamez-Alejo, L. Del Pozo-Yauner, E. Guevara-Hernandez, K. D. Garcia-Orozco, A. A. Lopez-Zavala, and A. Ochoa-Leyva. "High-Quality Draft Genomes of Two Vibrio Parahaemolyticus Strains Aid in Understanding Acute Hepatopancreatic Necrosis Disease of Cultured Shrimps in Mexico." *Genome Announcements* 2, no. 4 (2014).

Escalera-Zamudio, Marina, Martha I. Nelson, **Cobián Güemes, A. G.**, Irma López-Martínez, Natividad Cruz-Ortiz, Miguel Iguala-Vidales, Elvia Rodríguez García, et al. "Molecular Epidemiology of Influenza A/H3N2 Viruses Circulating in Mexico from 2003 to 2012." *PLoS ONE* 9, no. 7 (2014).

Vazquez-Perez, Joel, Pavel Isa, Darwyn Kobasa, Christopher E Ormsby, Jose E Ramírez-Gonzalez, Damaris P Romero-Rodríguez, Charlene Ranadheera, **Cobián Güemes, A. G.**et al. "A (H1N1) Pdm09 HA D222 Variants Associated with Severity and Mortality in Patients during a Second Wave in Mexico." *Virology Journal* 10, no. January (2013): 41.

Escalera-Zamudio, Marina, **Cobián Güemes, A. G.**, María de los Dolores Soto-del Río, Pavel Isa, Iván Sánchez-Betancourt, Aurora Parissi-Crivelli, María Teresa Martínez-Cázares, et al. “Characterization of Influenza A Virus in Mexican Swine That Is Related to the A/H1N1/2009 Pandemic Clade.” *Virology* 433, no. 1 (2012): 176–82.

Landa-Cardeña, Adriana, Jaime Morales-Romero, Rebeca García-Roman, **Cobián Güemes, A. G.**, Ernesto Méndez, Cristina Ortiz-Leon, Felipe Pitalúa-Cortés, Silvia Ivonne Mora, and Hilda Montero. “Clinical Characteristics and Genetic Variability of Human Rhinovirus in Mexico.” *Viruses* 4, no. 2 (2012): 200–210.

Arias, Carlos F., Marina Escalera-Zamudio, María De Los Dolores Soto-Del Río, **Cobián Güemes, A. G.**, Pavel Isa, and Susana López. "Molecular Anatomy of 2009 Influenza Virus A (H1N1)." *Archives of Medical Research* 40, no. 8 (2009): 643-54.

ABSTRACT OF THE DISSERTATION

Exploring the Global Virome and deciphering the role of phages in Cystic Fibrosis

by

Ana Georgina Cobián Güemes

Doctor of Philosophy in Biology

University of California San Diego, 2019

San Diego State University, 2019

Professor Forest Rohwer, Chair

Viruses are the most abundant and diverse life form on Earth. In Chapter 1 of this dissertation, a global census of the number of viral particles showed that 6.03×10^{31} viral particles are distributed across almost every ecosystem. The global census results showed that most viruses are in soils and sediments, two unexplored biomes for viral diversity. Accurate and fast bioinformatics methods to explore viral and bacterial composition in metagenomes are presented in Chapter 2 as "Fragment Recruitment Assembly Purification."

Phages are viruses that infect bacteria, their role in polymicrobial infections such as Cystic Fibrosis is explored. Cystic Fibrosis is a genetic disease in which the lung phenotype promotes mucus accumulation and microbial colonization. A multi-omics strategy to explore changes in the lung microbial community during acute pulmonary exacerbations is presented in Chapter 3 as “Cystic Fibrosis Rapid Response: Translating Multi-omics data into Clinically Relevant Information.” In Chapter 4, eight acute exacerbations were studied, and a trend of loss of diversity and viral lytic lifestyle was observed. Two Cystic Fibrosis patients studied in Chapter 4 were colonized by antibiotic resistant bacteria from the genera *Achromobacter*. Both patients suffered fatal exacerbations. This motivated the isolation and characterization of lytic phages to be used as antimicrobials against bacterial infections.

Chapter 1 : Viruses as Winners in the Game of Life

Abstract

Viruses are the most abundant and the most diverse life form. Their global abundance was previously estimated as 10^{31}, but their abundance remains uncertain and their global distribution vague. In this meta-analysis we estimated that there are 4.80×10^{31} phages on Earth. Further, 97% of them are in soil and sediments—two underinvestigated biomes that combined account for only ~2.5% of publicly available viral metagenomes. The majority of the most abundant phage sequences from all biomes are novel. Our analysis drawing on all publicly available viral metagenomes predicted a mere 257, 679 phage genotypes on Earth—an unrealistically low number— which attests to the current paucity of metagenomic data. Further advances in viral ecology and diversity call for a shift of attention to previously ignored major biomes and careful application of verified methods for viral metagenomic analysis.

J.B.S. Haldane observed, *"God has an inordinate fondness for stars and beetles."*

To which we would add, "…*and viruses."*

How many viruses?

Until the 1970s, viruses were of import only insofar as they were found to cause disease in us, our domesticates, and other eukaryotes of economic value. Bacteriophages — viruses that infect bacteria— were proving themselves to be eminently useful as model systems for molecular biology research, but were still thought to be of little significance to the functioning of the biosphere. There simply could not be enough environmental phages to

matter, given the then measurable number of potential prokaryotic hosts counted by culturing techniques. Since typically only ~ 1% of the *Bacteria* were culturable by methods at that time, bacterial populations were routinely underestimated by two orders of magnitude: the "great plate count anomaly" (Staley and Konopka, 1985) . In the late 1970s, the reported environmental bacterial populations jumped 100-fold or more with the development of improved direct counting methods employing epifluorescence microscopy (Hobbie, Daley, and Jasper 1977). In 1979 Torrella and Morita (Torrella and Morita 1979) concentrated particles larger than 0.2 μm from Oregon bay water by filtration and made direct counts of virus-like particles (VLPs) by transmission electron microscopy. Their counts, only about 10^4 mL^{-1} or one VLP per microbial cell, overlooked VLPs <0.2 μm, the majority of phages. This omission was corrected in 1989 when Bergh and associates centrifuged water samples directly onto grids for viewing by transmission electron microscopy (Bergh et al. 1989). They reported direct virion counts up to 1.5×10^7 mL^{-1}, an order of magnitude greater than their hosts. Moreover, based on several reasonable assumptions, they predicted that in marine environments, as much as one-third of the bacterial population suffered phage attack daily—hardly insignificant.

This was only the beginning as abundant microbes, and even more numerous phages, were then found in many environments, including inhospitable locations such as glacial ice (Anesio et al. 2007), bubbling acidic hot springs (Bolduc et al. 2015), deep-sea vents (Ortmann and Suttle 2005), and deep sediments (Engelhardt et al. 2014). Based on subsequent methodological developments, viruses are now recognized for what they are: the winners in

the game of life. They are the most abundant and the most diverse life forms, and are of great import for ecology and evolution.

A Global Census

How many viruses are there on Earth, and where are they located? One would expect to find the most viruses in the habitats with the most host organisms. The most abundant cellular organisms are the prokaryotes with an estimated 4.15×10^{30} cells, the majority of which are found in soil, subseafloor sediments, and marine waters (Table 1.1) (Whitman, Coleman, and Wiebe 1998; Williamson, Radosevich, and Wommack 2005; Kallmeyer et al. 2012; DeLong 2003). The number of microbial eukaryotes, including algae, is comparatively small, in the range of only 10^3 to 10^4 mL^{-1} in seawater and other planktonic environments (Rocke et al. 2015).

Taking a global phage census poses methodological challenges. Current methods for direct counting of VLPs commonly combine nucleic acid stains with EFM or flow cytometry. These methods were developed for aquatic environments, and their application to soil and sediment is problematic. Additional steps are required to detach prokaryotes and VLPs from particles and surfaces, and successful methods are specific to particular situations. Both may be obscured by opaque particles, and background fluorescence can interfere. The marine environment remains the most extensively sampled and best studied biome. Other intensively studied locations, such as the human gut and freshwater, make a relatively small contribution to the global total. Soils and the extensive subsurface sediments warrant more attention as their contribution is expected to be major.

Table 1.1 Estimated number of virus-like particles on Earth. Virus-to-microbe ratios (Supplemental Table 2 in Cobián et al., 2016) were extracted from 53 previous studies ratios (Supplemental Table 1 in Cobián et al., 2016) and used to calculate virus-like particles in each biome. Numbers of prokaryotic cells in marine, freshwater, other aquatic, sediment, and soil biomes are from Whitman at al., 1998. Number of cells per human is from Sender et al. 2016; human population is from United Nations, 2016. Median virus-to-microbe ratio for human biome is from Kim et al., 2011.

Biome	Number of prokaryotic cells	Median virus-to-microbe ratio	Virus-like particles per biome	Percentage of total virus-like particles
Marine	1.01×10^{29}	12.76	1.29×10^{30}	2.6828
Freshwater	1.26×10^{26}	14	1.76×10^{27}	0.0037
Other aquatic	2.44×10^{27}	30	7.32×10^{28}	0.1524
Sediment	3.80×10^{30}	11	4.18×10^{31}	87.0131
Soil	2.50×10^{29}	19.5	4.88×10^{30}	10.1481
Human-associated	2.80×10^{23}	0.1	2.80×10^{22}	0.0000
Other host-associated	Unknown	25	Unknown	Unknown
Total	**4.15×10^{30}**	**12**	**4.80×10^{31}**	

Marine sediment contains up to 10^5 times more organic matter than the water column above, supporting bacterial densities about 10^3 times higher (Kallmeyer et al. 2012). Summed globally, the total number of prokaryotes in subseafloor sediment (2.9×10^{29}) is roughly equal to the estimates of Whitman et al. (Whitman, Coleman, and Wiebe 1998) for the total number of prokaryotes in seawater (1.2×10^{29}) and in soil (2.6×10^{29}). Cell densities alone would predict good hunting for the phage, as the probability of non-specific phage-host contact would be ~10^3 times greater on average than in the water column above. Indeed, VLP counts of 10^9 ml^{-1} in sediment have been reported, three orders of magnitude greater than that

observed in the overlying water column at that depth (Danovaro and Serresi 2000). Similarly, the large numbers of prokaryotes in soil, the rhizosphere, and the rhizosheath, and their associated phages, remain largely terra incognita despite their accessibility and their importance to agriculture.

VLPs outnumber their hosts by approximately a factor of ten-to-one in various oceanic locations (Wommack and Colwell 2000), and similar ratios have been observed in some other biomes (Figure 1.1). It is not understood what drives this consistency, nor the observed variations. On this basis, earlier extrapolations from the prokaryote abundances in the major biomes yielded an estimated 1.2×10^{30} in open ocean, 2.6×10^{30} in soil, 3.5×10^{31} oceanic subsurface, and $0.25–2.5 \times 10^{31}$ in terrestrial subsurfaces, consistent with the rule-of-thumb estimate of 10^{31} viruses on Earth (Mokili, Rohwer, and Dutilh 2012). Here we revisited this question by combining the prokaryote populations drawn primarily from the classic 1998 paper by Whitman et al. (Whitman, Coleman, and Wiebe 1998) with the median measured VMRs from 53 previously-reported studies (Supplemental Table 1 in Cobian et al. 2016). In addition to the major biomes, we included some specialized environments of particular interest, such as niches associated with humans and other hosts. From a global perspective, the majority of prokaryotes inhabit soil, seawater, and the sediments. Thus, we expected that these biomes would also account for the majority of viruses.

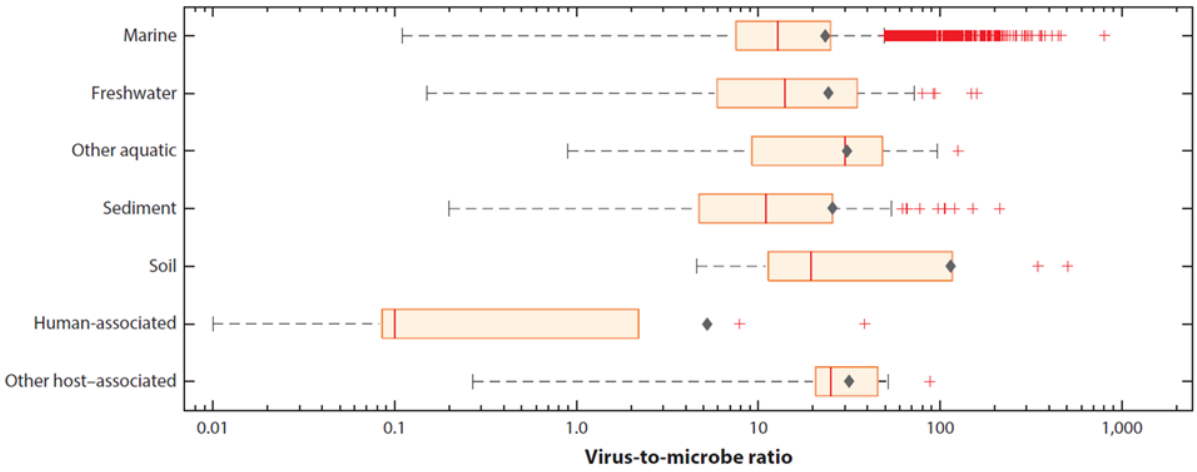


Figure 1.1 Virus-to-microbe ratios for major biomes. Box plots of biome virus-to-microbe ratios drawn from 53 previous studies (see Supplemental Methods). The means are indicated with gray diamonds. The medians are indicated with red lines.

This approach yielded 1.29×10^{30} VLPs in marine waters, 4.18×10^{31} in sediments, and 4.88×10^{30} in soil (Table 1.1). Adding in other lesser contributors brought the global total to 4.80×10^{31} VLPs (details in Supplemental Table 2 in Cobian et al. 2016). This confirms and slightly augments the oft-quoted number of 10^{31}. Given that the vast majority of cellular entities are prokaryotes, this number represents the global phage population. Significantly, sediments and soil combined accounted for 97% of the global total. In sum, phages are the most abundant life forms, exceeding the runner-up, the prokaryotes, by more than an order of magnitude. Populations of this magnitude dictate that phage evolution is driven by selection, with the contribution of drift being insignificant (Kimura 1962).

Caveats: To the best of our knowledge, there are only two published reports of VMRs for human-associated communities. One reported VMRs of 38.6 and 7.9 for gum-associated mucus and the adjacent milieu, respectively (Barr et al. 2013), while the other reported a mean VMR of 0.129 and a median VMR of 0.1 in fecal matter (M.-S. Kim et al. 2011). Since

the vast majority of human-associated microbes and VLPs are in the distal gut (Sender, Fuchs, and Milo 2016; Haynes and Rohwer 2011), the fecal VMR was used to calculate the total VLPs for the human-associated biome. Admittedly, that fecal VMR, being two orders of magnitude lower than the VMRs typical of most other biomes, was suspect. However, combining it with the recently revised estimate of 3.9×10^{13} human-associated prokaryotes (Sender, Fuchs, and Milo 2016) yields 3.9×10^{12} human-associated VLPs, in good agreement with the previous estimate of 3.0×10^{12} (Haynes and Rohwer 2011). Unraveling the phage-host dynamics in the human-associated microbiome will require further investigation.

Several factors may result in underestimations of phage abundance. 1) Equating the number of VLPs with the number of phages overlooks those residing as prophages in bacterial genomes—not an insignificant number (see below). 2) Some stains used to enumerate microbial cells and VLPs by EFM, such as SYBR Green, are relatively insensitive to single-stranded DNA and RNA, thus overlook many small viruses. A more representative assay can be made using SYBR Gold (Tuma et al. 1999). 3) VLP counts for soil and sediment are biased by reduced extraction efficiencies, while visualization and identification of VLPs can be compromised by particulate matter in samples of fecal matter, soil, sediment, etc. 4) Some VLP isolation methods for EFM include filtration through 0.2 μm filters which miss larger virions and also virions stuck to particulate matter. 5) The 0.02 μm grids used for TEM viewing miss the smallest virions.

Conversely, are these VLPs really viruses? In some environments, some VLPs may be gene transfer agents (GTAs), i.e., packaged cellular genes masquerading as tailed phage particles (Lang, Zhaxybayeva, and Beatty 2012). Although GTAs are produced by

roseobacters, a group that may account for more than 25% of the prokaryotes in some marine environments (Lang, Zhaxybayeva, and Beatty 2012), it remains unknown whether GTAs make a significant contribution to marine VLP counts. Also unknown is what fraction of the VLPs in various environments are infectious, and what percentage were defective upon release or subsequently inactivated by UV irradiation, physical damage, or enzymatic attack.

Phages Matter

How much do 10^{31} viruses matter? One measure of this is to consider the matter they contain. The mass of a typical phage virion containing 50 kbp DNA packaged inside an icosahedral capsid is calculated to be 0.0823 fg; of this, 0.054 fg is DNA and 0.0283 fg is the protein capsid (Antoni Luque, personal communication). For comparison, each *Prochlorococcus*, a particularly small autotrophic marine bacterium, has a mass of 300 fg. Based on this virion mass, the 4.80×10^{31} VLPs on Earth have a total mass of 3.95×10^{15} g or 3.95 Pg. The stoichiometry of carbon, nitrogen, and phosphorus (C/N/P) in this typical virion is 20/6/1 (Jover et al. 2014), which partitions that mass as approximately 0.06 fg C, 0.02 fg N, and 0.0075 fg P. Calculation based on these data suggests that the Earth's VLPs represent roughly 2.9 Pg C, i.e., two orders of magnitude less than the 350-500 Pg total carbon previously estimated for Earth's prokaryotes (Whitman, Coleman, and Wiebe 1998). Compared to microbial cells, virions are enriched in both nitrogen and phosphorus, with estimated global totals of 0.96 Pg N, and 0.36 Pg P. As a result of this enrichment, >5% of the total marine DOP and DON pools is estimated to reside in virions in some locations (Jover et al. 2014).

It is now widely recognized that microbes are of tremendous importance for global biogeochemical processes, and consequently, that which controls the microbes — the phages — runs the world. Prokaryotes are subject to predation by both heterotrophic protists and phages. Grazing by protist predators is size-selective, whereas lysis by phages is strain-specific. Moreover, in the oceans grazing moves the organic carbon and nutrients to higher trophic levels, whereas lysis routes these components instead through the viral shunt as dissolved organic matter (DOM). This DOM feeds the heterotrophic microbial community, thereby increasing net primary productivity and slowing the movement of carbon to the deep ocean (Fuhrman 1999; Weinbauer 2004; Wommack and Colwell 2000; Weitz et al. 2015; Proctor and Fuhrman 1990; C A Suttle 2005). Because of the stoichiometric mismatch between virions and their host cells, after lysis, a disproportionate amount of the P is found in the progeny virions, leaving the cellular debris that feeds the heterotrophs depleted in P (Jover et al. 2014). An estimated 10^{28} marine bacteria are lysed daily including ~30% of the cyanobacteria and ~60% of the heterotrophic bacteria (Curtis A. Suttle 2007; Proctor and Fuhrman 1990). This selective, strain-specific lysis profoundly impacts prokaryote community diversity increasing both richness and evenness (Wommack and Colwell 2000; Fuhrman 1999; T.F. Thingstad 2000; T.F. Thingstad et al. 2015; Sandaa et al. 2009).

Phages also add a horizontal dimension to microbial evolution by mediating the transfer of genes between cells (John H Paul 1999). They nab useful metabolic genes from their hosts, maintain them, evolve them further to suit their own needs, and then sometimes return the new version to the microbial gene pool (Frank et al. 2013; Lindell et al. 2004; Sullivan et al. 2006; D.B. Goldsmith et al. 2011). On an ecosystem level, the phage

community encodes environment-specific repertoires of microbial metabolic genes (Dinsdale et al. 2008).

Through lysogeny, phage genes are a continual presence in microbial communities. Identification of prophages in microbial genomes is challenging, and numerous bioinformatic methods have been developed (Bose and Barber 2006; Zhou et al. 2011; McNair, Bailey, and Edwards 2012; Akhter, Aziz, and Edwards 2012; Fouts 2006) Estimates of the percentage of prokaryotes in various biomes that carry one or more prophages have ranged between 0% and 100% (J.H. Paul 2008). Approximately 82% of the sequenced prokaryotic genomes available as of 2015 are predicted to contain at least one prophage (Katelyn McNair, personal communication). Resident prophages often account for the differences between strains within a bacterial species (Canchaya et al. 2003). Prophage-encoded exotoxin genes cause many notable human diseases, including cholera, diphtheria, and enterohaemorrhagic diarrhea (Casas et al. 2006) as well as diseases that plague our agriculture and aquaculture.

We do not yet understand the factors influencing the prevalence of lysogeny in any biome. Siphoviruses have long been associated with the temperate lifestyle, although lysogeny is not confined to that family. This group was observed to dominate the community in the Southern Ocean, marine sediment, desert, hypersaline ponds, and human fecal samples, accounting for 44% of the total in sediment (M. Breitbart et al. 2004; M Breitbart et al. 2002; M. Breitbart et al. 2003; Brum et al. 2015; Adriaenssens et al. 2015; Roux et al. 2016). However, whether a temperate phage will follow the lytic or lysogenic pathway is decided at the start of each infection in response to host and environmental factors. The common interpretation based primarily on studies of coliphage λ posits that host abundance, as sensed

by a low multiplicity of infection (MOI), favors lytic replication, while high MOI favors lysogeny (Herskowitz and Hagen 1980). A recent analysis of phage communities on coral reefs suggests more complex dynamics (Knowles et al. 2016).

How many different viruses?

Ever since the discovery of phages, it has been evident that there are different ones capable of killing different bacteria. However, one hundred years later we are still do not know the extent of this diversity, its biogeography, or its dynamic role in ecosystem function. While the diversity of all cellular life has been probed using universal genes such as the small subunit ribosomal RNA gene, the polyphyletic viruses have no gene in common, thus precluding a comprehensive PCR-based survey of viral diversity (F Rohwer and Edwards 2002; Dwivedi et al. 2012). Therefore, other methods had to be devised.

Two approaches based on data available in 2003 both yielded an estimated 100 million phage 'species' (F. Rohwer 2003). One conservatively assumed that 10 phage 'species' infect each of the estimated 10 million microbial species — thus 100 million different phages. This alone does not measure genetic diversity since two 'different' phages might differ in only one or two key proteins that determine host range. Similarly, the swapping of structural gene modules can yield a new 'species' that differs in virion morphology without any increase in the global genetic diversity, while two phages indistinguishable by morphology and host range can differ significantly in other genome modules. The second approach compared all the sequenced phage ORFs in GenBank at that time using BLAST and clustered the ORFs using a E-value of 10^{-4}. From this data the non-parametric estimator Chao1 (Chao 1984) predicted that 2×10^9 phage ORFs remained to be discovered. Assuming 50 ORFs per phage

genome with 50% of those being novel, calculations based on the Chao1 value predicted 100 million different phages.

A decade later, armed with more metagenomic data and new analysis tools, Sullivan and colleagues presented an analysis based on protein clustering that reduced the estimated total number of phage ORFs from 2×10^9 to only 3.9×10^6 (Ignacio-Espinoza, Solonenko, and Sullivan 2013), a demotion of almost three orders of magnitude. The debate continues.

Signature Genes

Although there is no universal phage gene, diversity within phage groups can be assessed using signature genes shared by all group members (Adriaenssens and Cowan 2014). For instance, the capsid portal gene (*g20*) is conserved among many of the large cyanomyophages that inhabit marine and freshwater environments. Several studies using *g20* reported rich community diversity, typically 100 or more OTUs and including clades for which there are no cultured isolates (Zhong et al. 2002; Jameson et al. 2011). Attempts to correlate variations in community composition with depth, host abundance, season, and geographic distance yielded inconsistent results. Some OTUs demonstrated consistent seasonal variation while others persisted in moderate abundance from year to year (Chow and Fuhrman 2012). Sampling across a north-south Atlantic Ocean transect found similar cyanomyophages to be widely distributed with no apparent geographical segregation (Jameson et al. 2011), while others were present in both marine and freshwater environments (Dorigo, Jacquet, and Humbert 2004). However, some diversity within this group eluded these surveys; of 39 cyanophage isolates from the Gulf of Mexico, only 63% carried detectable *g20* sequences (McDaniel, DelaRosa, and Paul 2006).

The entire T4 superfamily (*Myoviridae*) was similarly surveyed using their major capsid protein (MCP, gp23). In aquatic environments, these are primarily T4-like cyanomyophages. The results here echoed the same trends: diversity (more than 100 OTUs) exceeding that represented in cultured isolates (Comeau and Krisch 2008), seasonal succession patterns, and long-term persistence of some OTUs (Chow and Fuhrman 2012; Pagarete et al. 2013). In some cases persistent OTUs were also the most abundant (>4% relative abundance), thus contradicting predictions of both the Bank Model of viral community structure and the classic Kill-the-Winner dynamic (T. Thingstad and Lignell 1997; Pagarete et al. 2013; M. Breitbart and Rohwer 2005).

These analyses based on *g20* and gp23 are limited to the myophages (predominantly cyanomyophages) and miss the podophages and the abundant siphophages. A more inclusive assessment of cyanophage diversity, including myophages and podophages, used *psbA*, the gene that encodes the D1 protein of oxygenic photosystem II (Chenard and Suttle 2008). Phage-encoded sequences clustered by both phage family and by host, separate from the host *psbA* genes, and included clusters with no cultured isolates. Multiple clusters coexisted in some locales while some clusters extended over vast geographic distances. Moreover, one cyanopodovirus subcluster was found to be globally distributed based on its DNA polymerase gene (*pol*) (Huang et al. 2010).

Other signature genes can provide a more complete picture of marine phage diversity. For example, the phosphate-starvation gene *phoH* is present in nearly 40% of all marine phages compared to 4% of non-marine phages, reflecting the scarcity of phosphate in the marine environment (D.B. Goldsmith et al. 2011). It is not restricted to a single viral family

and is found in phages that infect heterotrophs as well as autotrophs, even in viruses of photosynthetic green algae. A *phoH*-based marine survey found, yet again, that most of the environmental diversity was not represented in cultured isolates, that *phoH* homologs are widely distributed with most clusters represented in multiple oceanic regions, and that marine phage community composition varies with depth and geographical location. A subsequent survey of the *phoH* genes in Sargasso Sea phage communities at depths of 0 to 1,000 m over a two-year period identified 3,619 OTUs (97% identical) and provided new insights into community dynamics (Dawn B Goldsmith et al. 2015). Approximately 96% of those OTUs were rare, each accounting for <0.01% of the total sequences, while more than 50% of the sequences were from five abundant OTUs. The presence of a few abundant OTUs (1-4 in any particular sample) and many rare ones is consistent with the Bank Model of phage community structure. However, whereas that model predicts the cycling of phages between the two groups over time, here the rare OTUs remained rare, and the most abundant OTUs persisted through seasons and years.

To date, signature genes have been developed for only some viral families. They are strikingly lacking for the *Siphoviridae*, the family that includes many temperate phages and that dominates both metagenomic datasets and cultured isolates. In even the best cases, signature genes fail to capture the full richness present in natural communities. However, they have provided insights into the global distribution of specific phage genes. The DNA polymerase conserved in the T7-like podophages and restricted to that group was found in multiple biomes (M. Breitbart, Miyake, and Rohwer 2004). Moreover, identical or nearly-identical 533 bp segments were recovered from different biomes, indicating that phages, or at

least phage genes, have moved between biomes in recent evolutionary time. Similarly, >98% identical sequences from algal virus DNA polymerase genes were found from the northern Pacific Ocean to Antarctica (Short and Suttle 2005). These observations could represent either the movement of phages or of individual phage genes. That phages from freshwater, sediment, and soil are able to infect marine prokaryotic communities suggests that phages can move successfully between biomes (Sano et al. 2004). These, and similar observations, suggest that the global viral diversity may be less than previously estimated based on the diversity in individual biomes.

Phage Metagenomics

Expansion of the field of view from signature genes to assessment of phage community diversity was made possible by the development of viral metagenomics. Earlier sequencing methods that required cloning of phage DNA had often encountered issues. Many phages carry genes that are lethal to the cloning host cells, or their DNA contains modified bases that block cloning. An alternative method, linker-amplified shotgun sequencing, was first used to assess near-shore marine communities (M Breitbart et al. 2002). Sample preparation included passage through a 0.16 μm tangential flow filter, purification of VLPs by CsCl gradient centrifugation, and subsequent DNA extraction. This procedure did not recover large viruses (e.g., algal Phycodnaviruses) or RNA phages. This initial marine survey found that more than 65% of the phage sequences were novel. Four years later, viromes prepared using next generation sequencing from four oceanic regions contained >90% unknowns (Angly et al. 2006). Even 15 or more years later, despite the increased number of sequenced viral genomes in the public databases, 60-99% of the sequences in viromes from diverse

biomes are still unknowns (Mokili, Rohwer, and Dutilh 2012; Brum and Sullivan 2015; Adriaenssens et al. 2015; Watkins et al. 2015; Roux et al. 2016). More than 99% of viral genetic diversity remains to be explored (Mokili, Rohwer, and Dutilh 2012).

Even though most sequences recovered are unknowns, bioinformatics methods can provide insights into community structure. Metagenomic reads are assembled into contigs *in silico*. The more diverse the community, the lower the probability of sequencing two overlapping fragments from the same genome, thus shorter contigs and more unassembled singleton reads. Plots of contig spectra (number of contigs versus contig length) were best represented by power-law based mathematical models and provided estimates of both the richness and evenness of the sampled community (M Breitbart et al. 2002). Application of this approach to near-shore marine communities estimated 374 – 7,114 genotypes present. Of these, the most abundant one represented only 2 – 3% of the total community, while only three contributed more than 1% of the reads. Subsequent metagenomic surveys of diverse environments reported genotypes numbering in the hundreds to tens of thousands [reviewed in (Youle, Haynes, and Rohwer 2012); data in (M. Breitbart et al. 2003; Angly et al. 2006; Tseng et al. 2013; Bolduc et al. 2015; Youle, Haynes, and Rohwer 2012)].

Virome Meta-analysis

We have developed a new bioinformatics method, FRAP (Fragment Recruitment, Assembly, Purification) (Figure 1.2, Supplemental Methods) and used it to assess global phage diversity by analyzing 1,623 publicly available viromes (Supplemental Table 3 in Cobian et al. 2016). To create a reference library of the observed phage genotypes on Earth, all reads from each of the viromes were assembled separately using SPAdes with the K value

adjusted to maximize the incorporation of base pairs into contigs (Bankevich et al. 2012). Comparative evaluations of assembler performance on metagenomic datasets had ranked SPAdes among the best based on contig accuracy (García-López, Vázquez-Castellanos, and Moya 2015). We assumed that each ≥ 1kbp contig represented a partial phage genotype that was relatively abundant in that biome. Rarely more than one non-overlapping contig might have been recovered from the same genotype. All of these ≥1 kbp contigs were added to the reference library. In addition, the 2,669 sequenced phage genomes (December 2015) and 67 archaeal virus genomes (February 2016) in the NCBI Viral Genomes database were added to the library, along with an additional 123 bacteriophage genomes from the Broad Institute Marine Phage Sequencing Project. Subsequent dereplication by CD-HIT (Li and Godzik 2006) at 98% identity yielded 2,267,978 phage contigs plus genomes.

Given this reference library, the fragment recruitment step could then retrieve the matching reads from any virome. In principle this FRAP method could be used to retrieve and thus purify sequences that are only minor components of any dataset. For our analysis all reads from all viromes were mapped to the reference library at 90%. 95%, and 99% identity, and the normalized number of hits to each contig was tallied for each biome (Table 1.2, Supplemental Methods). This yielded the fractional abundance in each biome of every contig that is also present in the reference library (Figure 1.3 b–g). Parallel mapping was performed on a pooled global virome prepared by proportionate subsampling of the individual biome files (Figure 1.3 a). Of these pooled reads, 87% were from sediment and 10% were from soil. The ten most abundant viral contigs in each case occupied the same rank for all three mapping identities, evidencing the robustness of the method.

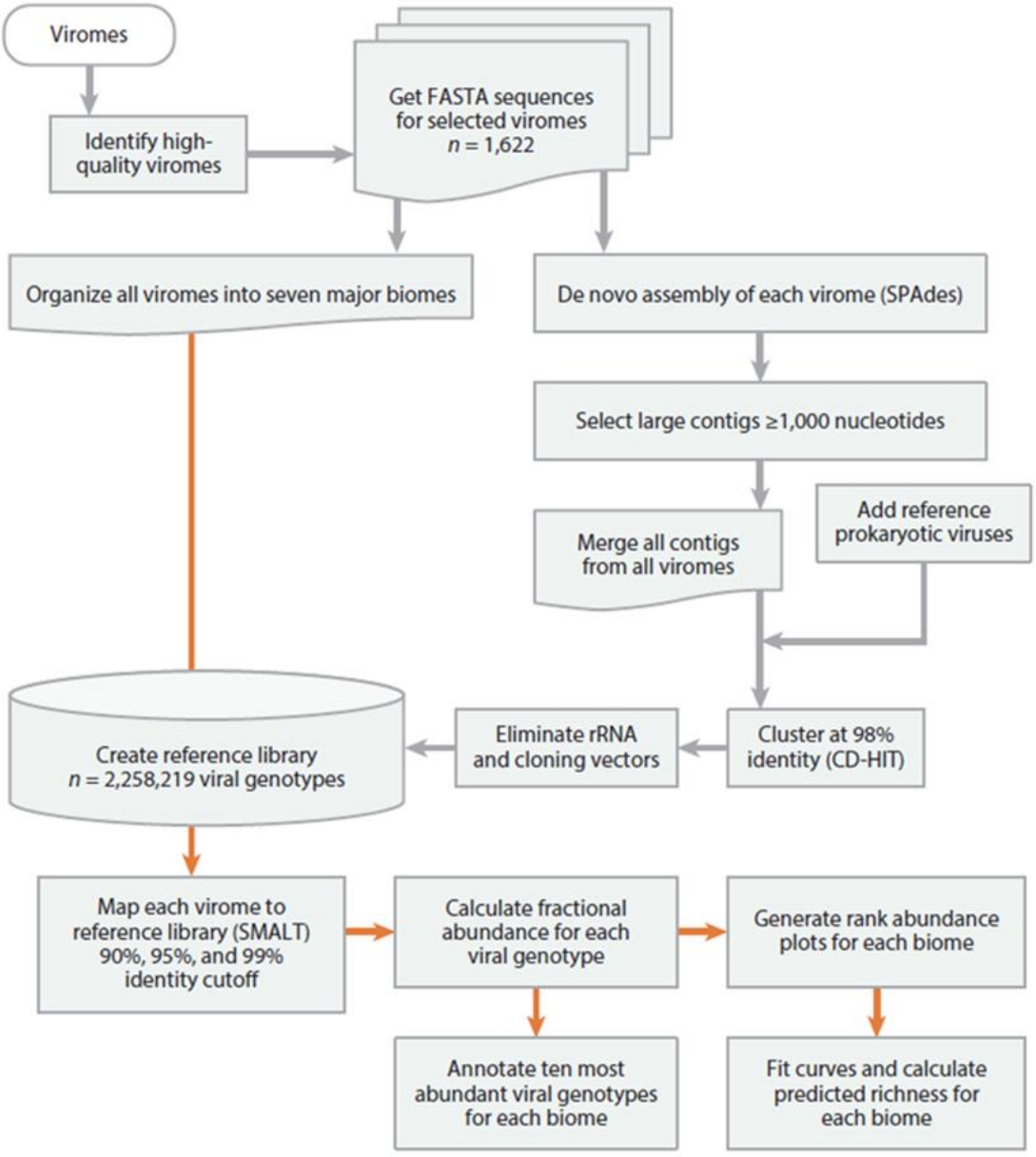


Figure 1.2 Bioinformatics pipeline for the FRAP (fragment recruitment, assembly, purification) method (see Supplemental Methods).

Table 1.2 Virome metrics. [1]For the Tara Oceans Virome data set, only 1% of the reads were used. Including all would increase the number of reads from the marine biome to 1,688,798,702.

Biome	Number of viromes	Number of reads	Percentage of reads assembled into ≥1 kbp contigs	Percentage of reads mapped to reference library		
				90% identity	95% identity	99% identity
Marine	192	56,676,517[1]	11.83	38.1	29.4	20.4
Freshwater	48	11,519,523	19.51	15.7	11.4	8.1
Other aquatic	19	3,342,537	14.04	40.2	35.8	27.3
Sediments	21	15,729,082	10.00	54.7	51.7	44.2
Soil	9	2,459,152	15.54	41.9	36.6	29.0
Human-associated	1158	481,172,486	25.16	30.3	27.2	12.4
Other host-associated	167	34,600,192	3.81	34.6	31.1	25.3
Other	7	396,889	28.22	44.3	36.8	17.4
Total	**1621**	**605,896,378**	**16.01**	**31.8**	**28.1**	**14.7**

Mapping at 99% identity provided the most conservative estimate of the number of identified viral contigs and was used to calculate the number of viral genotypes potentially encoded for each biome (Figure 1.3). Here we summed the lengths of all observed phage contigs and divided by the assumed average phage genome size of 50 kbp. The most viral genotypes were observed in the marine and human-associated biomes, the least in soil. This reflects the size of the datasets: 192, 1158, and 7 viromes for marine, human-associated, and soil, respectively. This meta-analysis demonstrated that even when using all the information currently available for DNA-containing phages, we detect only 1,160 viral genotypes in the

pooled global virome, likely due to the low number of viromes from soil and sediment biomes.

The rank abundance plots for each biome are best described by a power law model (Supplemental Table 5 in Cobian et al. 2016). Using this model, we calculated the predicted richness of each biome (Supplemental Methods). This approach predicted only 24,795 viral genotypes for the marine biome, which is not realistic (Supplemental Table 5 in Cobian et al. 2016). We conclude that we have insufficient information about the shape of the curve to calculate the actual viral richness of each biome. Further work is needed before we even know how much we still do not know, i.e., how far we are from understanding the viral dark matter of the biosphere.

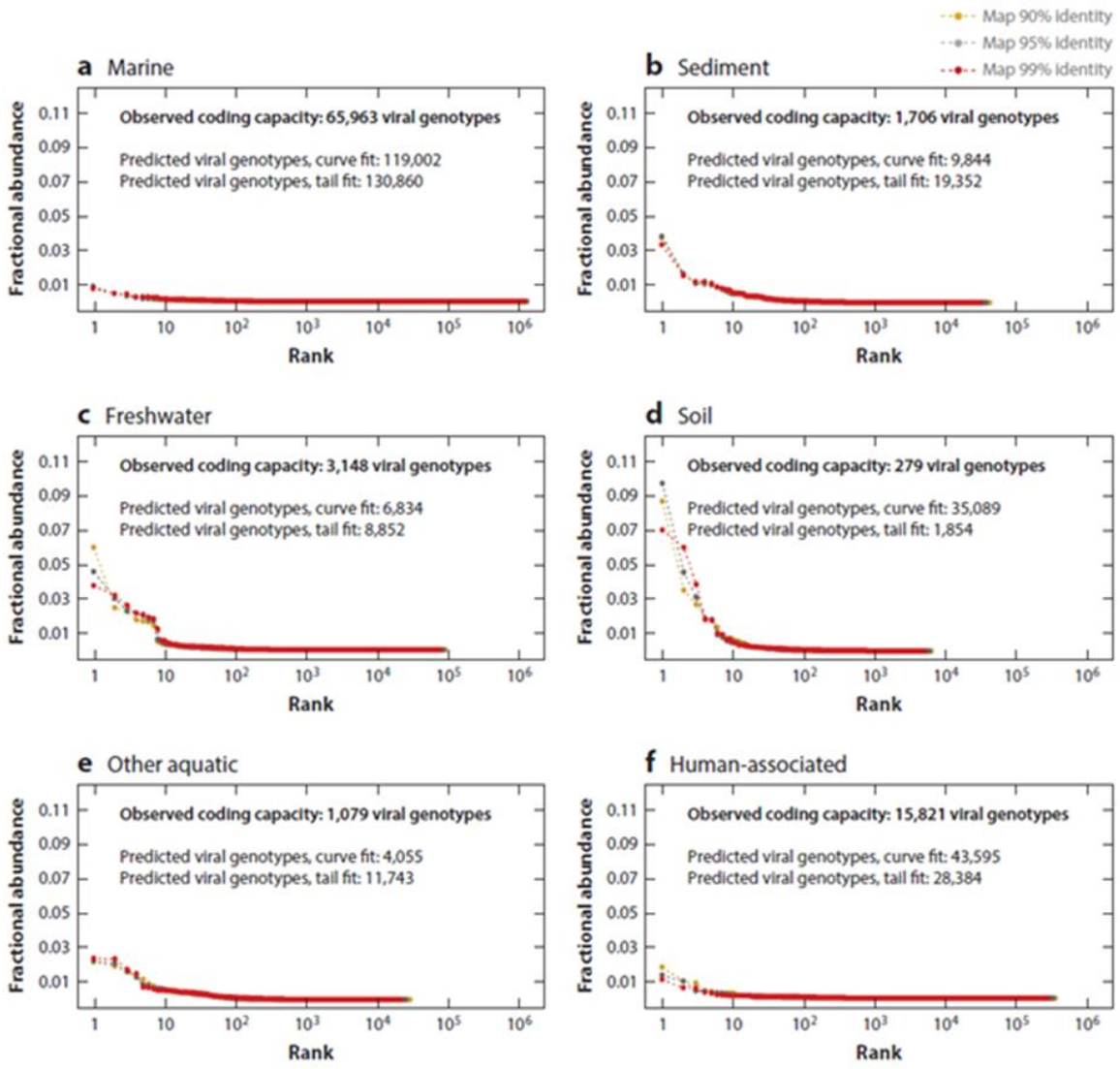


Figure 1.3 Viral rank abundance curves for (*a*) marine, (*b*) sediment, (*c*) freshwater, (*d*) soil, (*e*) other aquatic, and (*f*) human-associated biomes. Reads in these six major biomes were mapped to the genotypes present in the reference library at 90%, 95%, and 99% identity. In each case, the relative abundance of all recovered viral genotypes is shown as well as the observed coding capacity expressed as the potential number of different 50-kbp phage genomes and the predicted viral genotypes using both the curve fit and tail fit methods.

How many different phages on Earth?

Providing a direct answer to this question remains challenging, in part because we do not know whether phages are provincial or cosmopolitan. If a biome is sampled in two geographic locations, A and B, and the number of phage genotypes present in each is estimated, is the richness of the phage community in that biome equal to A + B, or is it significantly less? Likewise, if the number of phage genotypes in each biome is known, are we justified in summing them to calculate the global virome? Other studies have addressed this by assessing the fraction of genotypes shared between biomes or geographical regions. A three-pronged analysis of viromes from four oceanic regions (the Pacific off the coast of British Columbia, Arctic Ocean, Gulf of Mexico, and Sargasso Sea) found that a large fraction of the phage community is cosmopolitan, that is they are found in two, three, or even four of the surveyed regions (Angly et al. 2006). Within this global distribution, the phage communities showed regionalization in that community members, including the cosmopolitan and the most abundant, shifted in relative abundance from region to region. Assembly of reads from each region separately yielded a total of ~150,000 genotypes, whereas co-assembly reduced the total to only 57,600 different genotypes. Similarly, a recent study of hypersaline ponds on three continents found that community composition varied with the level of salinity but that communities in the same salinity are genetically connected across the globe (Roux et al. 2016). Conversely, phage communities present in three soil biomes shared essentially no genotypes (Fierer et al. 2007).

The cosmopolitan range of individual phage genes or gene modules is another factor that complicates determination of the ecological or geographical range of phage genotypes.

Studies described earlier that found nearly identical sequences of signature genes to be globally distributed indicated a global phage gene pool, at least within the marine biome. However, even within that biome, this does not distinguish between the global travels of phage genes by horizontal gene transfer and the presence of the same phage genotypes in geographically remote locations.

Here we used the number of phage encoding capacity observed in our meta-analysis to estimate the number of possible phage genotypes on Earth (Supplemental Methods). In brief, we used the sum of lengths of all the phage contigs as a proxy for all phage DNA currently available in the databases, and then divided by the assumed 50 kbp average phage genome length (Supplemental Table 5 in Cobian et al. 2016). On this basis we estimate that the observed phage DNA is sufficient to encode 257, 698 different phages.

This estimate is undoubtedly low. The sequenced samples so far represent only a minute fraction of the total phage-encoded information present in every biome. For example, including all 2.16×10^9 reads from the massive Tara Oceans Viromes dataset (average read length of ~101 bp) would provide 2×10^8 kbp of marine phage genomic sequence. To put this in perspective, 1.29×10^{30} marine VLPs with an average genome of 50 kbp would contain 6.5×10^{31} kbp of DNA.

Which are the most abundant phage genes?

To explore this, we annotated the 10 most abundant contigs in each biome and in the pooled global virome (Supplemental Table 4 in Cobian et al. 2016). The majority of them have no significant similarity to sequences in the NCBI database. For the soil biome, none of the 10 have similarities to known sequences. A podovirus polymerase is the most abundant

contig in the marine biome followed by several uncultured virus clones and unknown sequences. For the freshwater biome, nonstructural viral genes were detected among the most abundant ones, as well as others with no similarities to known sequences. Similarly, for the global virome the most abundant contigs have no significant hits in the databases, followed by contigs annotated as circovirus, a circular virus, and a cryptic MLU1 plasmid from *Micrococcus*.

Caveats: Our reference database was clustered at 98% identity. The 2 most abundant viral contigs in the marine biome were podovirus polymerase genes that share >98% identity among themselves. Methodological improvements allowing clustering at 100% identity sequences would eliminate this issue.

Further, all of our results are unavoidably skewed due to the biased distribution of the current viromes. Of the 1,623 viromes, 71% were from human-associated communities, 10% from communities associated with other animals or plants, and 11% from marine environments (Table 1.2). Thus 92% of the viromes explored 3% of the VLPs on Earth, while only 1.8% investigated the two biomes with 97% of the VLPs — soil and sediment (Table 1.1). Some biomes remain virtually unsampled. Even in the marine biome, only a very limited geographic territory and range of environmental parameters have been sampled. Despite the surging interest in the human microbiome, published counts for human-associated prokaryotic cells and VLPs are strikingly lacking.

The publicly available viromes are further compromised by poor methods. Of those viromes, 132 (7.5%) were omitted from our library because they were mislabeled microbial metagenomes or showed obvious contamination with human or microbial sequences. Some

contained abundant φX174 sequences due to either the amplification bias of multiple displacement amplification (K.-H. Kim and Bae 2011; Duhaime and Sullivan 2012) or, when sequencing on the Illumina platform, the failure to remove the PhiX quality control sequences prior to submission to the public databases. In the sediment biome, the highly abundant fragments from *Staphylococcus aureus* plasmids could be from phages or GTAs, or from bacterial contaminants of viromes due to inadequate viral purification during sample preparation. Many sources of error can be avoided and more quantitative data obtained by carefully adhering to current best practices (Duhaime and Sullivan 2012; Duhaime et al. 2012; K.-H. Kim and Bae 2011). Still needed is the comparable development of methods for RNA viruses and ssDNA phages, as well as VLP purification procedures that do not exclude the largest viruses.

The future for FRAP

Given a high-quality reference library, FRAP can be used to get rid of the crap, i.e., to fish out matching sequences from a metagenomic dataset, even when they comprise only a small percentage of the reads. This can potentially eliminate some sample purification steps or enable you to selectively retrieve different components from a mixed sample. However, FRAP's utility depends on the completeness and quality of the reference library. Library development, in turn, calls for clean sample preparation methods, sequencing of more bp, and longer read lengths (such as the 10 to 15 bp lengths now possible with PacBio SMRT® sequencing).

Recent advances in phage ecology have heightened our awareness that we live in a phage world. Much of that world remains a terra incognita. For the careful researcher equipped with today's technology, the opportunities for discovery are vast.

Summary Points

1. Phages are the winners: the most numerous and genetically diverse life forms on Earth. The estimated 4.80×10^{31} VLPs on Earth comprise at least 257,698 different phage genotypes.

2. We have barely begun to explore phage diversity. Metagenomic studies have focused on a few biomes and have ignored soil and sediment—the two that combined contain 97% of the global phage population. Sampling has been sparse at best, while numerous environments remain terra incognita.

3. Global phage diversity far exceeds that represented by cultured isolates.

4. One cannot fully understand the ecology or evolution of any ecosystem without including the phages.

5. Our current knowledge of the fractional abundances of phages on each biome is limited and we need better strategies to describe the population structure of the phages on each biome.

Future Issues

1. Metagenomic sampling has been narrowly focused on selected regions of the marine environment and on host-associated communities, primarily human-associated. Most

of the globe and most biomes await exploration. Initial surveys of the human-associated phage community uncovered anomalous dynamics that await resolution.

2. Viruses that have chromosomes of RNA or single-stranded DNA are significant components of some communities, but have been generally ignored. Correcting this requires new inclusive methods for both their direct VLP counts and their metagenomics.

3. Some phages and some phage genes are cosmopolitan, while others appear to be geographically or ecologically restricted. The question remains: do all phages share the same global gene pool, with varying levels of access?

4. Genomic analysis of both phages and their hosts indicates that lysogeny is commonplace. Awaiting further exploration are the prevalence of temperate phages in various environments, the lysis-lysogeny decision, and the impact of lysogeny on the ecology and evolution of both phages and their hosts.

5. Phages are the greatest reservoir of unexplored genetic diversity on Earth. Current estimates of the number of different proteins encoded by phages vary widely. After more than a decade of phage metagenomics, the majority of phage sequences remain novel. Sequencing technology has advanced rapidly, and now enhanced bioinformatics methods are essential to analyze the mushrooming metagenomic data.

6. Assignment of function to phage-encoded proteins based on sequence homology is limited by the rapid rate of phage evolution. Other approaches are called for in order to translate environmental metagenomic data into the metabolic potential of phage communities.

7. The species concept is not directly applicable to viruses. An alternative, generally-accepted metric is needed to facilitate discussion of viral diversity, ecology, and evolution.

8. In the current era of the microbiome, researchers are actively investigating the roles of microbes in processes including human health, ecosystem functioning, and global biogeochemical cycles. Now, a century after the discovery of phage, exploration of the role of phage in these and other activities is overdue.

Definition of terms

Virome: a viral metagenome

OTU: operational taxonomic unit

Phage: a virus that infects a prokaryote (Bacteria and Archaea)

Microbe: a prokaryote, i.e., an archaeon or a bacterium

PFU: plaque-forming unit

Virion: the intercellular transport form of a virus, typically comprising the chromosome(s) enclosed within a protein capsid

VLP: virus-like particle

VMR: the virus-to-microbe ratio

EFM: epifluorescence microscopy

fg: 10^{-15} g

Pg: 10^{15} g

Prophage: a phage chromosome that resides within a host cell (lysogen) without immediately engaging in lytic replication

Cyanophage: a phage that infects cyanobacteria

Myophage: a member of the family *Myoviridae*

Podophage: a member of the family *Podoviridae*

Siphophage: a member of the family *Siphoviridae*

DOP: dissolved organic phosphorus

DON: dissolved organic nitrogen

Acknowledgements

Chapter 1, in full, is published in Annual Review of Virology. Ana Georgina Cobián Güemes, Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton and Forest Rohwer; 2016. The dissertation author was the primiary investigator and author of this paper.

All calculations were made on Rob Edwards's lab cluster at San Diego State University, which is supported by National Science Foundation grant DBI-0850356 for computational resources. We are thankful to the members of the Biomath group at San Diego State University for critical discussion of this work; to Cynthia Silveira for early access to marine VMRs; to Rizki Wulandari, Jan Janouˇskovec, Nate Robinett, and Ben Knowles for suggestions and comments; and to Evelien Adriaenssens for data sharing. This work was partially supported by the National Council on Science and Technology (CONACyT), Mexico.

References

Adriaenssens, Evelien M, and Don A Cowan. 2014. "Using signature genes as tools to assess environmental viral ecology and diversity." Appl. Environ. Microbiol. 80 (15): 4470-4480.

Adriaenssens, Evelien M, Lonnie Van Zyl, Pieter De Maayer, Enrico Rubagotti, Ed Rybicki, Marla Tuffin, and Don A Cowan. 2015. "Metagenomic analysis of the viral community in Namib Desert hypoliths." Environ. Microbiol. 17 (2): 480-495.

Akhter, Sajia, Ramy K Aziz, and Robert A Edwards. 2012. "PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies." Nucleic Acids Res. 40 (16): e126-e126.

Anesio, Alexandre M, Birgit Mindl, Johanna Laybourn-Parry, Andrew J Hodson, and Birgit Sattler. 2007. "Viral dynamics in cryoconite holes on a high Arctic glacier (Svalbard)." Journal of Geophysical Research: Biogeosciences 112 (G4).

Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley, and H. Liu. 2006. "The marine viromes of four oceanic regions." PLoS Biol. 4 (11): e368.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, and Andrey D Prjibelski. 2012. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." J. Comput. Biol. 19 (5): 455-477.

Barr, Jeremy J, Rita Auro, Mike Furlan, Katrine L Whiteson, Marcella L Erb, Joe Pogliano, Aleksandr Stotland, Roland Wolkowicz, Andrew S Cutting, Kelly S Doran, P. Salamon, M. Youle, and F Rohwer. 2013. "Bacteriophage adhering to mucus provide a non-host-derived immunity." Proc. Natl. Acad. Sci. USA 110 (26): 10771-10776.

Bergh, Ø., K Y Børsheim, G Bratbak, and M Heldal. 1989. "High abundance of viruses found in aquatic environments." Nature 340: 467-468.

Bolduc, Benjamin, Jennifer F Wirth, Aurélien Mazurie, and Mark J Young. 2015. "Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis." ISME J. 9: 2162-2177.

Bolduc, Benjamin, Ken Youens-Clark, Simon Roux, Bonnie L Hurwitz, and Matthew B Sullivan. "iVirus." http://ivirus.us/.

Bose, Michael, and Robert D Barber. 2006. "Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences." In Silico Biol. 6 (3): 223-227.

Breitbart, M, P Salamon, B Andresen, J M Mahaffy, A M Segall, D Mead, F Azam, and F Rohwer. 2002. "Genomic analysis of uncultured marine viral communities." Proc. Natl. Acad. Sci. USA 99 (22): 14250-14255.

Breitbart, M., B. Felts, S. Kelley, J.M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2004. "Diversity and population structure of a near–shore marine–sediment viral community." Proc. R. Soc. Lond. B Biol. Sci. 271 (1539): 565.

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

Breitbart, M., J.H. Miyake, and F. Rohwer. 2004. "Global distribution of nearly identical phage encoded DNA sequences." FEMS Microbiol. Lett. 236 (2): 249-256.

Breitbart, M., and F. Rohwer. 2005. "Here a virus, there a virus, everywhere the same virus?" Trends Microbiol. 13 (6): 278-284.

Broad Institute. "Marine Phage Sequencing Project." www.broadinstitute.org/annotation/viral/Phage.

Brum, Jennifer R, Bonnie L Hurwitz, Oscar Schofield, Hugh W Ducklow, and Matthew B Sullivan. 2015. "Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics." ISME J.

Brum, Jennifer R, and Matthew B Sullivan. 2015. "Rising to the challenge: accelerated pace of discovery transforms marine virology." Nat. Rev. Microbiol. 13 (3): 147-159.

Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M.L. Dillmann, and H. Brüssow. 2003. "Phage as agents of lateral gene transfer." Curr. Opin. Microbiol. 6 (4): 417-424.

Casas, V., J. Miyake, H. Balsley, J. Roark, S. Telles, S. Leeds, I. Zurita, M. Breitbart, D. Bartlett, and F. Azam. 2006. "Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California." FEMS microbiology letters 261 (1): 141-149.

Chao, A. 1984. "Non-parametric estimation of the number of classes in a population." Scan. Stat. Theory Appl. 11: 265-270.

Chenard, C, and CA Suttle. 2008. "Phylogenetic diversity of sequences of cyanophage photosynthetic gene psbA in marine and freshwaters." Appl. Environ. Microbiol. 74 (17): 5317-5324.

Chow, Cheryl-Emiliane T, and Jed A Fuhrman. 2012. "Seasonality and monthly dynamics of marine myovirus communities." Environ. Microbiol. 14 (8): 2171-2183.

Comeau, André M, and Henry M Krisch. 2008. "The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere." Mol. Biol. Evol. 25 (7): 1321-1332.

Danovaro, R., and M. Serresi. 2000. "Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean." Appl. Environ. Microbiol. 66 (5): 1857.

DeLong, Edward F. 2003. "Oceans of archaea." ASM News 69 (10): 503-503.

Dinsdale, E.A., R.A. Edwards, D. Hall, F. Angly, M. Breitbart, J.M. Brulc, M. Furlan, C. Desnues, M. Haynes, and L. Li. 2008. "Functional metagenomic profiling of nine biomes." Nature 452 (7187): 629-632.

Dorigo, U., S. Jacquet, and J.F. Humbert. 2004. "Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget." Appl. Environ. Microbiol. 70 (2): 1017.

Duhaime, Melissa B, Li Deng, Bonnie T Poulos, and Matthew B Sullivan. 2012. "Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method." Environ. Microbiol. 14 (9): 2526-2537.

Duhaime, Melissa B, and Matthew B Sullivan. 2012. "Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline." Virology 434 (2): 181-186.

Dwivedi, Bhakti, Robert Schmieder, Dawn B Goldsmith, Robert A Edwards, and Mya Breitbart. 2012. "PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity." BMC Bioinformatics 13 (1): 37.

ENA. "ENA Sequence Search." https://www.ebi.ac.uk/metagenomics/.

Engelhardt, Tim, Jens Kallmeyer, Heribert Cypionka, and Bert Engelen. 2014. "High virus-to-cell ratios indicate ongoing production of viruses in deep subsurface sediments." ISME J. 8 (7): 1503-1509.

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

Fouts, Derrick E. 2006. "Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences." Nucleic Acids Res. 34 (20): 5839-5851.

Frank, Jeremy A, Don Lorimer, Merry Youle, Pam Witte, Tim Craig, Jan Abendroth, Forest Rohwer, Robert A Edwards, Anca M Segall, and Alex B Burgin. 2013. "Structure and function of a cyanophage-encoded peptide deformylase." ISME J. 7 (6): 1150-1160.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics 28 (23): 3150-3152.

Fuhrman, Jed A. 1999. "Marine viruses and their biogeochemical and ecological effects." Nature 399 (6736): 541-548.

García-López, Rodrigo, Jorge Francisco Vázquez-Castellanos, and Andrés Moya. 2015. "Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations." Front. Bioeng. Biotechnol. 3.

Goldsmith, D.B., G. Crosti, B. Dwivedi, L.D. McDaniel, A. Varsani, C.A. Suttle, M.G. Weinbauer, R.A. Sandaa, and M. Breitbart. 2011. "Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity." Appl. Environ. Microbiol. 77 (21): 7730-7739.

Goldsmith, Dawn B, Rachel J Parsons, Damitu Beyene, Peter Salamon, and Mya Breitbart. 2015. "Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea." PeerJ. 3: e997.

Haynes, Matthew, and Forest Rohwer. 2011. "The human virome." In Metagenomics of the Human Body, 63-77. Springer.

Herskowitz, Ira, and David Hagen. 1980. "The lysis-lysogeny decision of phage lambda: explicit programming and responsiveness." Ann. Rev. Genet. 14 (1): 399-445.

Hobbie, J El, R Jasper Daley, and STTI977 Jasper. 1977. "Use of nuclepore filters for counting bacteria by fluorescence microscopy." Appl. Environ. Microbiol. 33 (5): 1225-1228.

Huang, Sijun, Steven W Wilhelm, Nianzhi Jiao, and Feng Chen. 2010. "Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences." ISME J. 4 (10): 1243-1251.

Ignacio-Espinoza, J Cesar , Sergei A Solonenko, and Matthew B Sullivan. 2013. "The global virome: Not as big as we thought?" Curr. Opin. Virol.: 566-571.

Jameson, Eleanor, Nicholas H Mann, Ian Joint, Christine Sambles, and Martin Mühling. 2011. "The diversity of cyanomyovirus populations along a North–South Atlantic Ocean transect." ISME J. 5 (11): 1713-1721.

Jover, Luis F, T Chad Effler, Alison Buchan, Steven W Wilhelm, and Joshua S Weitz. 2014. "The elemental composition of virus particles: implications for marine biogeochemical cycles." Nature Reviews Microbiology 12 (7): 519-528.

Kallmeyer, Jens, Robert Pockalny, Rishi Ram Adhikari, David C Smith, and Steven D'Hondt. 2012. "Global distribution of microbial abundance and biomass in subseafloor sediment." Proc. Natl. Acad. Sci. USA 109 (40): 16213-16216.

Kim, Kyoung-Ho, and Jin-Woo Bae. 2011. "Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses." Appl. Environ. Microbiol. 77 (21): 7663-7668.

Kim, Min-Soo, Eun-Jin Park, Seong Woon Roh, and Jin-Woo Bae. 2011. "Diversity and abundance of single-stranded DNA viruses in human feces." Appl. Environ. Microbiol. 77 (22): 8062-8070.

Kimura, Motoo. 1962. "On the probability of fixation of mutant genes in a population." Genetics 47 (6): 713.

Knowles, B, CB Silveira, BA Bailey, Barott. K, A Cantu, AG Cobián-Güemes, FH Coutinho, E Dinsdale, B Felts, KA Furby, EE George, KT Green, GB Gregoracci, AF Haas, JM Haggerty, ER Hester, NG Hisakawa, LW Kelly, YW Lim, M Little, A Luque, T McDole-Somera, K McNair, LS de Oliveira, SD Quistad, NL Robinett, E Sala, P Salamon, SE Sanchez, S Sandin, GGZ Silva, J Smith, C Sullivan, C Thompson, MJA Vermeij, M Youle, C Young, B Zgliczynski, R Brainard, RA Edwards, J Nulton, F Thompson, and F Rohwer. 2016. "Lytic to temperate switching of viral communities." Nature: in press.

Lang, Andrew S, Olga Zhaxybayeva, and J Thomas Beatty. 2012. "Gene transfer agents: phage-like elements of genetic exchange." Nat. Rev. Microbiol. 10 (7): 472-482.

Li, Weizhong, and Adam Godzik. 2006. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics 22 (13): 1658-1659.

Lindell, D., M.B. Sullivan, Z.I. Johnson, A.C. Tolonen, F. Rohwer, and S.W. Chisholm. 2004. "Transfer of photosynthesis genes to and from Prochlorococcus viruses." Proc. Natl. Acad. Sci. USA 101 (30): 11013.

McDaniel, Lauren D, Michéle DelaRosa, and John H Paul. 2006. "Temperate and lytic cyanophages from the Gulf of Mexico." J. Mar. Biol. Assoc. U.K. 86 (03): 517-527.

McNair, K., B.A. Bailey, and R.A. Edwards. 2012. "PHACTS, a computational approach to classifying the lifestyle of phages." Bioinformatics 28 (5): 614-618.

Meyer, Folker, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, and Andreas Wilke. 2008. "The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics 9 (1): 386.

Mokili, J.L., F. Rohwer, and B.E. Dutilh. 2012. "Metagenomics and future perspectives in virus discovery." Curr. Opin. Virol.

Ortmann, Alice C, and Curtis A Suttle. 2005. "High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality." Deep Sea Research Part I: Oceanographic Research Papers 52 (8): 1515-1527.

Pagarete, A, C-ET Chow, T Johannessen, JA Fuhrman, TF Thingstad, and RA Sandaa. 2013. "Strong seasonality and interannual recurrence in marine myovirus communities." Appl. Environ. Microbiol. 79 (20): 6253-6259.

Paul, J.H. 2008. "Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas?" ISME J. 2 (6): 579-589.

Paul, John H. 1999. "Microbial gene transfer: an ecological perspective." J. Mol. Microbiol. Biotechnol. 1 (1): 45-50.

Proctor, L.M., and J.A. Fuhrman. 1990. "Viral mortality of marine bacteria and cyanobacteria." Nature 343: 60-62.

R Core Team. 2013. "R: A language and environment for statistical computing." http://www.R-project.org/.

RefSeq. "NCBI Viral Genomes." http://www.ncbi.nlm.nih.gov/genome/viruses/.

Rocke, Emma, Maria G Pachiadaki, Alec Cobban, Elizabeth B Kujawinski, and Virginia P Edgcomb. 2015. "Protist community grazing on prokaryotic prey in deep ocean water masses." PLoS One 10 (4): e012450.

Rohatgi, Ankit. "WebPlotDigitizer v. 3.9." http://arohatgi.info/WebPlotDigitizer.

Rohwer, F, and R Edwards. 2002. "The phage proteomic tree: a genome-based taxonomy for phage." J. Bacteriol. 184 (16): 4529-4535.

Rohwer, F. 2003. "Global phage diversity." Cell 113 (2): 141-141.

Roux, Simon, Francois Enault, Viviane Ravet, Jonathan Colombet, Yvan Bettarel, Jean-Christophe Auguet, Thierry Bouvier, Soizick Lucas-Staat, Agnès Vellet, and David Prangishvili. 2016. "Analysis of metagenomic data reveals common features of halophilic viral communities across continents." Environ. Microbiol.

Roux, Simon, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, Didier Debroas, and François Enault. 2011. "Metavir: a web server dedicated to virome analysis." Bioinformatics 27 (21): 3074-3075.

Sandaa, Ruth-Anne, Laura Gómez-Consarnau, Jarone Pinhassi, Lasse Riemann, Andrea Malits, Markus G Weinbauer, Josep M Gasol, and T Frede Thingstad. 2009. "Viral control of bacterial biodiversity–evidence from a nutrient-enriched marine mesocosm experiment." Environ. Microbiol. 11 (10): 2585-2597.

Sano, E., S. Carlson, L. Wegley, and F. Rohwer. 2004. "Movement of viruses between biomes." Appl. Environ. Microbiol. 70 (10): 5842.

Schmieder, Robert, and Robert Edwards. 2011. "Quality control and preprocessing of metagenomic datasets." Bioinformatics 27 (6): 863-864.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised estimates for the number of human and bacteria cells in the body." bioRxiv: 036103. https://doi.org/http://dx.doi.org/10.1101/036103.

Short, C.M., and C.A. Suttle. 2005. "Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments." Appl. Environ. Microbiol. 71 (1): 480.

SILVA. "SSU Ref NR." http://www.arb-silva.de/projects/ssu-ref-nr/.

SMALT. "Sanger Institute." http://www.sanger.ac.uk/science/tools/smalt-0.

SRA. "NCBI Sequence Read Archive." http://www.ncbi.nlm.nih.gov/sra.

Staley, James T, and Allan Konopka. 1985. "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats." Ann. Rev. Microbiol. 39 (1): 321-346.

Steward, G.F., J.L. Montiel, and F. Azam. 2000. "Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments." Limnol. Oceanogr. 45: 1697-1706.

Sullivan, M.B., D. Lindell, J.A. Lee, L.R. Thompson, J.P. Bielawski, and S.W. Chisholm. 2006. "Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts." PLoS Biol. 4 (8): e234.

Suttle, C A. 2005. "Viruses in the sea." Nature 437: 356-361.

Suttle, Curtis A.. 2007. "Marine viruses — Major players in the global ecosystem." Nat. Rev. Microbiol. 5: 801-812.

Thingstad, T Frede. 2000. "Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems." Limnol. Oceanogr. 45 (6): 1320-1328.

Thingstad, T Frede, Bernadette Pree, Jarl Giske, and Selina Våge. 2015. "What difference does it make if viruses are strain-, rather than species-specific?" Front. Microbiol. 6.

Thingstad, TF, and R Lignell. 1997. "Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand." Aquat. Microb. Ecol. 13 (1): 19-27.

Torrella, Francisco, and Richard Y Morita. 1979. "Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications." Appl. Environ. Microbiol. 37 (4): 774-778.

Tseng, Ching-Hung, Pei-Wen Chiang, Fuh-Kwo Shiah, Yi-Lung Chen, Jia-Rong Liou, Ting-Chang Hsu, Suhinthan Maheswararajah, Isaam Saeed, Saman Halgamuge, and Sen-Lin Tang. 2013. "Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances." ISME J. 7 (12): 2374-2386.

Tuma, Rabiya S, Matthew P Beaudet, Xiaokui Jin, Laurie J Jones, Ching-Ying Cheung, Stephen Yue, and Victoria L Singer. 1999. "Characterization of SYBR Gold nucleic acid gel stain: a dye optimized for use with 300-nm ultraviolet transilluminators." Anal. Biochem. 268 (2): 278-288.

UNIVEC. "NCBI Univec Database." http://www.ncbi.nlm.nih.gov/tools/vecscreen//univec/.

Watkins, Siobhan C, Neil Kuehnle, C Anthony Ruggeri, Kema Malki, Katherine Bruder, Jinan Elayyan, Kristina Damisch, Naushin Vahora, Paul O'Malley, and Brieanne Ruggles-Sage. 2015. "Assessment of a metaviromic dataset generated from nearshore Lake Michigan." Mar. Freshw. Res.

Weinbauer, M.G. 2004. "Ecology of prokaryotic viruses." FEMS Microbiol. Rev. 28 (2): 127-181.

Weitz, Joshua S, Charles A Stock, Steven W Wilhelm, Lydia Bourouiba, Maureen L Coleman, Alison Buchan, Michael J Follows, Jed A Fuhrman, Luis F Jover, and Jay T Lennon. 2015. "A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes." ISME J.

Whitman, William B., David C. Coleman, and William J. Wiebe. 1998. "Prokaryotes: The unseen majority." Proc. Natl. Acad. Sci. USA 95: 6578-6583.

Wilke, Andreas, Jared Bischof, Travis Harrison, Tom Brettin, Mark D'Souza, Wolfgang Gerlach, Hunter Matthews, Tobias Paczian, Jared Wilkening, and Elizabeth M Glass. 2015. "A RESTful API for accessing microbial community data for MG-RAST." PLoS Comput Biol 11 (1): e1004008.

Williamson, K.E., M. Radosevich, and K.E. Wommack. 2005. "Abundance and diversity of viruses in six Delaware soils." Appl. Environ. Microbiol. 71 (6): 3119.

Wommack, K.E., and R.R. Colwell. 2000. "Virioplankton: viruses in aquatic ecosystems." Microbiol. Mol. Biol. Rev. 64 (1): 69-114.

Youle, Merry, Matthew Haynes, and Forest Rohwer. 2012. "Scratching the surface of biology's dark matter." In Viruses: Essential agents of life, 61-81. Springer.

Zhong, Y., F. Chen, S.W. Wilhelm, L. Poorvin, and R.E. Hodson. 2002. "Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20." Appl. Environ. Microbiol. 68 (4): 1576.

Zhou, You, Yongjie Liang, Karlene H Lynch, Jonathan J Dennis, and David S Wishart. 2011. "PHAST: a fast phage search tool." Nucleic Acids Research.

Appendix for Chapter 1

Supplemental methods

Global VLP calculation

A total of 6,154 VMR measurements were collected from 53 research papers (Supplemental Table 2 in Cobian et al. 2016). When specific VMR values were not provided in the manuscript, the VMRs were extracted from the published plots using Web Plot Digitizer (Rohatgi). Each VMR value was assigned to one of 7 major biomes (marine, freshwater, other aquatic, sediments, soil, human-associated, and other host-associated) and both the mean and median VMRs were calculated for each biome. Boxplots for each biome VMR were generated using R (R Core Team 2013). Sources for the number of prokaryotic cells are noted on Figure 1. For each biome, the number of VLPs was estimated by multiplying the number of prokaryotic cells by the calculated median VMR. The global VMR was calculated by summing the VLPs for all 7 biomes and then dividing by the sum of the number of prokaryotic cells.

Virome collection

Virome FASTA sequences were obtained from MGRAST (Meyer et al. 2008), MetaVir (Roux et al. 2011), iVIRUS (Bolduc et al.), SRA (SRA), and ENA (ENA). Since there is overlap between these databases, virome duplicates were removed by manual curation. Only viromes accompanied by a peer-reviewed paper were included. For the large Tara Ocean Viromes dataset, one sequencing run per site was selected, of which a 1% subsample without replacement was used for analysis. FASTA sequences from MGRAST were downloaded after quality filtering using MGRAST API (Wilke et al. 2015); those from

SRA and ENA were quality filtered with PRINSEQ (Schmieder and Edwards 2011) (99% quality); raw FASTA files were used from MetaVir and iVIRUS. Six viromes (metavir_2726, metavir_2727, MGRAST: 4519681, 4519682, 4519683, 4519684) had been spiked with PhiX control DNA; those introduced sequences were eliminated from the FASTA files using in-house scripts (SMALT mapping at 95% identity to the ϕX174 genome; Genbank accession J02482.1). When viromes had been sequenced as paired-end, only one read was used for subsequent analysis. Of the 1,622 viromes collected, 1,615 were assigned to one of the 7 major biomes (marine, freshwater, other aquatic, sediments, soil, human-associated, and other host-associated). The remaining 7 viromes were classified as 'other' (fermented food and air viromes). The available metadata and major biome classification for each virome are provided in Supplemental Table 3.

Reference library creation

Each virome was assembled de novo using SPAdes (Bankevich et al. 2012) and all contigs ≥1000 nt from all viromes were merged into a single file for further analysis. A prokaryotic virus reference database was created from 2,699 bacteriophage genomes and 67 archaeal virus genomes from the NCBI RefSeq database (January 2016) (RefSeq), plus an additional 123 bacteriophage genomes from the Broad Institute Marine Phage Sequencing Project (Broad Institute). The virome contig file and the prokaryotic virus reference database were merged into a single file and clustered using CD-HIT-EST (Fu et al. 2012) at 98% identity. BLASTn (e-value cutoff of 0.001) was used to identify cloning vectors (UNIVEC) and rRNA (SILVA) sequences (using SSU Ref NR and LSURef_123_10_07_15). Removal of these sequences left a reference library containing 2,258,219 phage (sensu lato) contigs.

Size of the global virome

All assembled contigs <1000 nt were merged into a single file and clustered at 98% identity using CD-HIT-EST to generate the small_contigs_98 database. The lengths of every sequence in this database, and in the reference library were obtained and summed to yield the size of the global virome in nucleotides (1.29×10^{10}). This total was divided by 50 kbp, the assumed average phage genome size, to yield the number of viral genotypes on Earth.

Virome mapping to the reference library

All viromes were mapped individually to the reference library using SMALT (SMALT) in 3 separate runs using 99%, 95%, and 90% alignment identity cutoff values. The mapping at 99% alignment identity was used when calculating both observed and predicted viral genotypes.

In addition, a separate global analysis was performed by first creating a pool of all virome FASTA sequences for each of the 5 biomes with the most VLPs. A pooled global virome (1,000,000 reads) was then generated by subsampling with replacement from each biome pool, with the percentage contributed by each biome being the percentage of global VLPs present in that biome (i.e., 26,857 marine reads, 37 freshwater, 1,525 other aquatic, 870,833 sediments, and 101,667 soil). This pooled global virome was mapped to the reference library as above.

Fractional abundances of viral genotypes

Mapping as described above yielded the number of hits to each genotype in the reference library for each virome. The fractional abundance (*f*) of each viral genotype in each virome was then calculated as:

Equation 1.1 Fractional abundance of viral genotypes

$$f(i) \cong \frac{r(i)}{T(j)} * \frac{L(mean)}{L(i)}$$

where

$f(i)$: the fractional abundance of contig *i* in this virome

$r(i)$: the number of reads that map to contig *i*

$T(j)$: the total number of reads in virome *j*

$L(mean)$: the mean genome length (bp), assumed to be 50,000 bp (Steward, Montiel, and Azam 2000)

$L(i)$: the length (bp) of contig *i*

Then the fractional abundance of contig *i* in a biome was calculated as:

Equation 1.2 Fractional abundance of contigs

$$f(i_{biome}) \cong \frac{f(i)}{N}$$

where

$f(i_{biome})$: the fractional abundance of contig *i* in the biome

N: number of viromes in biome

Parallel calculations were made to calculate the fractional abundance of contig *i* in the global virome sample. Rank abundance plots for each biome and the global virome sample were generated using R.

Annotation

The 10 viral genotypes with the greatest fractional abundance in each biome and in the pooled global virome were annotated through online NCBI BLASTn against the nr/nt database.

Observed viral genotypes

For each biome and the pooled global virome, the sum of the lengths of every contig with a fractional abundance at 99% identity > 0 was divided by 50,000 (the assumed average phage genome length) to estimate the number of observed viral genotypes.

Predicted viral genotypes

Curve fitting was performed for the 99% identity rank abundance plots for each biome using the Matlab Curve Fitting Toolbox™ R2015b (Mathworks, Inc.). Several curve fit models were evaluated when applied to the first 100, 1,000, and 10,000 fractional abundance ranks for each sample (Supplemental Table 5 in Cobian et al. 2016). The power law model gave the best fit and was therefore used to calculate the predicted number of viral genotypes for each biome. Estimated a and b for each curve were used to calculate the number of VLPs on each rank of the curve as follow:

Equation 1.3 Estimated a and b for rank abundance curves

$$V_i = ai^b \times V_{total}$$

where V_i is the number of VLPs on rank i, and V_{total} is the total number of VLPs in the biome (from Table 1.1). The values for V_i were summed until the following condition was met:

Equation 1.4 Number of viral genotypes prediction

$$V_{total} = \sum_{i=1}^{x} ai^b \times V_{total}.$$

At that point, the sum was considered to be the predicted number of viral genotypes. For the soil virome, the curve was asymptotic, and *x* was not calculated.

Chapter 2: Fragment Recruitment Assembly Purification (FRAP): Put bioinformatics back into the hands of biologists.

Abstract

A strict or exact match strategy for the analysis of metagenomes is presented as Fragment Recruitment Assembly Purification. This strategy reduces the problem of metagenomic assignments to text matching using strict or exact hits from a dataset to a database. Such databases are constructed by the user and tailored to adrress specific biological questions. With the increasing number of available genomes, and the high number of reads generated by current sequencing technologies, unreliable assignments can be eliminated from the analysis. Reliable assignments can be identified by using only strict and exact matches, such strategy limits the occurrence of false positives. Here the implementation of Fragment Recruitment Assembly Purification is presented as well as a set of auxiliary tools that allow the user to quickly build heatmaps, coverage plots, and fragment recruitment plots; and then use these data to make biological inferences. The performance of the presented algorithm was tested in aquatic, and host associated metagenomes. Fragment Recruitment Assembly Purification was used for the exploration of The Global Virome (Chapter 1) and in the analysis of Cystic Fibrosis metagenomes and viromes (Chapter 4).

Introduction

Current bioinformatics

Current bioinformatic methods have become a convoluted land for biologists. In metagenomic studies, once next generation sequencing (NGS) data (Singer et al. 2016) is generated, there are a myriad of options (Fonseca et al. 2012; Naccache et al. 2014; Huttenhower 2019; Huson et al. 2016) and opinions (Greninger 2018) about the best way to analyze (Sczyrba et al. 2017; 2019, n.d.; McIntyre et al. 2017) such data.

Bioinformatic metodologies to compare a dataset (i.e. a metagenome) to a database (i.e. viral reference genomes) rely heavily on statistics in calculating the probability that a DNA fragment originated from a given organism (reference genome). Variability in the results of metagenomic analyses using different methods and pipelines is embedded in several steps such as: 1) the scoring systems used for each application (i.e. BLAST family algorithms developed by Altschul et al. 1990), 2) the construction of k-mer profiles with different databases, k-mer sizes and inference methods (i.e. KAIJU(Menzel, Ng, and Krogh 2016) and FOCUS2 (Silva, Dutilh, and Edwards 2016)), and 3) the use of different sets of marker genes for taxonomical inferences (i. e. MetaPhlan2, MetaVir (Roux et al. 2011)). Such variability in the search methods can result in distinct biological interpretations when comparing the analysis of the same dataset through different software or platforms.

To overcome these issues, Fragment Recruitment Assembly Purification (FRAP) was developped, as a simple and reproducible method to compare a dataset to a database through strict or exact hits. FRAP assign strict (96% identity over 100% of the read) or exact hits

(100% identity over 100% of the read) in the dataset to a custom database specific to the biological problem at hand.

Why use FRAP?

FRAP aims to be a simple approach to bioinformatics with a fast learning curve. Exact hits are reliable and reproducible, which is defined as 100% of the read matching to the database with 100% identity. As the probability of having a specific 100 nucleotides sequence by chance is $1/4^{100}$, it is hard to argue that an exact hit is not a true positive. The exact hits approach is possible because the amount of sequences in the databases and in the datasets is large enough so that non perfect matches can be ignored or used for separate analysis. Databases and datasets size will continue to increase(Stephens et al. 2015), so much so that eventually we will have the Earth's metagenome. As the databases become bigger, the more FRAP will become relevant.

The FRAP approach has an adaptable design in which datasets and databases are any collection of sequences in fasta format. This makes the strategy directed towards answering biological questions and researchers can easily construct databases relevant to their questions. FRAP reduces metagenomic assignments to text strings comparisons which have been implemented in several operating systems (such as Unix) as low level functions. This allows FRAP to be a "never break" pipeline which should be stable across platforms and versions. String matching algorithms have at worst lineal time complexity(Pandiselvam, Marimuthu, and Lawrance 2014). Regardless of the method used to obtain exact hits, FRAP results should always be the same.

How to use FRAP?

Databases and datasets are nucleotide fasta files. Datasets are usually short DNA fragments (reads) from metagenomes, metatrascriptomes or viromes. Databases are usually a set of genomes of interest such as bacteria representative genomes, archaea reference sequences, and viral reference sequences; or a set of genes of interest such as bacteria metabolic genes (i.e. the SEED (Overbeek et al. 2014) database), virulence factors (Sayers et al. 2019; Chen et al. 2016), or antibiotic resistance genes (Jia et al. 2017). These are popular examples, but a dataset and a database can be any collection of sequences in a fasta file format.

Several exact hit implementations were explored in this study, such as the use of mapping algorithms with adjusted parameters to obtain exact hits such as smalt ("SMALT Manual" 2010) and HISAT (Kim, Langmead, and Salzberg 2015); the implementation of a search via a hash table, called hash exact hit (heh), and the use of the Unix command grep.

FRAP has two main applications: basic FRAP and blind FRAP. Basic FRAP compares a set of reads to a set of known reference genomes or genes. Examples include comparing aquarium metagenomes to bacteria and archaea representative genomes and comparing lung metagenomes to virulence factors genes and antibiotic resistance genes. In basic FRAP the reported fractional abundances are interpreted as taxonomical or functional assignments. Basic FRAP can also be used to filter out or retrieve reads present in a database of interest, for example to filter out the human genome reads present in a lung metagenome and keep the remaining reads for further analysis, or to retrieve the viral reads in a lung metagenome and assemble all of them together to discover viral like contigs.

Furthermore, basic FRAP aids in the detection of contamination in samples or in databases. For example, a marine bacteria was detected in a lung metagenome. Upon closer inspection, all the detected reads map to a single short region of the marine bacteria genome, in this case the sample was contaminated with a marine bacteria PCR product. Contamination in the databases can be detected, for example a human biopsy sample had exact hits to a bacteria and all hits were in a very short region, further exploration led to the identification of a human genome region in the assembly of the bacteria in the reference database.

Blind FRAP compares a dataset to itself via contigs or reads. In blind FRAP the database has no identified function or taxa and the databases are usually contigs assembled from the dataset or contigs assembled from another dataset. One example of blind FRAP is presented in Chapter 1, where contigs assembled from viromes were used as a database and reads from the same viromes were used as datasets. The resulting fractional abundances for each contig represent the abundance of viral-like elements that have no functional or taxonomical annotation yet. A second example of blind FRAP is obtaining viromes and metagenomes from the same sample and using the contigs from a virome as database and the reads from metagenomes as datasets, in this case the fractional abundances of virome derived contigs is interpreted as the presence of viral-like elements in metagenomes.

Challenges in the development of FRAP

FRAP needs further development to become portable and fast. ideally FRAP should be easy to use, not requiring the use of the terminal and where dataset and database files can be dragged and dropped into the analysis window.

Computational challenges in the development of FRAP are: 1) The development of efficient search algorithms (So 2017) for text search which is currently limited by database size, and 2) Determine the amount of computational power needed for large scale analysis, for example, how much computing power do we need to use the Earth's metagenome?

Biological challenges in the development of FRAP rely on the compilation of reliable databases that are vetted for reproducible results.

Results and Discussion

FRAP was used in several projects. FRAP generalized pipeline (Figure 2.1) inputs are a single fasta file containing a database and one or multiple fasta files containing the reads in a dataset or datasets. The implemented methods to obtain exact hits are smalt at 100% identity, HISAT at 100% identity, hash exact hit, and grep. FRAP outputs are a tab delimited file containing the fractional abundances of each element in the database in each dataset, a tab delimited file containing the number of hits to each element in the database in each dataset and fragment recruitment plots for the 10 database elements with the highest fractional abundances. In addition, frap-tools can be used to obtain a fasta file containing the reads that have hits to the database (si_hits.pl) or a fasta file containing the reads that have no hits to the database (no_hits.pl). FRAP can be accessed through the GitHub repository https://github.com/yinacobian/frap

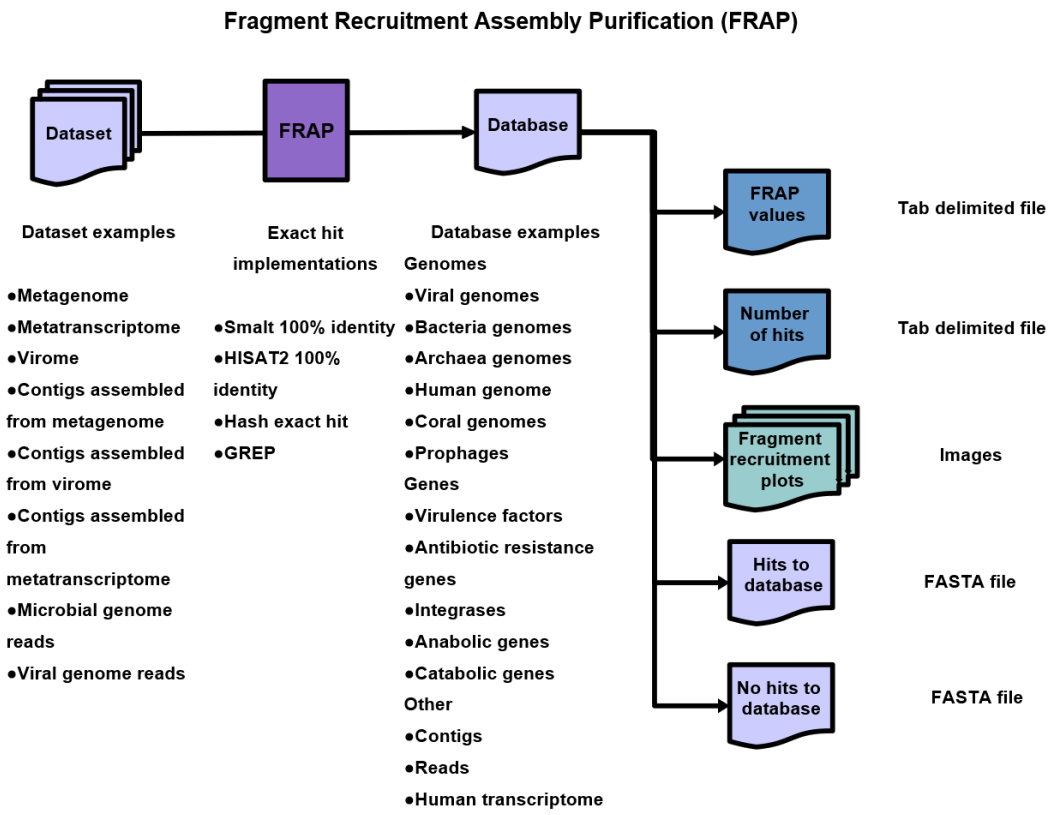


Figure 2.1 Fragment Recruitment Assembly Purification general case.

Basic FRAP

The simplest implementation of FRAP is the "just map FRAP" approach (Figure 2.2-A). FRAP normalization to obtain fractional abundances from exact hits (Figure 2.2-B) is based on the principle that datasets have different sizes and the length of each element in the database is different. The fractional abundance of each element in the database (*f(i)*) is calculated by dividing the number of hits to an element in the database (*r(i)*) by the total number of sequences in the dataset (*T(j)*), next this term is multiplied by the mean length of

the elements in the dataset (*L(mean)*) divided by the length of the database element (*L(i)*). Fractional abundances are then scaled by a factor of 1,000,000.

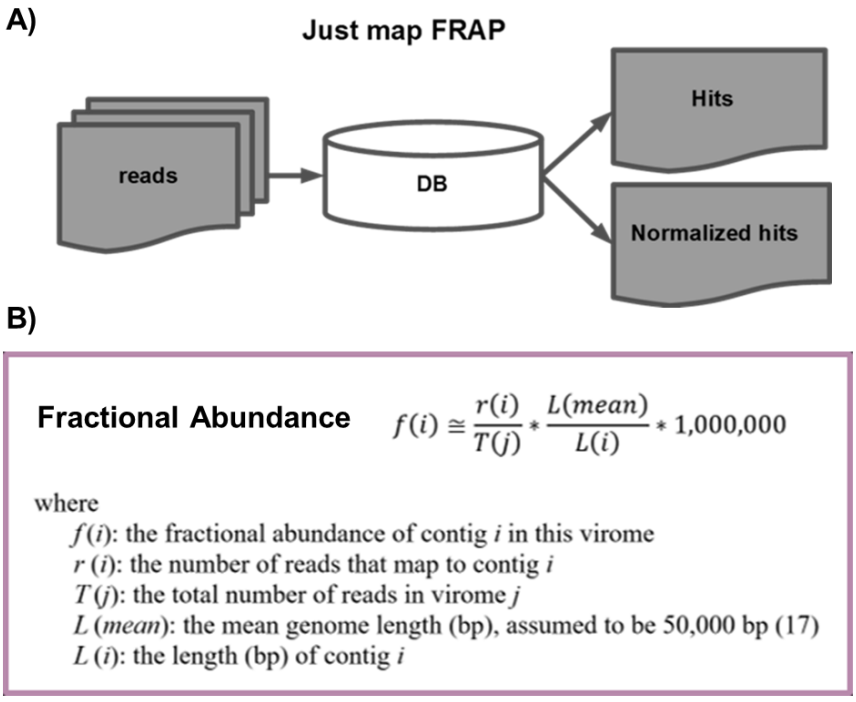


Figure 2.2 FRAP basic concept and normalization. A) FRAP implementation as "Just map FRAP" which is implemented in the script jmf.pl B) FRAP normalization to fractional abundances

L(mean) values represent the average genome length. When the database is composed of complete genomes, the *L(mean)* value can be pre-determined and used for the calculations. Microbial average genome length was calculated from publicly available complete genomes included in RefSeq-release80 (Table 2.1). Based on this set of genomes, virus average genome length is 29,936 bp, bacteria average genome length is 4,028,000 bp and archaea average genome length is 2,730,000 bp. For fungi and protozoa, the average genome length

was calculated using locus length, which can be the complete genome, or a chromosome (Table 2.2). Subsequent calculations are needed to get accurate average genome lengths for this organisms. *L(mean)* values represent the average gene length when the database is composed of genes. The average gene length for virus, bacteria, archaea, fungi and protozoa was calculated (Table 2.3). The average gene length for virus is 741 bp, for bacteria 899 bp, for archaea 855 bp, for fungi 1,660 bp and for protozoa 1,798 bp.

Table 2.1 Average genome length for FRAP normalization. RefSeq-release80 published on January 9, 2017. Genomes/ASSEMBLY_REPORTS/ published on March 6, 2017. Virus: obtained from all viral genomes available at refseq. Bacteria: obtained from assembly reports, the average genome length per taxid group was used as input, 67,706 genomes are distributed in 964 taxid groups. Archaea: obtained from assembly reports, the average genome length per taxid group was used as input, 192 genomes are distributed in 11 taxid groups. Information not available for fungi and protozoa, everything is together in an eukaryotes genome size file.

	GENOME LENGTH						
	N	**minimum**	**1st quartile**	**median**	**mean**	**3rd quartile**	**maximum**
Virus	8,321	200	2,737	7,233	29,936	38,124	2,473,870
Bacteria	67,706	162,600	2,425,000	3,912,000	4,028,000	5,193,000	11,930,000
Archaea	192	174,700	2,150,000	2,486,000	2,730,000	3,046,000	4,568,000

Table 2.2 Average locus length for FRAP normalization. Locus length for L-mean in FRAP normalization. RefSeq-release80 published on January 9, 2017. A LOCUS is a refseq entry, in the case of viruses one locus is one genome, for the rest of the groups a locus can be either a compelte genome, a plasmid, a contig or any refseq entry.

LOCUS LENGTH

	N	minimum	1st quartile	median	mean	3rd quartile	maximum
Virus	8,321	200	2,737	7,233	29,936	38,124	2,473,870
Bacteria	8,067,030	16	886	4,656	36,891	25,351	14,782,125
Archaea	20,952	4	2,851	11,354	30,810	32,179	4,064,496
Fungi	74,670	86	664	2,226	89,939	14,047	11,880,248
Protozoa	247,163	22	922	1,158	14,559	2,514	13,391,543

Table 2.3 Average gene length for FRAP normalization. [1]min represent the start site of regulatory regions in the database, therefore it does not represent an accurate minimum gene length.

GENE LENGTH

	N	min[1]	1st quartile	median	mean	3rd quartile	maximum
Virus	308,153	1	260	443	747	860	262,387
Bacteria	291,916,012	1	437	773	899	1,179	110,417
Archaea	1,638,741	1	419	731	855	1,124	32,447
Fungi	2,195,385	3	857	1,394	1,660	2,100	186,159
Protozoa	976,901	3	649	1,221	1,798	2,168	131,440

FRAP-tools are a set of auxiliary scripts to visualize and generate fasta files from jmf.pl outputs. Heatmaps are generated from fractional abundances (Figure 2.3). Fragment recruitment plots (Figure 2.4) are generated for the 10 elements in the database with the highest fractional abundances. In the case of complete genomes as a database, the confidence of the identification of a bacteria genome in the dataset is assessed by the distribution of fragments across most of the genome.

A basic FRAP approach was used in a set of metagenomes from an aquarium water column microbial community establishment process, in which 47 samples were obtained over a 6 months period. The databases used were bacteria representative genomes (Figure 2.3-A) and archaea representative genomes (Figure 2.3-B). The dominant bacteria during the microbial community establishment was *Phaeobacter galleciensis*, whose hits are distributed across the whole genome (Figure 2.4) which provides evidence of the presence of the organism in the system. Results from this study will be further discussed and published by Calhoun S. et. al.

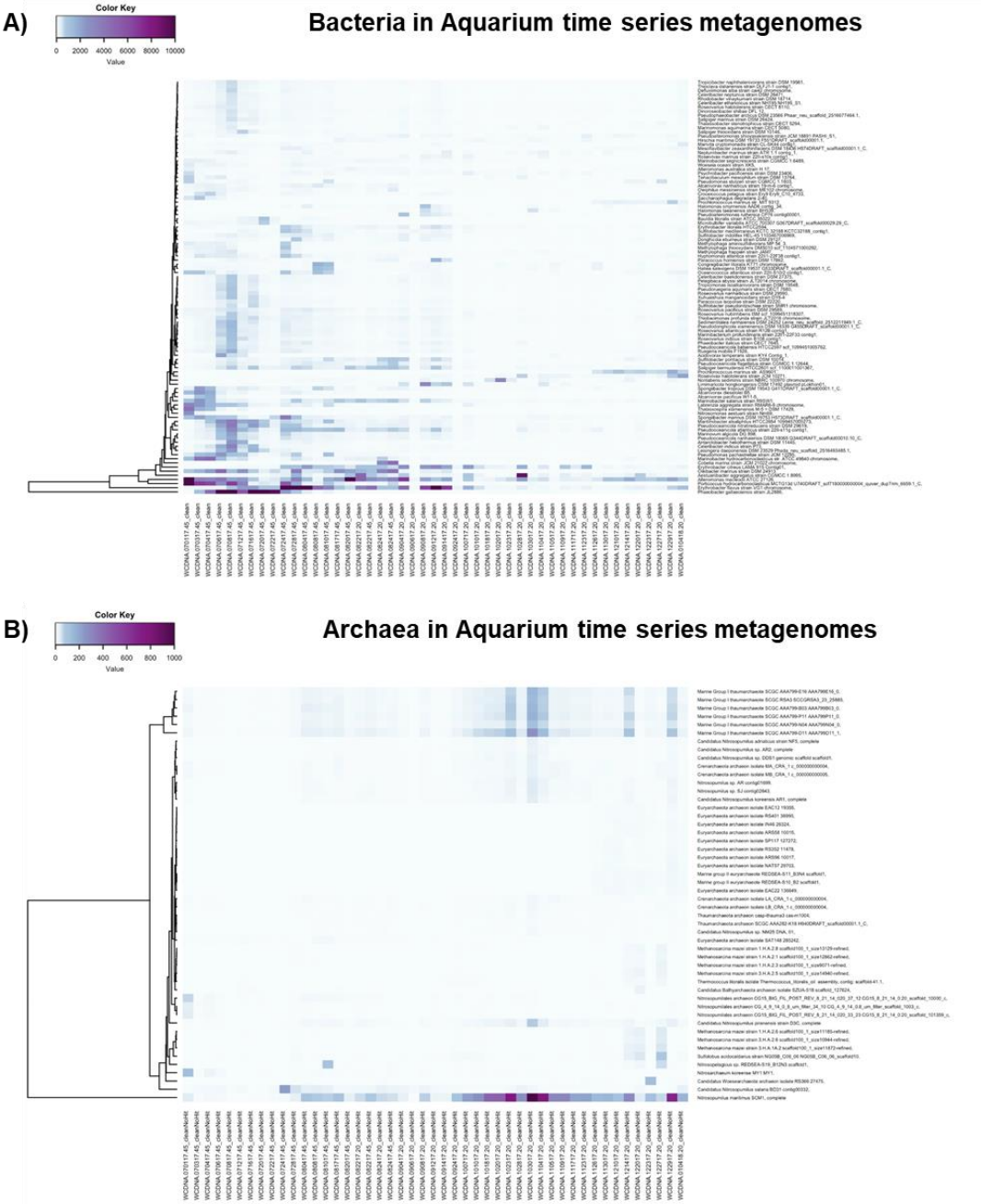


Figure 2.3 Heatmap of the 100 most abundant bacteria in the aquarium metagenomes. FRAP at 96% identity. A) Heatmap of the most abundant bacteria found in the aquarium metagenomes. B) Heatmap of the most abundant archaea in the aquarium metagenomes. *Propionibacterium acnes* was manually removed from the analysis since it is not a marine bacteria and is a possible contamination.

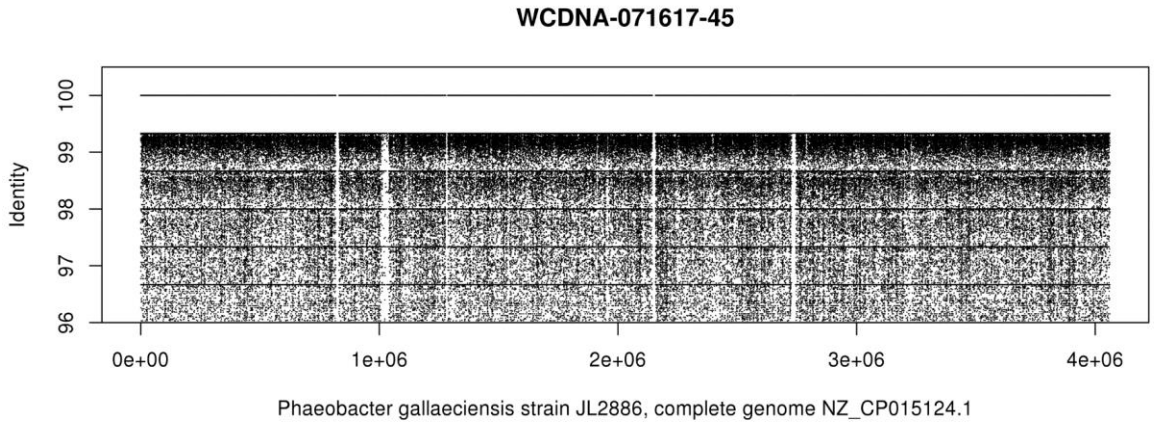


Figure 2.4 Fragment recruitment plot of *Phaeobacter gallaciensis*, the most abundant bacteria in the aquarium metagenomes.

A basic FRAP approach was used in the analysis of CF associated microbial metagenomes obtained from sputum samples of acute exacerbations. These results are discussed in detail in Chapter 4 of this dissertation. Human DNA in the sputum metagenomes was removed using FRAP and FRAP-tools, in which the remaining sequences were compared to a set of bacteria reference genomes. In a sample from patient CF418, the microbial community was dominated by *Achromobacter xylosoxidans*, which is the closest reference genome and evidence for its presence in the clinical sample is supported by hits across the genome. In Chapter 4 of this dissertation a detailed analysis of the genome deletions supports the hypothesis that the genome deletions are phage regions which may be excising from the genome. In clinical metagenomic studies is essential to use exact hits with curated reference databases so clinicians receive complete information that they can trust for trearment consideration.

Visual inspection of fragment recruitment plots allows the identification of contaminants in the dataset or in the databases. In a sputum metagenome from patient CF01, sample contamination was detected when for three bacteria genomes hits in a 1,500 nt region were identified (Figure 2.5). Such bacteria were *Delftia acidovorans, Bordetella petrii*, and *Marinomonas aquamarina* and the identified regions are the result of amplicons amplification of such regions. Contamination in an assembled reference bacteria genome present in the bacteria representative genomes database was identified when a metatranscriptome from a human lung biopsy had hits to a small region of *Microbacterium barkeri* (Figure 2.6), in this case the region in the bacteria reference genome had a human origin, which implies it was contaminated during the sequencing and assembly process. It is important to identify and flag this reference genomes to avoid further misassignments.

The reproducibility of FRAP was assessed using replicates of subsamples with replacement from a human plasma metagenome spiked with 10^4 copies of human immunodeficiency virus (Naccache et al. 2014). Five subsamples with replacement were compared to viral refseq (Supplemental Table 4.1), overall the fractional abundances of viruses in each replicates have low variability. The mean standard deviation from all datasets was 0.0109, which means the replicates show reproducibility.

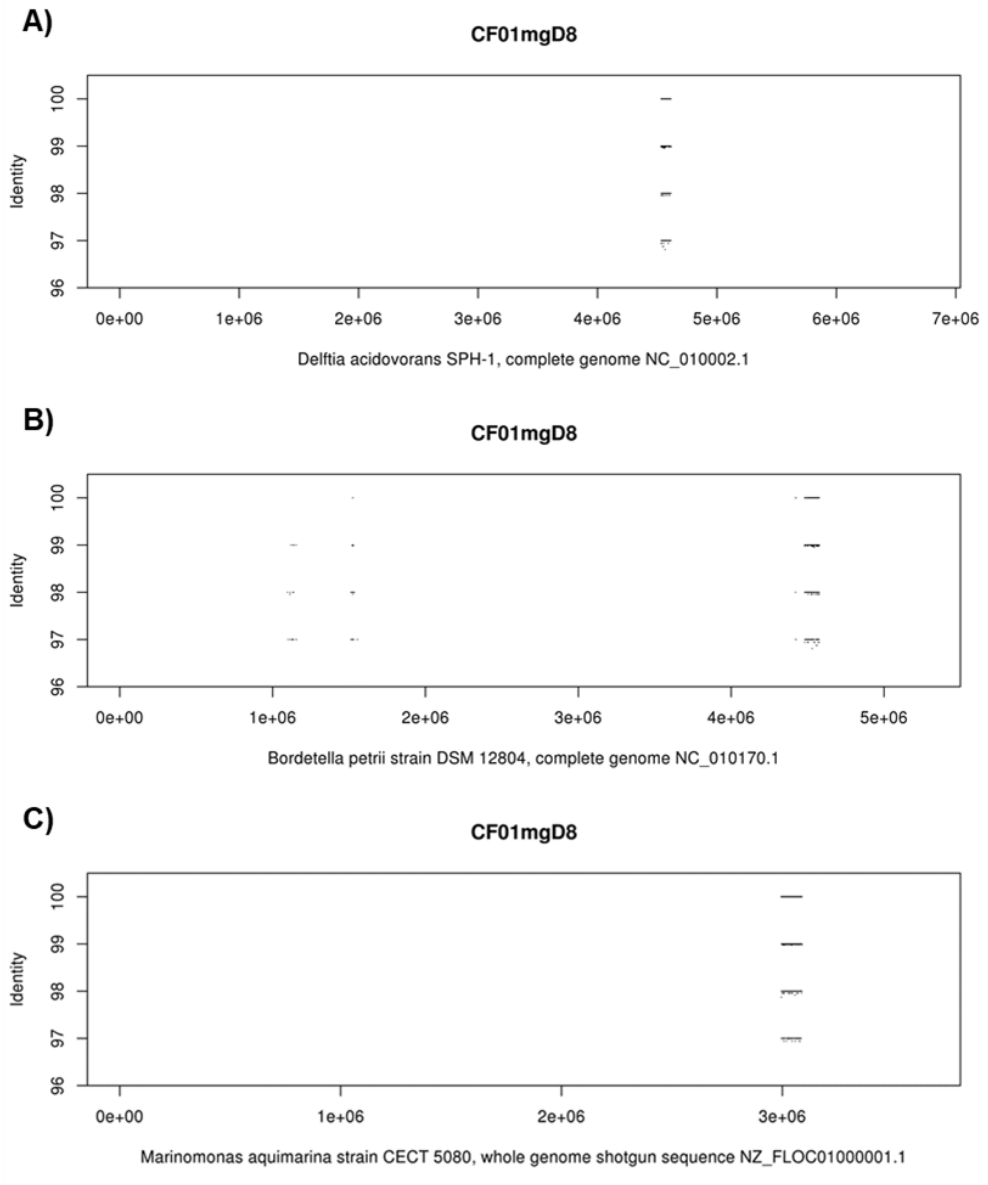


Figure 2.5 Contamination detection in CF samples. The most abundant bacteria found in CF01 were *Delftia acidovorans*, *Bordetella petrii* and *Marinomonas aquimarina*. In these 3 bacteria, all the hits mapped in one isolated position, supporting the idea of a sample contamination.

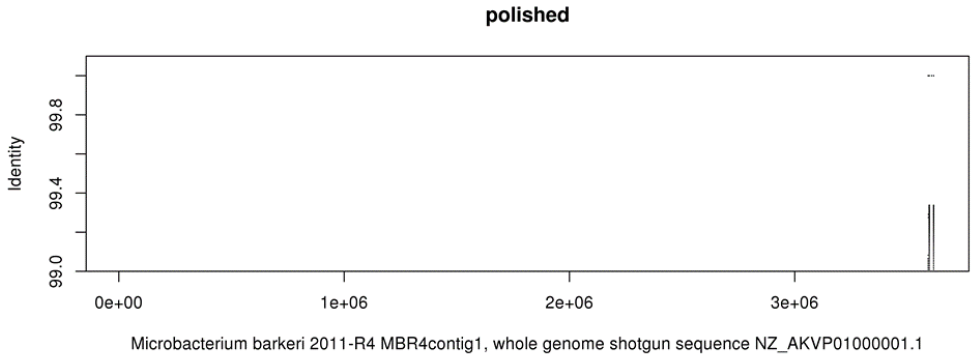


Figure 2.6 Genome contamination detection. Recruitment plot of the most abundant bacteria (*Microbacterium barkeri*) in a cancer biopsy sample. All the 2,295,340 reads map in the same position.

Blind FRAP

The second family of FRAP uses is the blind FRAP strategy. Its basic principle is the use of a database that does not have taxonomical or functional assignments to obtain fractional abundances for each contig in the datasets. The database are contigs assembled either from the dataset or an extended dataset. Functional assignments are obtained for the contigs with highest fractional abundances though distant evolutionary relationships assignments such as tBLASTx.

A blind FRAP strategy was used in the global virome calculations presented in Chapter 1 of this dissertation. In that study a global virome database was constructed using available public viromes from several biomes. Datasets were viromes in each biome, the result was the fractional abundances of contigs per biome. The contigs with highest fractional abundance on each biome were annotated using tBLASTx, those contigs represent new diversity which has not yet been described.

A second blind FRAP strategy is proposed for the study of pairs of metagenomes and viromes from the same sample (Figure 2.7). The rationale behind this strategy is to estimate the fractional abundances of viral-derived elements in metagenomes, which otherwise would be missed since they are unrepresented viruses that are not yet in reference databases. Viromes to metagenomes FRAP was used in a test dataset of coral and algae interactions (Figure 2.8). The samples are punches from a coral-algae interaction transect in which samples A and B are coral punches, sample C is the interaction zone between coral and algae and samples D and E are algae tissue. Metagenomes were obtained from all samples and viromes were obtained from samples A, B and C. Hits to viral like elements derived from coral punches were identified in all the coral derived samples and in the interaction zone, in algae samples, two contigs had hits to the coral-derived viral elements. This data will be explored in further detail and published by Little M. et. al.

A blind FRAP strategy in which reads and contigs originated from the same sample are compared to each other is proposed. In this case fractional abundance of contigs can be obtained from the samples themselves. This strategy was applied to two cheese rinds viromes. Viromes from Winnimere cheese and Bailey cheese rinds were obtained and contigs were assembled and used as a database, the reads were used as dataset. The percentage of exact hit reads to the contigs database was 77.5 % for Winnimere and 63.8% for Bailey (Table 2.4). From these estimates we can calculate false negatives as the number of fragments that don't recruit to the assembly, which is 23% for Winnimere and 36% for Bailey. An assessment of how accurate the exact hit search algorithm is to compare reads to reads. In the case of cheese

rind viromes, the search strategies are good, with 99.8% reads recruited back to reads for Winnimere and 99.98% for Bailey (Table 2.5).

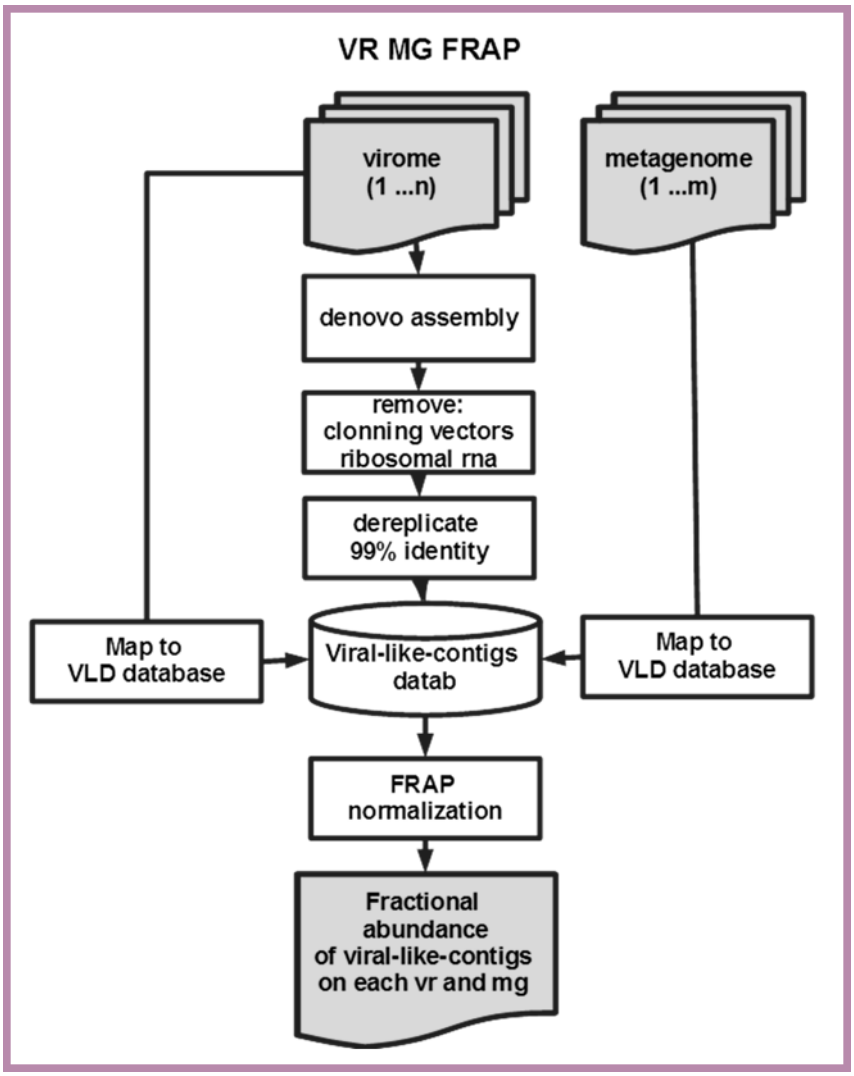


Figure 2.7 Generalized FRAP pipeline to map metagenomes to viral assembled contigs.

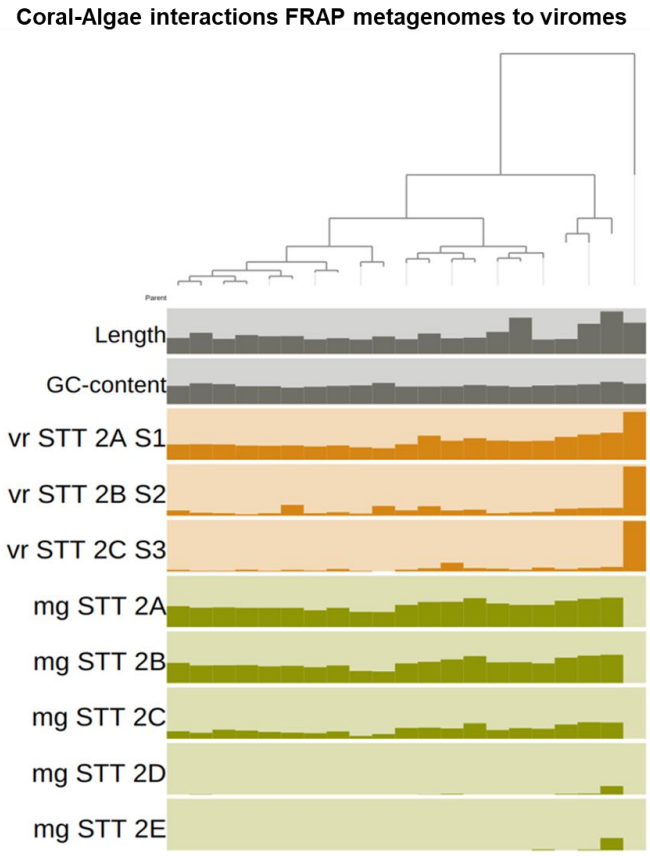


Figure 2.8 FRAP metagenomes to viromes in coral-algae interactions. Samples 2A and 2B are coral punches, sample 2C is a punch of the coral-algae interaction zone, samples 2D and 2E are algae punches. Viromes were obtained for samples 2A, 2B and 2C. Metagenomes were obtained for all samples. Contigs fractional abundances were visualized in Anvi'o (Eren et al. 2015).

Table 2.4 Exact hit, using smalt at 100% identity. Dataset: polished reads, q-phred>30 (99.9% base call accuracy), no low complexity (Shannon entropy 0.5) single end. Database: all assembled contigs (spades --only-assembler) from single end.

Exact hit reads to contigs

	Winnimere cheese virome	Bailey cheese virome
number of reads	3,256,476	1,714,998
number of nucleotides	497,990,024	262,232,327
number of contigs	30,484	23,315
number of nucleotides in contigs	26,541,438	13,239,529
number of reads that map to contigs	2,523,962	1,095,563
number of contigs with no reads mapped back	533	1,253
percentage of reads that map back to contigs	**77.51**	**63.88**
percentage of contigs with no reads	1.75	5.37

Table 2.5 Exact hit, using smalt at 100% identity. Dataset: polished reads, q-phred>30 (99.9% base call accuracy), no low complexity (Shannon entropy 0.5) single end. Database: polished reads, q-phred>30 (99.9% base call accuracy), no low complexity (Shannon entropy 0.5) single end.

Exact hit reads to reads

	Winnimere cheese virome	Bailey cheese virome
# reads	3,256,476	1,714,998
# reads that map to reads	3,249,940	1,714,714
% reads map back to reads	**99.80**	**99.98**

The future of FRAP

A third family of FRAP uses is proposed, in which two or more sample groups are compared using blind FRAP, then the fractional abundances are used as input for machine learning algorithms such as random forests and a group of contigs that better differentiate among samples is identified. This strategy needs further exploration.

FRAP aims to enable easy, robust and meaningful bioinformatics to accelerate biological discoveries. To get to this point, FRAP needs to be adapted as an easy to use platform and optimization of exact hits search are needed to be able to use large databases and datasets and obtain results in a short time.

Biological insights from FRAP

Inferences about biological mechanisms are enabled through FRAP analysis, ecological and evolutional processes can be elucidated using FRAP. For example, the co-occurrence of changes across contigs shows an ecological relation among such contigs. Contigs that increase by the same amount are part of an ecological unit that co-varies. Also, the co-occurrence of changes at specific sites shows an evolutionary relation between such sites, viral quasispecies(Lauring and Andino 2010) can be identified in this way.

Methods

File types

The input for FRAP are nucleotide fasta files. Datasets are usually fasta files containing multiple reads (DNA fragments between 35 and 1000 nucleotides originated from an NGS instrument), several datasets can be used at the same time as long as each sample is in

a single fasta file and have a unique identifier. Databases are usually fasta files containing several reference genomes or genes. Databases or Datasets can also be contigs.

FRAP outputs are tab delimited files, each column is a dataset and each row is an element of the database. Both files are provided as outputs in the basic version of FRAP, just map frap (jmp.pl). Such files are the hits file which contains the number of exact hits to each element of the database (hits.tab) and the normalized file (normalized.tab) which contains the FRAP values to each element of the database.

Two graphical outputs are generated using frap-tools. The first one is a generic heatmap of the FRAP values. The second are fragment recruitment plots in which the x axis is the element in the database and the y axis is the identity of each hit.

A fasta output with the reads that have exact hits to the database can be generated using the script si_hit.pl, a fasta file containing the reads that do not have exact hits to the database can be generated using the script no_hit.pl.

FRAP implementations

FRAP implementations using existing mapper algorithms are FRAP-smalt and FRAP-HISAT, both can be used through the perl script jmf.pl. The arguments of jmf.pl are the complete path to the database fasta file, the path to the folder containing the datasets, the path to the results folder that would be created, the mapper to use, and the average L(mean) to use.

FRAP-smalt is a FRAP implementation that uses the mapper smalt ("SMALT Manual" 2010) to find exact hits between dataset and database. To use smalt, the database is indexed using a k-mer of 10 (-k 10) with a step size of 5 (-s 5), next the dataset is mapped to

the database using an identity filter of 100% (y=1), the selected output format is samsoft (-o samsoft). From the samsoft file, the exact hits are counted, and tab delimited files generated.

FRAP-HISAT is a FRAP implementation that uses the mapper HISAT (Kim, Langmead, and Salzberg 2015) to find exact hits between dataset and database. HISAT uses scores to evaluate if a read map to the database. To obtain a score equivalent to 100% identity over 100% of the read, the scores were back calculated to identity the corresponding identity to score (Figure 2,9). In a read of length 100, a score of 0 represents 75% identity, a score of 120 represents 90% identity, a score of 160 represents 95% identity, a score of 200 represents 100% identity. A score of 200 is used in FRAP.

FRAP was implemented in Perl using a hash strategy, this implementation is called hash exact hit (heh). FRAP-heh is available in the repository https://github.com/yinacobian/frap-heh, to construct the index a k-mer size is selected, viral refseq index was contructed using 50, 100 and 200 k-mers, index construction is made using the program heh-db2.pl. The database in then indexed in a hash that is used to compare the dataset to, which is implemented in the program heh-compa.pl. Further development is needed to make FRAP-heh efficient.

FRAP was implemented using the command grep. FRAP-grep is available in the repository https://github.com/yinacobian/FRAP-grep, the database is converted to a single line, then each read is compared to the database. This approach in portable since grep is a broadly used command available in many operating systems. In an Ubuntu operating system with 156 Gb of RAM, each read took 0.5 seconds to be searched against viral refseq (7,194

genomes and 218Mb), which is very slow. FRAP-grep needs further development to be more efficient.

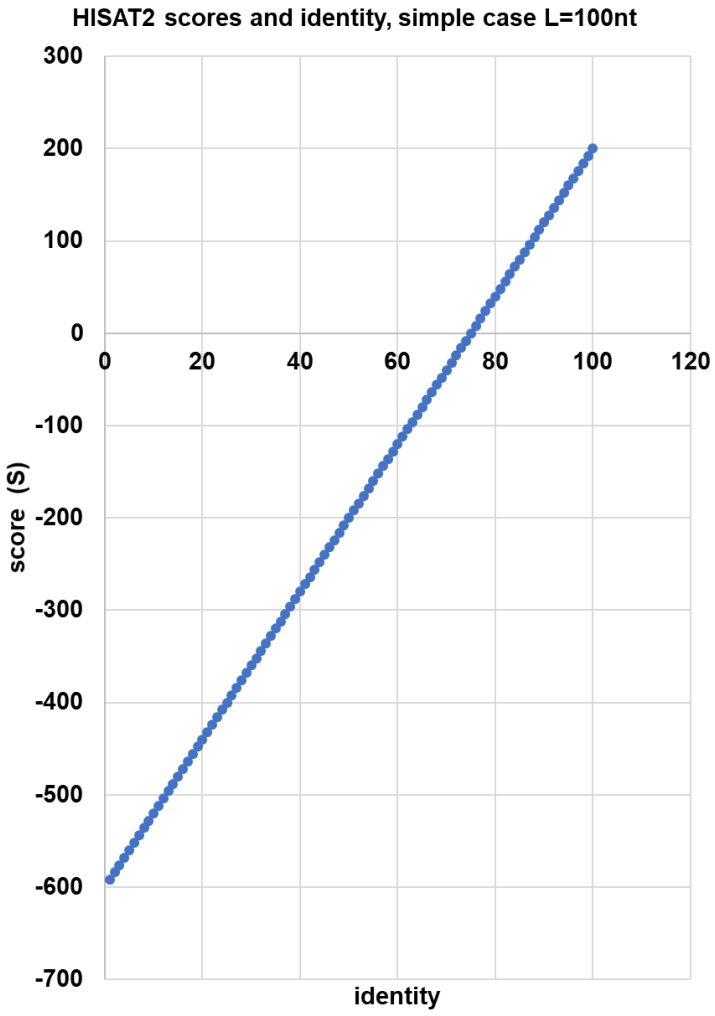


Figure 2.9 Relation between HISAT2 score and identity for reads of length 100nt. HISAT2 (Kim, Langmead, and Salzberg 2015) scores are calculated as following: score = ((number of mismatches * (match weight=2)) + (length – number of mismatches))*(mismatch weight=-6). In a read of length 100, a score of 0 represents 75% identity, a score of 120 represents 90% identity, a score of 160 represents 95% identity, a score of 200 represents 100% identity.

FRAP-tools

Graphic outputs from FRAP are obtained using frap-tools, which uses R and python. The script Plot_Heatmap.R creates a heatmap from FRAP values, this is a generic heatmap that can be further customized by the user. The script fragplot2.py creates individual fragment recruitment plots. The program si_hit.pl is used to generate a fasta file containing the reads that map the dataset, the program no_hit.pl is used to generate a fasta file containing the reads that do not map to the database.

Reference databases

Three main databases were used in the presented FRAP examples: bacteria representative genomes, archaea refseq, and viral refseq. Bacteria representative genomes contains 5,460 complete genomes which include at least one genome from each branch of the bacteria phylogenetic tree as calculated by NCBI(O'Leary et al. 2016). Archaea refseq contains 192 complete genomes. Viral refseq contains 8,321 complete genomes. All databases are part of RefSeq-release80 and were accessed on 01/09/2017.

Denovo assemblies

Denovo assemblies were performed using SPAdes (Bankevich et al. 2012) with the parameter –only-assembler. For some of the test cases, only contigs >900 nucleotides were used as database.

Acknowledgments

We are grateful to Sandi Calhoun and Mark Little for data sharing during FRAP tests and implementation. We are grateful to Dr. Rachel Dutton for providing cheese samples for

viromes generation and for viromes sequencing. Thanks to the bioinformatics breakfast group and the biomath group for fruitful discussion during the development of this project.

References

2019, CAMI. n.d. “CAMI II Challenge.” https://data.cami-challenge.org/cami2.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” Journal of Computational Biology 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Chen, Lihong, Dandan Zheng, Bo Liu, Jian Yang, and Qi Jin. 2016. “VFDB 2016: Hierarchical and Refined Dataset for Big Data Analysis - 10 Years On.” Nucleic Acids Research 44 (D1): D694–97. https://doi.org/10.1093/nar/gkv1239.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for ‘omics Data.” PeerJ 3: e1319. https://doi.org/10.7717/peerj.1319.

Fonseca, Nuno A., Johan Rung, Alvis Brazma, and John C. Marioni. 2012. “Tools for Mapping High-Throughput Sequencing Data.” Bioinformatics 28 (24): 3169–77. https://doi.org/10.1093/bioinformatics/bts605.

Greninger, Alexander L. 2018. “The Challenge of Diagnostic Metagenomics.” Expert Review of Molecular Diagnostics 18 (7): 605–15. https://doi.org/10.1080/14737159.2018.1487292.

Huson, Daniel H., Sina Beier, Isabell Flade, Anna G??rska, Mohamed El-Hadidi, Suparna Mitra, Hans Joachim Ruscheweyh, and Rewati Tappu. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.” PLoS Computational Biology 12 (6): 1–12. https://doi.org/10.1371/journal.pcbi.1004957.

Huttenhower, Curtis. 2019. “HUMANN2 Pipeline.” 2019. http://huttenhower.sph.harvard.edu/humann.

Jia, Baofeng, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, et al. 2017. “CARD 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database.” Nucleic Acids Research 45 (D1): D566–73. https://doi.org/10.1093/nar/gkw1004.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. “HISAT : A Fast Spliced Aligner with Low Memory Requirements” 12 (4). https://doi.org/10.1038/nmeth.3317.

Lauring, Adam S., and Raul Andino. 2010. “Quasispecies Theory and the Behavior of RNA Viruses.” PLoS Pathogens 6 (7): 1–8. https://doi.org/10.1371/journal.ppat.1001005.

McIntyre, Alexa B. R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, et al. 2017. “Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers.” Genome Biology 18 (1): 182. https://doi.org/10.1186/s13059-017-1299-7.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. “Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju.” Nature Communications 7: 1–9. https://doi.org/10.1038/ncomms11257.

Naccache, Samia N., Scot Federman, Narayanan Veeraraghavan, Matei Zaharia, Deanna Lee, Erik Samayoa, Jerome Bouquet, et al. 2014. “A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from next-Generation Sequencing of Clinical Samples.” Genome Research 24 (7): 1180–92. https://doi.org/10.1101/gr.171934.113.

O’Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” Nucleic Acids Research 44 (D1): D733–45. https://doi.org/10.1093/nar/gkv1189.

Overbeek, Ross, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, et al. 2014. “The SEED and the Rapid Annotation of Microbial Genomes Using Subsystems Technology (RAST).” Nucleic Acids Research 42 (D1): 206–14. https://doi.org/10.1093/nar/gkt1226.

Pandiselvam, P, T Marimuthu, and R Lawrance. 2014. “A Comparative Study on String Matching Algorithms of Biological Sequences.” International Conference on Intelligent Computing, no. January 2014: 1–5. https://arxiv.org/pdf/1401.7416.pdf%0Ahttp://arxiv.org/ftp/arxiv/papers/1401/1401.7416.pdf.

Roux, Simon, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, Didier Debroas, and François Enault. 2011. “Metavir: A Web Server Dedicated to Virome

Analysis." Bioinformatics 27 (21): 3074–75. https://doi.org/10.1093/bioinformatics/btr519.

Sayers, Samantha, Li Li, Edison Ong, Shunzhou Deng, Guanghua Fu, Yu Lin, Brian Yang, et al. 2019. "Victors: A Web-Based Knowledge Base of Virulence Factors in Human and Animal Pathogens." Nucleic Acids Research 47 (D1): D693–700. https://doi.org/10.1093/nar/gky999.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation - A Benchmark of Metagenomics Software." Nature Methods 14 (11): 1063–71. https://doi.org/10.1038/nmeth.4458.

Silva, Genivaldo, Bas Dutilh, and Robert Edwards. 2016. "FOCUS2: Agile and Sensitive Classification of Metagenomics Data Using a Reduced Database." BioRxiv, 046425. https://doi.org/10.1101/046425.

Singer, Esther, Bill Andreopoulos, Robert M. Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, et al. 2016. "Next Generation Sequencing Data of a Defined Microbial Mock Community." Scientific Data 3: 160081. https://doi.org/10.1038/sdata.2016.81.

"SMALT Manual." 2010. October, no. 1: 1–7.

So, Martin. 2017. "Sequence Analysis Edlib : A C / C 11 Library for Fast , Exact Sequence Alignment Using Edit Distance" 33 (January): 1394–95. https://doi.org/10.1093/bioinformatics/btw753.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. "Big Data: Astronomical or Genomical?" PLoS Biology 13 (7): 1–11. https://doi.org/10.1371/journal.pbio.1002195.

Appendix for Chapter 2

Supplemental tables

Supplemental Table 2.1. FRAP vs viral refseq of plasma metagenome spiked with HIV. Five subsample replicates. spikeHIV-mg-SRR1106548-2.

id	name	R1	R2	R3	R4	R5	Mean	SD
gi|9628705|ref|NC_001710.1|	GB virus C/Hepatitis G virus, complete genome	899.54	899.54	899.54	899.54	899.54	899.54	0.0000
gi|9629357|ref|NC_001802.1|	Human immunodeficiency virus 1, complete genome	324.90	324.90	324.90	324.90	324.90	324.90	0.0000
gi|56718463|ref|NC_003287.2|	Enterobacteria phage M13, complete genome	17.59	17.70	17.59	17.33	17.73	17.59	0.1420
gi|295413923|ref|NC_014075.1|	Torque teno virus 12, complete genome	13.68	13.68	13.68	13.68	13.68	13.68	0.0000
gi|9627425|ref|NC_001604.1|	Enterobacteria phage T7, complete genome	7.96	7.96	7.96	7.96	7.96	7.96	0.0000
gi|339832375|ref|NC_015783.1|	Torque teno virus, complete genome	6.36	6.36	6.36	6.36	6.36	6.36	0.0000
gi|29502191|ref|NC_002076.2|	Torque teno virus 1, complete genome	5.72	5.72	5.72	5.72	5.72	5.72	0.0000
gi|730977588|ref|NC_025824.1|	Enterobacteria phage fd strain 478, complete genome	5.13	5.01	5.13	5.39	4.99	5.13	0.1420
gi|295413965|ref|NC_014082.1|	Torque teno mini virus 7, complete genome	4.29	4.29	4.39	4.34	4.34	4.33	0.0382
gi|9634957|ref|NC_002195.1|	Torque teno mini virus 9, complete genome	3.93	3.93	3.83	3.88	3.88	3.89	0.0387
gi|9626243|ref|NC_001416.1|	Enterobacteria phage lambda, complete genome	3.13	3.09	3.13	3.13	3.13	3.12	0.0166
gi|295413834|ref|NC_014073.1|	Torque teno virus 28, complete genome	1.83	1.83	1.83	1.83	1.83	1.83	0.0000
gi|295413918|ref|NC_014074.1|	Torque teno virus 27, complete genome	1.74	1.74	1.74	1.78	1.74	1.75	0.0162
gi|295413958|ref|NC_014081.1|	Torque teno virus 3, complete genome	1.77	1.77	1.77	1.77	1.77	1.77	0.0000
gi|295413928|ref|NC_014076.1|	Torque teno virus 10, complete genome	1.56	1.56	1.56	1.56	1.56	1.56	0.0000
gi|295441877|ref|NC_014089.1|	Torque teno mini virus 5, complete genome	1.56	1.56	1.56	1.56	1.56	1.56	0.0000
gi|418487627|ref|NC_019445.1|	Escherichia phage TL-2011b, complete genome	1.27	1.27	1.27	1.27	1.27	1.27	0.0000
gi|295441896|ref|NC_014094.1|	Torque teno virus 6, complete genome	0.85	0.85	0.85	0.85	0.85	0.85	0.0000

Supplemental Table 2.1. FRAP vs viral refseq of plasma metagenome spiked with HIV. Five subsample replicates. spikeHIV-mg-SRR1106548-2. (Continued)

id	name	R1	R2	R3	R4	R5	Mean	SD
gi|727071839|ref|NC_025726.1|	Torque teno mini virus ALA22, complete genome	0.72	0.72	0.72	0.72	0.72	0.72	0.0000
gi|295441905|ref|NC_014096.1|	Torque teno virus 15, complete genome	0.48	0.48	0.48	0.48	0.48	0.48	0.0000
gi|209427726|ref|NC_011356.1|	Enterobacteria phage YYZ-2008, complete prophage genome	0.68	0.68	0.68	0.68	0.68	0.68	0.0000
gi|46401626|ref|NC_005856.1|	Enterobacteria phage P1, complete genome	0.59	0.63	0.64	0.62	0.60	0.62	0.0179
gi|557307526|ref|NC_022749.1|	Shigella phage SfIV, complete genome	0.70	0.66	0.66	0.68	0.71	0.68	0.0197
gi|744692686|ref|NC_026013.1|	Microviridae IME-16, complete sequence	0.71	0.71	0.71	0.71	0.71	0.71	0.0000
gi|971482474|ref|NC_028748.1|	Bacillus phage BMBtpLA, complete genome	0.46	0.42	0.40	0.42	0.44	0.43	0.0206
gi|374531645|ref|NC_016765.1|	Pseudomonas phage vB_PaeS_PMG1, complete genome	0.28	0.28	0.28	0.28	0.28	0.28	0.0025
gi|124300942|ref|NC_008376.2|	Geobacillus phage GBSV1, complete genome	0.26	0.26	0.26	0.26	0.26	0.26	0.0000
gi|221328618|ref|NC_011976.1|	Salmonella phage epsilon34, complete genome	0.32	0.32	0.32	0.32	0.32	0.32	0.0000
gi|849250250|ref|NC_027339.1|	Enterobacteria phage SfI, complete genome	0.18	0.16	0.15	0.16	0.15	0.16	0.0109
gi|82700933|ref|NC_007623.1|	Pseudomonas phage EL, complete genome	0.16	0.16	0.16	0.16	0.16	0.16	0.0000
gi|295441884|ref|NC_014091.1|	Torque teno virus 16, complete genome	0.16	0.16	0.16	0.16	0.16	0.16	0.0000
gi|209447126|ref|NC_011357.1|	Stx2-converting phage 1717, complete prophage genome	0.15	0.15	0.15	0.15	0.15	0.15	0.0000
gi|937456792|ref|NC_027991.1|	Staphylococcus phage SA1, complete genome	0.16	0.16	0.16	0.16	0.16	0.16	0.0000
gi|939482395|ref|NC_002484.2|	Bacteriophage D3, complete genome	0.13	0.13	0.14	0.14	0.14	0.14	0.0024
gi|428782011|ref|NC_019711.1|	Enterobacteria phage HK629, complete genome	0.14	0.13	0.15	0.13	0.15	0.14	0.0096
gi|428782787|ref|NC_019723.1|	Enterobacteria phage HK630, complete genome	0.12	0.19	0.16	0.17	0.16	0.16	0.0220
gi|29134936|ref|NC_004629.1|	Pseudomonas phage phiKZ, complete genome	0.11	0.11	0.11	0.11	0.11	0.11	0.0000

Supplemental Table 2.1. FRAP vs viral refseq of plasma metagenome spiked with HIV. Five subsample replicates. spikeHIV-mg-SRR1106548-2. (Continued)

id	name	R1	R2	R3	R4	R5	Mean	SD
gi|428783345|ref|NC_019769.1|	Enterobacteria phage HK542, complete genome	0.04	0.03	0.03	0.03	0.04	0.04	0.0045
gi|428783215|ref|NC_019767.1|	Enterobacteria phage HK544, complete genome	0.04	0.06	0.03	0.02	0.01	0.03	0.0169
gi|428782316|ref|NC_019716.1|	Enterobacteria phage mEp460, complete genome	0.07	0.07	0.07	0.07	0.07	0.07	0.0000
gi|966201269|ref|NC_028694.1|	Propionibacterium phage PA1-14, complete genome	0.03	0.04	0.05	0.05	0.07	0.05	0.0117
gi|408905847|ref|NC_018852.1|	Propionibacterium phage P100D, complete genome	0.06	0.05	0.06	0.06	0.04	0.05	0.0088
gi|543171632|ref|NC_022342.1|	Propionibacterium phage PHL111M01, complete genome	0.05	0.05	0.05	0.05	0.05	0.05	0.0021
gi|330858351|ref|NC_015453.1|	Propionibacterium phage PAS50 endogenous virus, complete genome	0.02	0.05	0.04	0.05	0.07	0.04	0.0185
gi|849254459|ref|NC_027373.1|	Propionibacterium phage PHL030N00, complete genome	0.06	0.04	0.06	0.04	0.02	0.04	0.0150
gi|408905837|ref|NC_018842.1|	Propionibacterium phage P1.1, complete genome	0.08	0.06	0.04	0.05	0.05	0.06	0.0128
gi|682123269|ref|NC_024787.1|	Listeria phage LMTA-148, complete genome	0.02	0.02	0.02	0.02	0.02	0.02	0.0000
gi|238801880|ref|NC_012753.1|	Streptococcus phage 5093, complete genome	0.02	0.02	0.02	0.02	0.02	0.02	0.0000
gi|89152530|ref|NC_007805.1|	Pseudomonas phage F10, complete genome	0.03	0.03	0.03	0.03	0.03	0.03	0.0000
gi|431809676|ref|NC_019914.1|	Staphylococcus phage StB27, complete genome	0.00	0.00	0.00	0.00	0.00	0.00	0.0000
gi|849251120|ref|NC_027346.1|	Propionibacterium phage PHL171M01, complete genome	0.00	0.00	0.01	0.01	0.01	0.00	0.0038
gi|543171262|ref|NC_022334.1|	Propionibacterium phage PHL112N00, complete genome	0.00	0.01	0.00	0.00	0.01	0.00	0.0025
gi|906475910|ref|NC_027624.1|	Propionibacterium phage SKKY, complete genome	0.01	0.00	0.01	0.00	0.00	0.00	0.0025
gi|422933554|ref|NC_019491.1|	Cyprinid herpesvirus 1 strain NG-J1, complete genome	0.00	0.00	0.00	0.00	0.00	0.00	0.0002
gi|363539767|ref|NC_016072.1|	Megavirus chiliensis, complete genome	0.00	0.00	0.00	0.00	0.00	0.00	0.0000
							Mean of all SD	**0.0109**

Chapter 3 : Cystic Fibrosis Rapid Response: Translating Multi-omics Data into Clinically Relevant Information

Abstract

Pulmonary exacerbations are the leading cause of death in cystic fibrosis (CF). To track microbial dynamics during acute exacerbations, a CF Rapid Response (CFRR) strategy was developed. The CFRR relies on viromics, metagenomics, metatranscriptomics and metabolomics data to rapidly monitor active members of the viral and microbial community during acute CF exacerbations. To highlight CFRR, a case study of a CF patient is presented, in which an abrupt decline in lung function characterized a fatal exacerbation. The microbial community in the patient's lungs was closely monitored through the multi-omics strategy, which led to the identification of pathogenic shigatoxigenic *Escherichia coli* (STEC) expressing shiga toxin. This case study illustrates the potential for CFRR to deconstruct complicated disease dynamics and provide clinicians with alternative treatments to improve the outcomes of pulmonary exacerbations and expand the lifespans of individuals with CF.

Introduction

Cystic fibrosis (CF) is a recessive genetic disease in which defects or deficits in the cystic fibrosis transmembrane conductance regulator (CFTR) protein result in disease phenotypes of the pancreas, sweat glands and reproductive, respiratory and digestive systems (Knowles and Drumm, 2012). In the lungs of individuals with CF, mucociliary clearance is impaired, which promotes chronic polymicrobial infections (Laguna et al., 2016). Antibiotic treatments and proper disease management have extended the average lifespan of CF patients;

nevertheless, these polymicrobial lung infections are still the primary cause of morbidity and mortality (Alexander et al., 2016). Common bacteria that colonize CF lungs over the long-term include *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Haemophilus influenzae*, *Burkholderia cepacia* complex, *Rothia mucilaginosa* and *Streptococcus* spp. (Surette 2014; LiPuma 2010; Lim et al. 2013; Whiteson et al. 2014), but every CF individual presents a unique microbial community that changes over time (Lim et al., 2014a; Whelan et al., 2017; Zhao et al., 2012). This highlights the need to characterize the microbial communities in each CF individual.

Microbial community dynamics in CF lungs follow the Climax Attack Model (CAM) (Conrad et al., 2013; Quinn et al., 2014), in which a climax community is acclimated to the host and dominates during stable periods, and a transient attack community is associated with exacerbations. Attack communities are virulent and colonize the CF lungs either from an external source or are already present in the CF lungs and become active during exacerbations. In the CAM, attack communities lead to Cystic Fibrosis Pulmonary Exacerbations (CFPE), declines in lung function, and eventually death. Preventing CFPE relies on quickly identifying attack viral and microbial communities and the genes they carry and express, such as those encoding specific toxins (Gallant et al., 2015), to efficiently tailor antimicrobial therapies.

Herein we propose Cystic Fibrosis Rapid Response (CFRR), a strategy for determining microbial dynamics during CFPE. This strategy is a personalized multi-omics approach that uses viromes (Willner et al., 2012), metagenomes, metatranscriptomes (Lim et al., 2013b), and metabolomes (Quinn et al., 2016a; Whiteson et al., 2014) from longitudinal

samples to monitor the whole microbial community, particularly its active members and their metabolic products. Using CFRR to obtain personalized taxonomic and functional profiles of the lung microbial communities would provide clinicians with comprehensive information about each patient's viral and microbial ecosystem. This information allows clinicians to generate testable hypotheses, test those hypotheses using standard clinical tests, and propose specific clinical interventions (e.g., precisely targeted antibiotic therapy) to improve CFPE outcomes.

The ability to generate multi-omic datasets and analyze large amounts of data in a clinically relevant time frame (i.e., ≤ 48 hours) makes the CFRR approach applicable in CF clinical practice, especially in clinics closely related to research institutions. It requires access to a sequencing instrument, mass spectrometer, computational resources and specialized personnel in each one of these areas. In an optimal situation, the time between sample collection and data interpretation is 30 hours for metabolomes (Quinn et al. 2016), 38 hours for metagenomes and metatranscriptomes, and 48 hours for viromes. These times are expected to shorten as technologies improve. The rapid decrease in sequencing costs (National Human Genome Research Institute, 2018) and incorporation of sequencing cores within hospitals (Deurenberg et al., 2017) will increase CF patients accessibility to the CFRR in the foreseeable future.

A case study is presented to demonstrate the potential of the CFRR strategy. A 37-year-old male CF patient (CF01) was monitored over a two-year period with metagenomes, metatranscriptomes and metabolomes. Integrating the information from these sources led to

the identification of an attack community in which a strain of *E. coli* that likely produced shiga-toxin was detected during a fatal exacerbation.

Results

Patient CF01 fatal exacerbation expedited monitoring: metatranscriptomes and metabolomes

An overall decline in lung function was observed in patient CF01 during his last year of life and four CFPEs were reported. In the last month of life 10% of predicted FEV_1 was lost (Figure 3.1-A). During the last exacerbation, the patient was hospitalized at the intensive care unit (ICU) for seven days and then died. The fatal exacerbation was characterized by severe lung tissue damage (Figure 3.1- D, Supplemental Table 3.1A), an increase in white blood cell counts (Figure 3.1-B, Supplemental Table 3.1B) and a general decline in health. During the fatal exacerbation, clinical microbiology lab cultures from sputum samples tested positive for *P. aeruginosa*, *Stenotrophomonas maltophilia*, *Aspergillus terreus* and yeast (Figure 3.1-C, Supplemental Table 3.1C). Treatment alternated between the antibiotics aztreonam, azithromycin, in addition to a sulfonamide and a quinolone; at the ICU, colistin and meropenem were administered (Supplemental Table 3.1D) but no improvement was observed.

The CFRR strategy was launched to rapidly identify the cause of the CFPE. Sputum samples were collected 7 and 8 days before death (samples D-7 and D-8). In samples D-7 and D-8 active members of the microbial community were determined using metatranscriptomics. In sample D-8 small molecule profiles (using metabolomics) were characterized and a total DNA metagenome was sequenced.

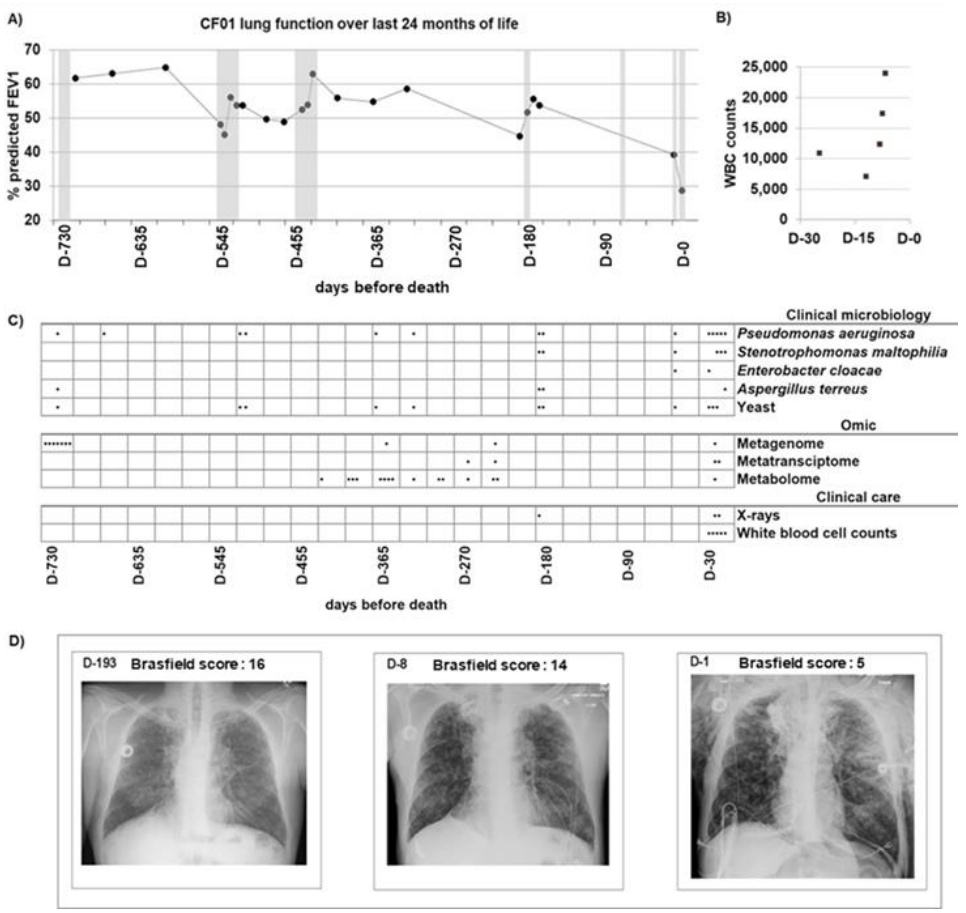


Figure 3.1 Clinical data for the last 24 months of patient CF01's life. (A) Percentage of predicted FEV1 of patient CF01 over the last 24 months of life. Solid dots are FEV1 measurements. The line is included to highlight lung function dynamics and does not represent measurements. Seven exacerbation periods were reported and are shown in gray. (B) White blood cell (WBC) counts for the last month of life. (C) Clinical microbiology positive cultures from patient CF01's sputum samples over the last 24 months of life. Dots represent days where cultures were positive for each microbe tested in the clinical microbiology panel. Omics sampling points for metagenomes, metatranscriptomes, and metabolomes are indicated by dots in the Omic panel. Performed X rays and WBC measurement days are indicated with dots in the clinical care panel. (D) Patient CF01 chest X rays in a frontal view with quantitative disease severity evaluation using Brasfield scores. D-193, mild exacerbation; D-8, acute exacerbation, the time point where CFRR data were obtained; D-1, 1 day before death. A lower Brasfield score represents a higher disease severity. The Brasfield score scale is from 25 to 0, where 25 is lower disease severity and 0 is higher disease severity. Parameters used for Brasfield scores calculations are air trapping, linear markings, nodular cystic lesion, large lesions, and general severity, and individual scores are shown in Supplemental Table 3.1A in the supplemental material.

Metatranscriptomics data from sample D-8 showed that the most abundant microbial ribosomal RNAs (rRNA) belonged to the genera *Bacillus* (29.9%), *Escherichia-Shigella* (23.9%), *Streptococcus* (11.6%), *Salmonella* (6.9%) and *Lactococcus* (4.4%) among other genera (23.3%) (Supplemental Figure 3.1-A). The microbial messenger RNA (mRNA) composition was dominated by the genera *Pseudomonas (97.1%)*, followed by *Stenotrophomonas* (1.9%) and *Escherichia* (0.07%) (Supplemental Figure 3.1-B). At species level resolution, the most abundant bacterial genomes (based on total RNA, Figure 3.2) were: *Bacillus* sp., *E. coli* (STEC), *Salmonella enterica* serovar Infantis, *P. aeruginosa* and *S. maltophilia*. Enterobacteria phage SP6, *Pseudomonas* phages and *Stenotrophomonas* phage S1 were also detected. Two members of the phylum Ascomycota were identified: *Candida albicans* and *Aspergillus fumigatus*. Metagenomics data of sample D-8 detected *Pseudomonas* (98.5%) as the dominant bacterial genus (Supplemental Figure 3.5-A).

	Pseudomonas	Escherichia	Bacillus	Salmonella	Stenotrophomonas
rRNA	1.03	23.99	29.94	6.88	0.02
mRNA	97.15	0.07	0.00	0.00	1.92
all RNA	21.30	18.95	23.62	5.43	0.42
reference genome	*Pseudomonas aeruginosa* LESlike4	*Escherichia coli* O157:H7	*Bacillus sp.* WP8	*Salmonella enterica* serovar Infantis	*Stenotrophomonas maltophilia* K279a
GenBank ID	CP006985.1	AE005174.2	CP010075.1	LN649235.1	AM743169.1
reference bp	6,524,053	5,528,445	3,708,888	4,710,675	4,851,126

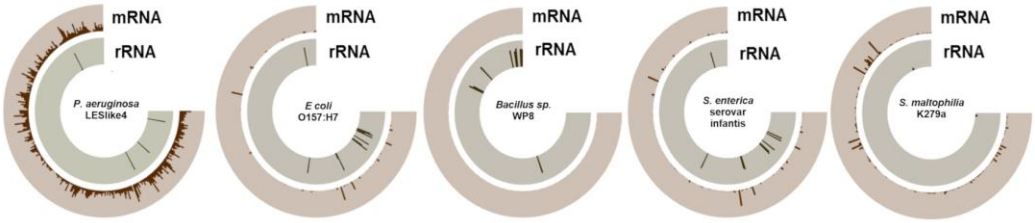


Figure 3.2 The most abundant bacterial genera of fatal exacerbation sample D-8. The relative abundances of each genus as determined by rRNA, mRNA, and total RNA are shown. A reference genome from each genus was selected based on the number of reads recruited in the rRNA (Escherichia, Bacillus, and Salmonella) or mRNA (Pseudomonas and Stenotrophomonas) category. Fragment recruitment was visualized using Anvi'o, showing a logarithmic scale for mRNA and rRNA from 1 to 1,000. Anvi'o plots show reads mapped along the genome coordinates. Nonribosomal microbial reads were recruited against each reference genome using SMALT with an identity cutoff of 80% and are shown in brown along the external ring. rRNA reads were classified into each genus by BLASTn, were recruited against the corresponding reference genome using SMALT with an identity cutoff of 60% and are shown in gray along the internal ring.

The presence of *Escherichia-Shigella* in the lungs of a CF patient is unusual and thus a detailed analysis was performed to further resolve the taxonomy at strain level. Strain level analysis identified that *E. coli* present in CF01's lungs was most closely related to the genome of *E. coli* (STEC) B2F1. This strain typically carries the shiga toxin 1 and shiga toxin 2 genes, both of which were identified in the metatranscriptomes (Figure 3.3-B, C). Furthermore, the shiga toxin receptor, globotriaosylceramide (Gb3), was detected in the metabolome from sample D-8 (Figure 3.3-A). This suggests that shiga toxin, and its Gb3 target, were being produced in the lungs of CF01. Gb3 is produced in human cells by Gb3 synthase, which adds a sugar to a lactosylceramide molecule. Ceramide is produced by sphingomyelinase (SMase)

in the host cell or by the action of bacterial encoded SMase (Figure 3.5-B). The gene that encodes a *P. aeruginosa* secreted SMase, the hemolytic phospholipase C (PlcH) (Vasil et al., 2009), was detected in the sample D-8 metatranscriptome (Supplemental Figure 3.2-B).

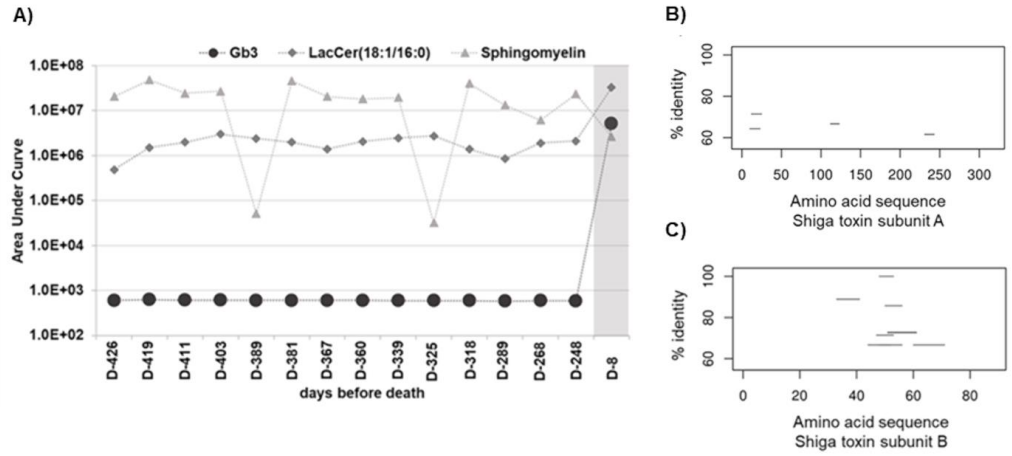


Figure 3.3 Shiga toxin and its human receptor globotriaosylceramide (Gb3). (A) The masses of globotriaosylceramide and its precursors lactosylceramide and sphingomyelin from exacerbation sample D-8 and 14 historical nonexacerbation samples were determined by parent mass searching and validated by MS/MS matching. The fatal exacerbation sample is shown in gray. (B) STEC BRF1 was used as a reference genome for fragment recruitment to the Shiga-like toxin 2 subunit A protein sequence. The amino acid sequence position is shown on the x axis, and percent identity is shown on the y axis. The nucleotide sequences from patient CF01 metatranscriptome exacerbation sample D-8 were mapped to proteins using BLASTx with an E value cutoff of 0.001 and filtered by an identity of ¬60%. (C) Metatranscriptome recruitment as explained above for panel B, except that in this case, reads were recruited to the Shiga-like toxin 2 subunit B amino acid sequence.

In a longitudinal metabolomics dataset, Gb3 was highly abundant ($p < 0.0001$) in sample D-8 but was in low abundance in the prior samples (Figure 3.3-A). The Gb3 precursor lactosylceramide (18:1/16:0) (Sandvig et al., 2012) and its ceramide donor sphingomyelin (18:1/16:0) (Obrig et al., 2003) were abundant in all samples throughout the longitudinal dataset (Figure 3.3-A, Supplemental Table 3.2A). These data demonstrate that Gb3 precursors

were present for at least a year before the fatal exacerbation, but Gb3 was produced in significantly high quantities eight days before death (sample D-8).

Gb3 levels positively correlate with shiga toxin levels (Boyd et al., 1993) although the mechanism behind this positive correlation is not clear. Gb3 is the only known functional receptor for shiga toxins (Aigal et al., 2015) and shiga toxins induce reorganization of lipids in the epithelial cell's membrane. Shiga toxin B can bind up to 15 Gb3 molecules (Ling et al., 1998) and this binding result in the aggregation of Gb3 in lipid rafts. The aggregation of Gb3 in lipid rafts promotes a negative membrane curvature and internalization of shiga toxin (Betz et al., 2011). Spatial distribution of Gb3 in the cell membrane has a regulatory role in its presentation (Lingwood et al., 2010), thus higher recruitment of Gb3 in lipid rafts may induce the production of more Gb3.

Antibiotic resistance genes were detected in the metatranscriptomes of D-8 and D-7 samples. Transcripts encoding all the protein components were identified for two RND-type multidrug exporters, MexGH1-OpmD (Aendekerk et al., 2002) and MexA-MexB-OprM (Li et al., 1995) previously described in *Pseudomonas*, as well as the tetracycline efflux pump tet(C) previously described in *Achromobacter*. Transcripts encoding several beta-lactamases were identified, such as TEM-116, PDC-3, OXA-50 and BEL-3 (Jia et al., 2017), which are typically found in *Pseudomonas*, and CTX-M-21 (Saladin et al., 2002), which is usually found in Enterobacteriaceae. Transcripts encoding for enzymes that are involved in resistance to macrolides, aminoglycosides, lincosamide, diaminopyrimidine and glycopeptide antibiotics were detected; these enzymes were previously described in *Pseudomonas, Achromobacter*,

Escherichia, Streptomyces, Paenibacillus, Clostridium and *Morganella* (Supplemental Table 3A).

A partial *P. aeruginosa* genome sequence was recovered by assembling reads from the fatal exacerbation metatranscriptomes (D-8 and D-7) into contigs and then mapping those contigs to the *P. aeruginosa* PAO1 reference genome (Supplemental Figure 3.2-A). In the resulting *P. aeruginosa* CF01 contigs, 38 genes related to resistance to antibiotics and toxic compounds were identified (Supplemental Table 3.3B). Two prophages were also identified in the assembled *P. aeruginosa* CF01 contigs (D-8 and D-7); one was complete and the second one was a partial prophage (Supplemental Figure 3.2-C and 3.2-D).

Bacterial small molecule profiles before and during fatal exacerbation.

Longitudinal metabolomic data from CF01 historical samples and fatal exacerbation sample D-8 were compared to metabolic profiles from six pathogenic bacterial isolates previously detected in CF sputum (*P. aeruginosa* VVP172, *Enterococcus* sp. VVP100, *E. coli* VVP427, *Streptococcus* sp. VVP047, *Stenotrophomonas* sp. VVP327 and *S. aureus* VVP270). The goal was to identify metabolites produced by pathogenic bacteria and track how changes in their abundances might have preceded the fatal exacerbation. Metabolites from these pathogens were consistently detected throughout the longitudinal samples. In sample D-8 there was an increase ($p < 0.001$) in the number of metabolites that matched *P. aeruginosa* VVP172, *E. coli* VVP427, *Streptococcus* sp. VVP047 and *S. aureus* VVP270 (Supplemental Figure 3.3, Supplemental Table 3.2B).

Active members of the microbial community during a stable period and the fatal exacerbation.

Analysis of metatranscriptomes from a stable period 10 and 9 months before the fatal exacerbation event (samples D-303 and D-279) identified several differences between this stable period and the fatal exacerbation. First, Firmicutes was the most active phylum during the stable period whereas Proteobacteria was the most active during exacerbation (Figure 3.4-A). Second, samples from the stable period showed an active microbial community that was more even and diverse than the community in exacerbation samples (Figure 3.4-D). Third, transcripts from *Pseudomonas* were detected at very low levels in stable samples (average relative abundance 3%), but at high levels (average relative abundance 37%) in exacerbation samples (Supplemental Figure 3.2-A). Fourth, the percentages of unclassified sequences were higher in the stable samples D-303 and D-279 (40.9% and 39.0%) than in the exacerbation samples D-8 and D-7 (27.6% and 17.0%). Fifth, a higher fractional abundance of bacteriophages was detected in the fatal exacerbation samples relative to the stable ones. Enterobacteria phage SP6, several *Pseudomonas* phages (Figure 3.4-B) and sarcoma viruses (Figure 3.4-C) were the dominant viruses in samples D-8 and D-7.

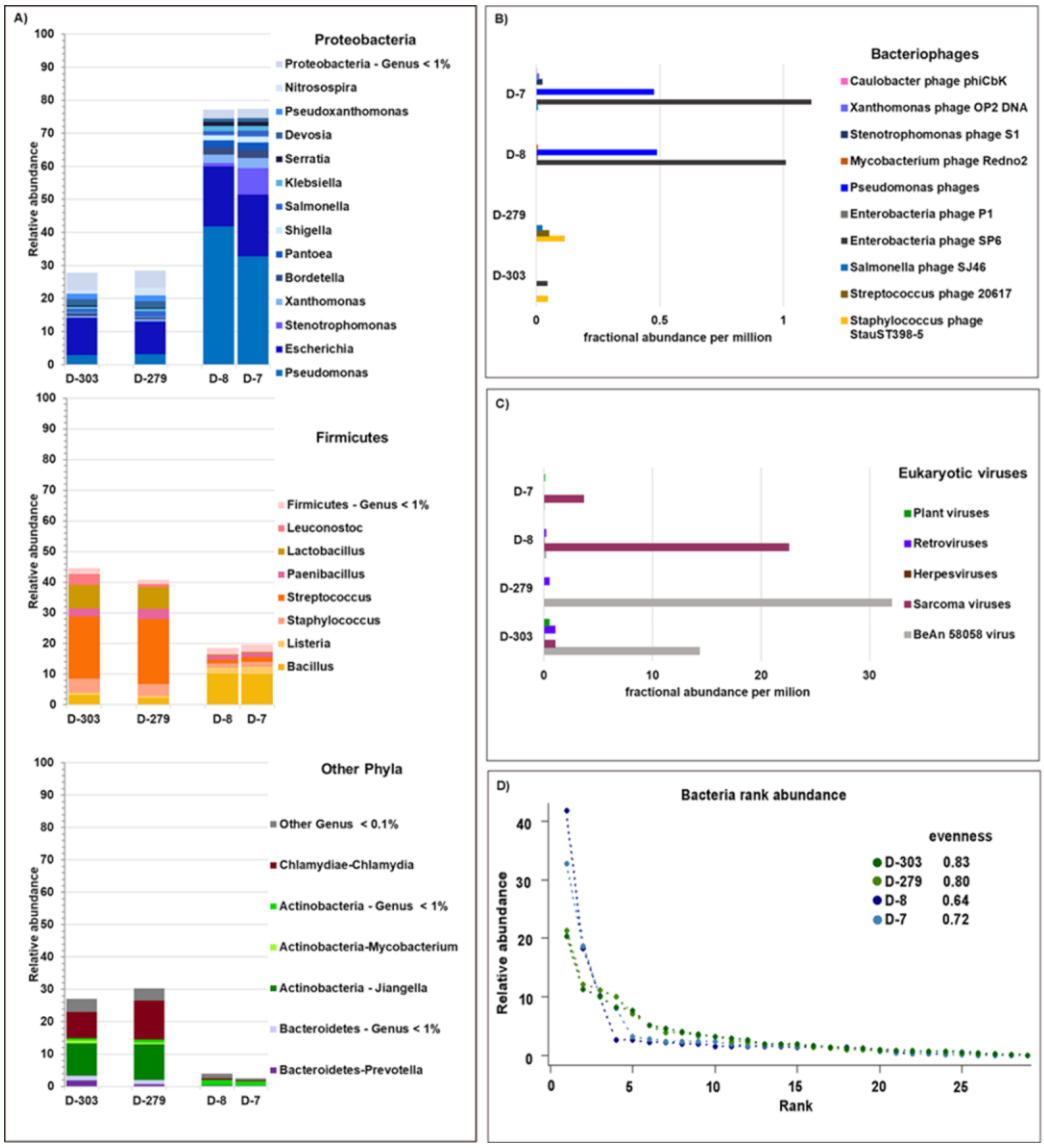


Figure 3.4 Actively transcribing members of the viral and bacterial communities in sputum samples of patient CF01. Metatranscriptomes from two exacerbations and two stable samples were obtained. (A) Bacterial taxonomical assignments were made using KAIJU at the genus level and are color-coded by phylum. (B) Fractional abundances of bacteriophages based on viral RefSeq mapping and FRAP normalization. (C) Fractional abundances of eukaryotic viruses based on viral RefSeq mapping and FRAP normalization. (D) Bacterial rank abundance plot generated using relative abundances at the genus level. Evenness was

calculated as H/In(S), where H is the Shannon diversity index and S is the total number of species.

Microbial community dynamics during a non-fatal exacerbation.

Two years before the fatal exacerbation, CF01's lung function declined faster than in previous years (Supplemental Figure 3.4-A). The rate of lung function change in the last two years of life was -9.75 FEV1%/year (Supplemental Figure 3.4-C). The overall rate of lung function change during CF01 last 14 years of life was -1.39 FEV1%/year. During a two-year period of 4 and 3 years before death, the rate of lung function change was 1.30 FEV1%/year (Supplemental Figure 3.4-B).

During the two-year period leading up to the fatal exacerbation, seven exacerbation events were reported, and sputum samples were periodically screened for fungi and bacteria at the clinical microbiology lab (Supplemental Table 3.1-C). *P. aeruginosa* was detected in all samples. Six months before the fatal exacerbation *S. maltophilia* was detected, and during the last two months of life, *Enterobacter cloacae* was detected. *A. terreus* was detected in two samples in the last six months of life. Yeast was detected in all screened samples, except for the final exacerbation samples. Based on this information, several antibiotics were prescribed to manage the exacerbations (Supplemental Table 3.1-D); these included monobactams, macrolides, quinolones, beta-lactams, sulfonamides and a cationic polypeptide.

Two years before CF01's death, metagenomics was used to monitor the microbial composition of the respiratory tract during an exacerbation event, the subsequent antibiotic treatment (samples D-724 to D-718), and a stable period that followed (samples D-409 and D-286) (Supplemental Figure 3.5-A). The bacterial genera that best differentiated between samples collected during antibiotic treatment (D-722 to D-718) and no antibiotic treatment

(D-724 and D-723) were *Rothia, Campylobacter, Veilonella* and *Prevotella* (Supplemental Figure 3.6). The antibiotics prescribed during this exacerbation were a fluoroquinolone (ciprofloxacin), and a tetracycline (doxycycline). Clinical microbiology lab tests performed on D-719 were positive for *P. aeruginosa, Pseudomonas fluorescens, A. fumigatus* and yeast (Supplemental Table 3.1C). Exacerbation and stable samples had *Streptococcus* phages, *Staphylococcus* phages and *Pseudomonas* phages, whereas only exacerbation samples had a shigatoxin-converting phage (Supplemental Figure 3.5-B), and stable samples had higher abundances of Herpesviruses (Supplemental Figure 3.5-D).

Discussion

CF01 fatal exacerbation mechanism

The unusually fast decline of patient CF01 led to the implementation of the CFRR. During CF01 fatal exacerbation, *E. coli* mRNA, rRNA and metabolites were detected, which demonstrated not only the presence but also the activity of shigatoxigenic *E. coli*. The identification of a shigatoxinogenic *E. coli* is supported by ribosomal RNA (36,590 unique ribosomal RNA sequences in metatranscriptome D-8), messenger RNA (1,412 E. coli mRNA reads in metatranscriptome D-8, and 11 partial mRNA reads at 60% identity to STEC BRF1), and metabolites (10 matched metabolome spectra to E. coli in D-8). The presence of STEC in the lungs of a CF patient was alarming as this strain causes severe damage to the lung epithelium (Bergan et al., 2012; Uchida et al., 1999). Moreover, interactions between shiga toxin and the host epithelium were inferred from metabolomes. The molecule

globotriaosylceramide (Gb3), the receptor for shiga toxin, showed an increase of three orders of magnitude during the fatal exacerbation (sample D-8), compared to previous samples.

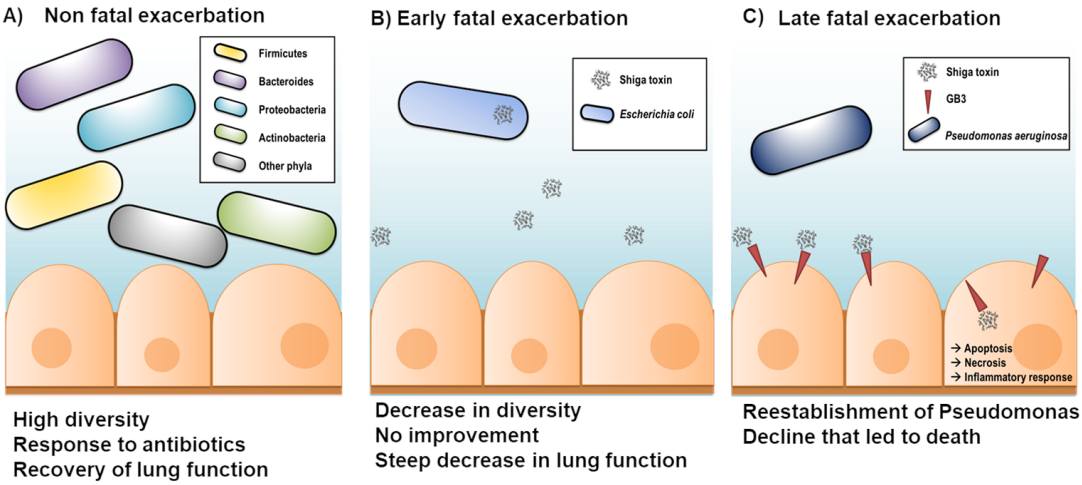


Figure 3.5 Proposed model of lung dynamics resulting in patient CF01's death. (A) A nonfatal exacerbation (days ⩾724 to ⩾718) was followed by a recovery of lung function, and attack and climax communities were diverse. (B) The fatal exacerbation was triggered by colonization by STEC, which is supported by the presence of its rRNA in the metatranscriptomes. This bacterium encodes Shiga toxin, which was likely taken into host cells by the human receptor globotriaosylceramide. (C) Later during the fatal exacerbation, Shiga toxin was internalized and then induced apoptosis, necrosis, and inflammation. P. aeruginosa was reestablished and came to dominate the community, as suggested by its abundant mRNA.

Altogether, these multi-omics data support the following model of microbial dynamics that caused patient CF01's death. At the beginning of the fatal exacerbation, STEC produced shiga toxin that remained inside the bacterial cells. Later in the exacerbation, STEC's cell membranes were disrupted and the shiga toxin was released (Figure 3.5-B). This release may have been triggered by the action of the cationic polypeptide colistin (Gupta et al., 2009). Next, the toxin was taken up by lung epithelial cells through the host membrane receptor globotriaosylceramide (Gallegos et al., 2012) (Figure 3.5-C). Inside the lung epithelial cells (Uchida et al., 1999), shiga toxin inhibited host translation by blocking the ribosomes, thereby

inducing cell death, necrosis and an acute inflammatory response (Melton-Celsa, 2014; Obrig, 2010; Uchida et al., 1999). The immune response and lung tissue damage was evident in the chest X-rays and the increase in white blood cells (Figure 3.1-D, D-8 and D-1).

During the fatal exacerbation, STEC led the attack community that ultimately destabilized the climax community, a phenomenon previously reported in CF exacerbations (Conrad et al., 2013); this resulted in a decline of evenness (diversity index that quantifies how equal the community is (Pielou, 1979)) and diversity (the number of different species in a community (Shannon and Weaver, 1949)), a switch from a community dominated by Firmicutes to one dominated by Proteobacteria, and transcription of Enterobacteria and *Pseudomonas* bacteriophages and sarcoma viruses. This event was followed by a *Pseudomonas* and *Stenotrophomonas* bloom, characterized by active transcription, as both rRNA and mRNA were detected, as well as an increase in their metabolites. *Pseudomonas* was the most active member of the microbial community with an mRNA abundance of 97%, followed by 1.92% of *Stenotrophomonas* mRNA. *Bacillus* was either lysed or dormant as only rRNA was detected. A feature that may have contributed to the success of *Pseudomonas* was its resistance to multiple antibiotics, as detected by the transcription of over 38 antibiotic resistance genes. This scenario is congruent with the one described by the clinical laboratory, as positive cultures for *Pseudomonas* and *Stenotrophomonas* were reported during the fatal exacerbation.

Additional dynamics such as bacteriophage induction may have happened during the fatal exacerbation, as active transcription was detected from Enterobacteria phage SP6 and

Pseudomonas bacteriophages. Bacteriophage induction is known to play a role in the control of bacterial populations in CF lungs (James et al., 2015).

CFRR for polymicrobial infections management, the importance of historical samples and a fast sample to results strategy.

The CFRR emerged from the need to investigate the cause of acute exacerbations. The power of the CFRR is shown in the information obtained for CF01 case study. The CFRR is ideal for medical centers closely associated with research facilities where the equipment is available. However, as technologies improve and become more accessible, CFRR could be implemented within the clinic.

A key component of the CFRR strategy is the comparison between acute exacerbation and stable periods. Because CF microbial communities are heterogeneous, a baseline needs to be determined for each patient. Longitudinal samples are essential to identify the changes in the microbial community and metabolites during acute exacerbations.

In the presented CF01 case study, historical samples were essential to differentiate the attack community that led to a fatal exacerbation from the attack community associated with a non-fatal exacerbation. The increase in Gb3 abundance during CF01 fatal exacerbation (Figure 3.3) was detected when comparing its abundances in historical samples. In the case of metabolites, a baseline is necessary for each CF patient because for many compounds the basal levels are not known. Accumulation of ceramides and sphingomyelins is observed in CF lungs (Seitz et al., 2015). In particular, sphingomyelins, ceramides and lactosylceramide are significantly higher in CF lungs compared to non-CF ones (Quinn et al., 2015).

A challenging component of the CFRR is the collection and storage of historical samples. Sputum samples intended for viromes, metagenomes and metabolomes (Wandro et al., 2017) analysis are stable if stored at -20 ºC or -80 ºC. Metatranscriptomes are prone to RNA degradation and sputum collection intended for this purpose requires RNA stabilization prior to -20 ºC or -80 ºC storage. Given these considerations, each patient can be provided with a non-thaw cycle -20 ºC freezer where individual raw sputum samples can be stored for viromes, metagenomes and metabolomes (Supplemental Figure 3.7-A). Sputum samples for metatranscriptomes can be collected during the patient's visit to the CF clinic, where immediately after collection the RNA integrity is preserved by adding TRizol or RNAlater. RNA should then be extracted as soon as possible. A proposed sampling scheme, in which higher resolution of samples is desired close to an acute exacerbation and fewer samples far away from the exacerbation event is proposed (Supplemental Figure 3.7).

Historical samples collected by the patient at home or during routine visits to the clinic are a valuable resource in the event of an acute exacerbation. In these cases, historical samples would be processed along with those from acute exacerbations in the CFRR pipeline (Figure 3.6) and valuable information would be obtained in less than 48 hours. This information is then analyzed by a multidisciplinary scientific team along with the clinician to 1) validate the multi-omics findings with approved clinical tests and 2) identify appropriate therapeutic options.

The information presented by the CFRR to the clinician is more detailed than that provided by classical clinical microbiology. A clear understanding of how this information is obtained and the exploratory nature of the findings needs to be considered when interpreting

the results. Discussion among clinicians and experts on the benefits and limitations of each ‘omics approach is essential to identify the elements causing CF acute exacerbations and then select the course of action to prevent a fatality. The final treatment decision is always in the hands of the clinician, who evaluates the different lines of evidence for each finding and considers the cost to benefit ratio of possible therapeutic interventions. The application of the CFRR in a clinical context gives CF patients the opportunity for a better outcome based on an informed treatment decision. Another consideration when implementing the CFRR in the clinic is the availability of financial resources to perform the multi-omics strategy in exacerbation and historical samples.

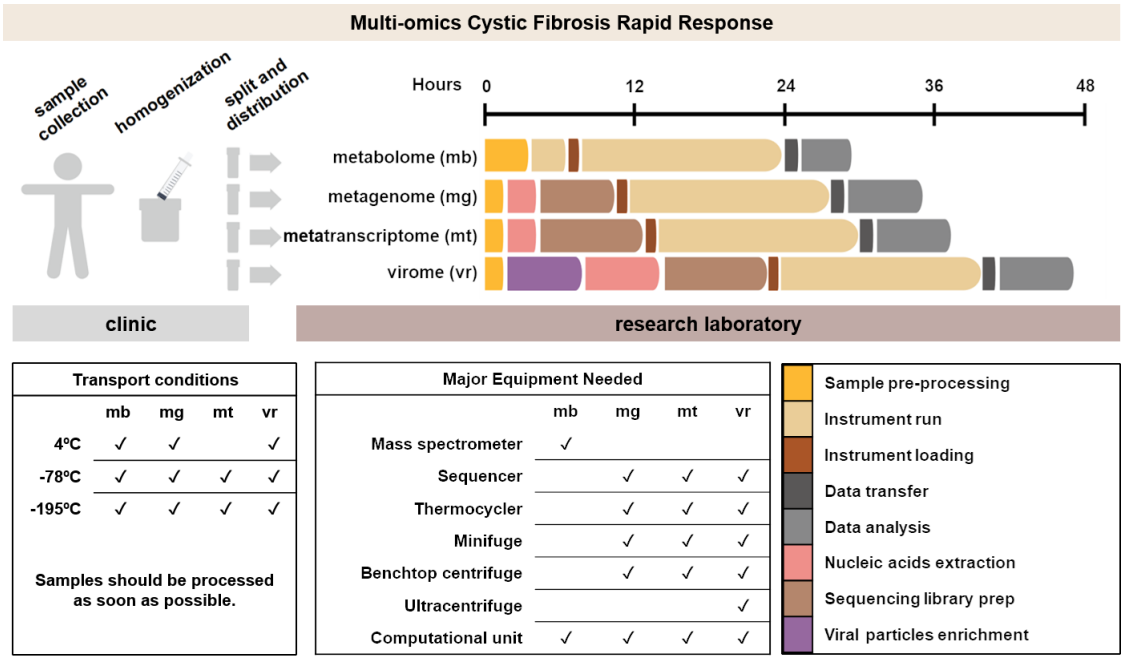


Figure 3.6 Cystic fibrosis rapid response. Our proposed multi-omics strategy is to analyze sputum samples from cystic fibrosis patients, in which metabolomes, metagenomes, metatranscriptomes, and viromes are obtained from a single sputum sample. Estimated times and equipment for each omics step are included, as are recommended transport conditions. Recommended transport condition temperatures can be achieved by using ice, dry ice, or liquid nitrogen.

Considerations about implementing the Cystic Fibrosis Rapid Response.

This was a retrospective study in which the patient's treatment was not modified based on the presented meta-omics results. The course of action of the CFRR strategy is to provide information to clinicians so that they can evaluate and confirm the findings before proceeding with pertinent treatment modifications.

In the case of CF01's fatal exacerbation, the information obtained from the CFRR strategy could have informed the course of action in the treatment with the following modifications: 1) use of different antibiotics, since colistin mechanism of action results in liberation of the bacteria cell contents such as shiga toxin, and 2) administration of neutralizing antibodies against shiga toxin. Colistin is a cationic polypeptide that disrupts the cell membrane of gram-negative bacteria through a detergent-like mechanism and it is often used in the treatment of multidrug-resistant exacerbation in patients with CF (Wishart et al., 2018).

In the presented case study only metatranscriptomes, metabolomes and metagenomes were used to elucidate the cause of a fatal exacerbation. In future CFRR case studies the use of viromes could be incorporated. The combination of metagenomes and viromes allows the identification of viral induction events, for example of prophages carrying toxins. Shigatoxigenic phages are capable of lysogenic conversion (Krüger and Lucchesi, 2015; Moons et al., 2013), and in the case of the CF01 fatal exacerbation, an early detection of shigatoxin in the viromes of historical samples could have provided valuable information about the coding potential of the viral community.

Time is crucial during the management of CF exacerbations. The estimated execution time of the CFRR in an ideal situation with specialized staff working 24/7 is 48 hours. Each step has room for improvement that would shorten the execution times. For example, real time direct sequencing, such as Oxford Nanopore, can eventually be used for the CFRR metagenomes, metatranscriptomes and viromes. These technologies provide genomic information as it is being sequenced (Greninger et al., 2015; Schmidt et al., 2017), which will be ideal for CFRR once sample preparation and data analysis are optimized for human DNA removal (Gu et al., 2016), and once high amounts of sputum starting material (400 ng of DNA needed for a Nanopore run) are no longer necessary for DNA for sequencing.

Combining data from multiple 'omics sources enabled the identification of shigatoxigenic *E. coli* as the likely cause of patient CF01's fatal exacerbation. Although these 'omics data were not used to alter clinical treatment of CF01, future applications of CFRR are expected to provide information that is essential for improving therapy, e.g., antibiotic resistance predictions and gene expression in major attack community pathogens. Although each individual's CF community is unique, these methods will allow for the observation of overarching trends within and between patients, for example a loss in diversity in acute exacerbations.

Materials and methods

Clinical data

Sample collection procedures and access to clinical data were approved by the Institutional Review Board of University of California San Diego (HRPP 081510) and San Diego State University (IRB#1711018R). Clinical microbiology, hematology, and X-rays

were taken during the normal care of the patient at UCSD medical center. Spirometry tests were used to calculate the percentage of predicted FEV_1 as previously described (Hankinson et al., 1999). Clinical status (exacerbation or stable) was determined by the clinician. Lung function dynamics were modeled using splines and linear model fitting as previously described (Conrad et al., 2017).

Metagenome and Metatranscriptome shotgun sequencing

Sputum samples were collected by expectoration in a sterile cup and processed for metagenomes or metatranscriptomes as previously described (Lim et al., 2014b). Metagenome libraries were constructed using Nextera DNA library preparation kit. Metatranscriptome libraries were constructed using TruSeq RNA library preparation kit. All libraries were sequenced on the Illumina GAIIx platform. Metatranscriptomes D-7 and D-8 were prepared using a modified procedure to obtain rRNA and mRNA in a single sequencing step, where half of the sample was depleted of rRNA using Ribo-Zero Gold kit (Lim et al., 2013b) while total RNA was extracted from the other half. Both fractions were pooled in a proportion of 4:1, and then a single Illumina library was constructed and sequenced.

Sequencing data processing

Quality filtering and dereplication was done using PRINSEQ (Schmieder and Edwards, 2011) (-min_qual_mean 20 -derep 1245 -lc_method entropy -lc_threshold 50 -ns_max_p 1 -out_bad null). Cloning vector sequences were removed using SMALT (-y 0.8 -x) with 80% identity against the UniVec database (National Center for Biotecnology Information, 2016a), possible sources of cloning vector sequences are reagents used in the library preparation (National Center for Biotecnology Information, 2016b; Woyke et al.,

2011). Human genome sequences were removed using BLASTn (E value of 0.1) against the human reference genome GRCh38. The resulting FASTA files are available on NCBI SRA with the following accession numbers: (SAMN10605049 to SAMN10605062, n=12). Metagenome and metatranscriptome datasets presented in this study are summarized in Supplemental Table 3.1E. Microbial taxonomy assignments at the genus level were made from BLASTn against NT (E value of 0.001, the hit with lowest E value out of 10 hits was kept) for metagenomes and KAIJU (Menzel et al., 2016) for metatranscriptomes. Viral assignments were made by mapping reads against the viral reference genomes database (NCBI RefSeq, release 87) using SMALT (2010) with 80% identity. Fractional abundances were calculated using FRAP as previously described (Cobián Güemes et al., 2016) and expressed per million reads. After quality filtering and removing reads that mapped to the human genome, metatranscriptome D-8 reads were compared to the SILVA SSU database using BLASTn with an E-value cutoff of 0.001, and taxonomy was assigned at the genus level using the best hit from 10000 subsample replicates. Non-ribosomal reads were compared to the NCBI nucleotide database (NT) using BLASTn with an E-value cutoff of 0.001. The best hit was selected and used to assign bacterial taxonomy at the genus level. Species level assignments were determined by the genome that recruited the most reads for each genus either at the rRNA (*Bacillus, Escherichia,* and *Salmonella*) or mRNA (*Pseudomonas* and *Stenotrophomonas*) levels. The bacterial genome with more hits in the BLASTn analysis was selected as the closest strain and used as reference genome. rRNA and mRNA reads were mapped against each one of the reference genomes using SMALT with an identity cutoff of 60% and 80%, respectively, and the results were visualized using Anvi'o (Eren et al., 2015).

Reads from metatranscriptomes D-8 and D-7 were together assembled *de novo* using SPADES (Bankevich et al., 2012), and all resultant contigs were compared to NT using BLASTn with an E-value cutoff of 0.001; taxonomies were assigned using MEGAN6 (Huson et al., 2016). Contigs identified as *Pseudomonas* in all metatranscriptomes were separately mapped to the reference genome *P. aeruginosa* PAO1 using SMALT with an identity cutoff of 80%. *Pseudomonas* contigs (n=4965, a total of 2,686,355 base pairs) were annotated using PATRIC (Wattam et al., 2017); genes identified by subsystems classification as resistance to antibiotics and toxic compounds were summarized in Supplemental Table 3.3A . All contigs were screened for antibiotic resistance genes using the Resistance Gene Identifier implemented in the CARD database (Jia et al., 2017). All perfect and strict hits were retained, as was any hit with an identity ≥ 80%. Metatranscriptomes D-8 and D-7 reads were mapped to the proteins shiga-like toxin subunit A and subunit B using BLASTx with an E-value cutoff of 0.1 and identity of 60%. Fragment recruitment plots were generated using custom python scripts.

Samples comparison

Random forest, a non-parametric statistical method, was used to determine the bacterial genera that best differentiated between (1) antibiotic treatment and no antibiotic treatment in the metagenomes and (2) stable from exacerbation states in the metatranscriptomes. The importance of each variable was assessed using the R implementation of the algorithm random forest (Wiener, 2002) using 2000 trees.

The R package vegan (Jari et al., 2017) was used with the metatranscriptomes to calculate Pielou's evenness using Shannon diversity.

Metabolomics

LC-MS/MS metabolomics data were generated on the sputum sample D-8 and compared to a set of 15 samples routinely collected from the previous 426 days. Metabolite extraction (ethyl acetate and methanol), LC-MS/MS methods and data analysis were performed as described in (Quinn et al., 2016a). Data from these same sputum samples have been published previously (Quinn et al., 2016a), but the metabolites reported here were not presented in that study making these data novel (metabolomic data for this project are available under MassiVE dataset ID MSV000079444).

Metabolomics data processing

Metabolomics data were analyzed using molecular networking (Watrous et al., 2012) and GNPS (Wang et al., 2016). Molecular networking parameters were altered for this study and are as follows: cosine minimum of 0.7, 6 minimum matched peaks for spectral clustering, and a precursor mass and fragment ion mass tolerance of 0.1 Da. Molecular networks were visualized using the Cytoscape® software (Shannon et al., 2003). Molecules were annotated by searching the GNPS libraries and specific metabolites of interest were searched for using the MS^1 parent mass and then compared to the Metlin MS/MS spectral libraries (Smith et al., 2005). Area under the curve abundances of metabolites in the LC-MS/MS data were calculated using the mzMine 2 software (Pluskal et al., 2010) using selected masses. The parameters of the feature finding were as follows: minimum time span of 0.05 min, a minimum feature height of 2 and an m/z tolerance of 0.05 m/z or 15.0ppm. The chromatograms were deconvoluted, isotope peaks were grouped, and the peaks were aligned with the same ion mass tolerance and a retention time tolerance of 1 min. The final matrix of

features was gap filled. All metabolite annotations based on spectral alignment are considered level 2 according to the metabolomics proposed minimum reporting standards (Sumner et al., 2007).

Isolates of CF pathogens *P. aeruginosa* VVP172, *Enterococcus* sp. VVP100, *Escherichia coli* VVP427, *Streptococcus* sp. VVP047, *Stenotrophomonas maltophilia* VVP327 and *Staphylococcus aureus* VVP270 were obtained from the UCSD Center for Advanced Laboratory Medicine. These isolates were grown in artificial sputum medium according to the method of (Quinn et al., 2014) and their metabolomes were extracted using an ethyl acetate and methanol sequential extraction (the same method as for the sputum samples described in (Quinn et al., 2016a)). The LC-MS/MS data were generated with the same protocols as the sputum samples and the data were uploaded to GNPS. The MS/MS data from these bacterial isolates were used individually as a reference for searching for matching spectra in the CF01 longitudinal sputum data. Spectral matching parameters were as follows: parent and fragment mass tolerance of 0.1, minimum matched peaks of 6, cosine of 0.7 and a minimum spectral count of 3 in the dataset. Spectral matches between a sputum sample file and a bacterial isolate were summed for each sample for each bacterium and plotted to identify metabolite matches through the longitudinal datasets from pathogens known to be present in CF01 from clinical culture history (it must be noted these isolates were obtained from CF patients in the same clinic as CF01, but not from CF01). It is unknown if specific bacteria molecules were detected.

Data availability

Sequencing data is available on SRA with the study number SRP173673 (Nacional Center for Biotechnology Information, 2016). Metabolomics data is available on GNPS with the MassiVE dataset ID MSV000079444 (GNPS, 2019).

Acknowledgments

Chapter 3, in full, is published in mBIO. Ana Georgina Cobián Güemes, Yan Wei Lim, Robert A. Quinn, Douglas J. Conrad, Sean Benler, Heather Maughan, Rob Edwards, Thomas Brerrin, Vito Adrian Cantú, Daniel Cuevas, Rohaum Hamidi, Pieter Dorrestein and Forest Rohwer; 2019. The dissertation author was the primiary investigator and author of this paper.

We are grateful for the support received by Argonne National Laboratories staff members for their time and access to their large-scale systems, in particular to Tomas Brettin, Rick Stevens, Ross Overbeek and Robert Olson. We are thankful to Ty Roach, Nate Robinett, Douglas Naliboff, Sandi Calhoun and Mark Little for critical discussion of this work. This work was supported by Spruance Foundation. Ana Cobian was supported by CONACyT and UCMEXUS. The authors declare no competing interests.

References

Aendekerk, Séverine, Bart Ghysels, Pierre Cornelis, and Christine Baysse. 2002. "Characterization of a New Efflux Pump, MexGHI-OpmD, from Pseudomonas Aeruginosa That Confers Resistance to Vanadium." Microbiology 148 (8): 2371–81. https://doi.org/10.1099/00221287-148-8-2371.

Aigal, Sahaja, Julie Claudinon, and Winfried Römer. 2015. "Plasma Membrane Reorganization: A Glycolipid Gateway for Microbes." BBA - Molecular Cell Research 1853 (4): 858–71. https://doi.org/10.1016/j.bbamcr.2014.11.014.

Alexander, Bruce Marshall, Elbert Kristofer Petren, Samar Rizvi, Aliza Fink, Josh Ostrenga, Ase Sewall, and Deena Loeffler. 2016. "Cystic Fibrosis Foundation Patient Registry." Annual Data Report. Bethesda, Maryland. https://www.cff.org/Our-Research/CF-Patient-Registry/2015-Patient-Registry-Annual-Data-Report.pdf.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." Journal of Computational Biology 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bergan, Jonas, Anne Berit Dyve Lingelem, Roger Simm, Tore Skotland, and Kirsten Sandvig. 2012. "Shiga Toxins." Toxicon 60 (6): 1085–1107. https://doi.org/10.1016/j.toxicon.2012.07.016.

Betz, Josefi, Martina Bielaszewska, Andrea Thies, Hans-ulrich Humpf, Klaus Dreisewerd, Helge Karch, Kwang S Kim, Alexander W Friedrich, and Johannes Müthing. 2011. "Shiga Toxin Glycosphingolipid Receptors in Microvascular and Macrovascular Endothelial Cells: Differential Association with Membrane Lipid Raft Microdomains." Journal of Lipid Research 52. https://doi.org/10.1194/jlr.M010819.

Boyd, Beth, Gregory Tyrrell, Mark Maloney, Carlton Gyles, James Brunton, and Clifford Lingwood. 1993. "Alteration of the Glycolipid Binding Specificity of the Pig Edema Toxin from Globotetraosyl to Globotriaosyl Ceramide Alters In Vivo Tissue Targetting and Results in a Verotoxin L-like Disease in Pigs." Journal of Experimental Medicine 177 (June).

Brasfield, Dana, Guy Hicks, Seng-jaw Soong, James Peters, and Ralph Tiller. 1980. "Evaluation of Scoring System of the Chest Radiograph in Cystic Fibrosis." American Journal of Reoentgenology, 1195–98.

Cobián Güemes, Ana Georgina, Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton, and Forest Rohwer. 2016. "Viruses as Winners in the Game of Life."

Annual Review of Virology 3 (1): 197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Conrad, Douglas, Barbara Bailey, Jon A. Hardie, Per S. Bakke, Tomas M. L. Eagan, and Bernt B. Aarli. 2017. “Median Regression Spline Modeling of Longitudinal FEV1 Measurements in Cystic Fibrosis (CF) and Chronic Obstructive Pulmonary Disease (COPD) Patients.” Plos One 12 (12): e0190061. https://doi.org/10.1371/journal.pone.0190061.

Conrad, Douglas, Matthew Haynes, Peter Salamon, Paul B. Rainey, Merry Youle, and Forest Rohwer. 2013. “Cystic Fibrosis Therapy: A Community Ecology Perspective.” American Journal of Respiratory Cell and Molecular Biology 48 (2): 150–56. https://doi.org/10.1165/rcmb.2012-0059PS.

Deurenberg, Ruud H, Erik Bathoorn, Monika A Chlebowicz, Natacha Couto, Mithila Ferdous, Silvia García-cobos, Anna M D Kooistra-smid, et al. 2017. “Application of next Generation Sequencing in Clinical Microbiology and Infection Prevention.” Journal of Biotechnology 243: 16–24. https://doi.org/10.1016/j.jbiotec.2016.12.022.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for ‘omics Data.” PeerJ 3: e1319. https://doi.org/10.7717/peerj.1319.

Gallant, Claude V, Tracy L Raivio, Joan C Olson, Donald E Woods, and Douglas G Storey. 2015. “Pseudomonas Aeruginosa Cystic Fibrosis Clinical Isolates Produce Exotoxin A with Altered ADP- Ribosyltransferase Activity and Cytotoxicity,” no. 2000: 1891–99.

Gallegos, Karen M., Deborah G. Conrady, Sayali S. Karve, Thusitha S. Gunasekera, Andrew B. Herr, and Alison A. Weiss. 2012. “Shiga Toxin Binding to Glycolipids and Glycans.” PLoS ONE 7 (2). https://doi.org/10.1371/journal.pone.0030368.

GNPS. 2019. “GNPS Dataset MSV000079444.” https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0e8c1c0bc22745519be9d7147e74eff4&view=advanced_view.

Greninger, Alexander L, Samia N Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, et al. 2015. “Rapid Metagenomic Identification of Viral Pathogens in Clinical Samples by Real-Time Nanopore Sequencing Analysis.” Genome Medicine 7 (1): 99. https://doi.org/10.1186/s13073-015-0220-9.

Gu, W., E. D. Crawford, B. D. O’Donovan, M. R. Wilson, E. D. Chow, H. Retallack, and J. L. DeRisi. 2016. “Depletion of Abundant Sequences by Hybridization (DASH):

Using Cas9 to Remove Unwanted High-Abundance Species in Sequencing Libraries and Molecular Counting Applications." Genome Biology 17 (1): 41. https://doi.org/10.1186/s13059-016-0904-5.

Gupta, Sachin, Deepak Govil, Prem N Kakar, Om Prakash, Deep Arora, and Shibani Das. 2009. "Colistin and Polymyxin B: A Re-Emergence." Ijccm 13 (2): 49–53. https://doi.org/10.4103/0972-5229.56048.

Hankinson, Iohn L, John R Odencrantz, and Kathleen B Fedan. 1999. "Spirometrie Reference Values from a Sample of the General U . S . Population." American Journal of Respiratory and Critical Care Medicine 159: 179–87. https://doi.org/10.1164/ajrccm.159.1.9712108.

Huson, Daniel H., Sina Beier, Isabell Flade, Anna G??rska, Mohamed El-Hadidi, Suparna Mitra, Hans Joachim Ruscheweyh, and Rewati Tappu. 2016. "MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data." PLoS Computational Biology 12 (6): 1–12. https://doi.org/10.1371/journal.pcbi.1004957.

James, Chloe E., Emily V. Davies, Joanne L. Fothergill, Martin J. Walshaw, Colin M. Beale, Michael A. Brockhurst, and Craig Winstanley. 2015. "Lytic Activity by Temperate Phages of Pseudomonas Aeruginosa in Long-Term Cystic Fibrosis Chronic Lung Infections." ISME Journal 9 (6): 1391–98. https://doi.org/10.1038/ismej.2014.223.

Jari, Oksanen, Guillaume Blanchet F., Friendly Michael, Kindt Roeland, Legendre Pierre, and McGlinn Dan. 2017. "Vegan: Community Ecology Package." https://cran.r-project.org/package=vegan.

Jia, Baofeng, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, et al. 2017. "CARD 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database." Nucleic Acids Research 45 (D1): D566–73. https://doi.org/10.1093/nar/gkw1004.

Knowles, Michael R., and Mitchell Drumm. 2012. "The Influence of Genetics on Cystic Fibrosis Phenotypes." Cold Spring Harbor Perspectives in Medicine 2 (12): 1–13. https://doi.org/10.1101/cshperspect.a009548.

Krüger, Alejandra, and Paula M.A. Lucchesi. 2015. "Shiga Toxins and Stx Phages: Highly Diverse Entities." Microbiology (United Kingdom) 161 (3): 1–12. https://doi.org/10.1099/mic.0.000003.

Laguna, Theresa A., Brandie D. Wagner, Cynthia B. Williams, Mark J. Stevens, Charles E. Robertson, Cole W. Welchlin, Catherine E. Moen, Edith T. Zemanick, and Jonathan K. Harris. 2016. "Airway Microbiota in Bronchoalveolar Lavage Fluid

from Clinically Well Infants with Cystic Fibrosis." PLoS ONE 11 (12): 1–15. https://doi.org/10.1371/journal.pone.0167649.

Li, X. Z., H. Nikaido, and K. Poole. 1995. "Role of MexA-MexB-OprM in Antibiotic Efflux in Pseudomonas Aeruginosa." Antimicrobial Agents and Chemotherapy 39 (9): 1948–53. https://doi.org/10.1128/AAC.39.9.1948.

Lim, Yan Wei, Jose S. Evangelista, Robert Schmieder, Barbara Bailey, Matthew Haynes, Mike Furlan, Heather Maughan, Robert Edwards, Forest Rohwer, and Douglas Conrad. 2014. "Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis." Journal of Clinical Microbiology 52 (2): 425–37. https://doi.org/10.1128/JCM.02204-13.

Lim, Yan Wei, Matthew Haynes, Mike Furlan, Charles E. Robertson, J. Kirk Harris, and Forest Rohwer. 2014. "Purifying the Impure: Sequencing Metagenomes and Metatranscriptomes from Complex Animal-Associated Samples." Journal of Visualized Experiments, no. 94: 1–15. https://doi.org/10.3791/52117.

Lim, Yan Wei, Robert Schmieder, Matthew Haynes, Mike Furlan, T. David Matthews, Katrine Whiteson, Stephen J. Poole, et al. 2013. "Mechanistic Model of Rothia Mucilaginosa Adaptation toward Persistence in the CF Lung, Based on a Genome Reconstructed from Metagenomic Data." PLoS ONE 8 (5). https://doi.org/10.1371/journal.pone.0064285.

Lim, Yan Wei, Robert Schmieder, Matthew Haynes, Dana Willner, Mike Furlan, Merry Youle, Katelynn Abbott, et al. 2013. "Metagenomics and Metatranscriptomics: Windows on CF-Associated Viral and Microbial Communities." Journal of Cystic Fibrosis 12 (2): 154–64. https://doi.org/10.1016/j.jcf.2012.07.009.

Ling, Hong, Amechand Boodhoo, Bart Hazes, Maxwell D Cummings, Glen D Armstrong, James L Brunton, and Randy J Read. 1998. "Structure of the Shiga-like Toxin I B-Pentamer Complexed with an Analogue of Its Receptor Gb3." Biochemistry 2960 (97): 1777–88. https://doi.org/10.1021/bi971806n.

Lingwood, Clifford A, Adam Manis, Radia Mahfoud, Fahima Khan, Beth Binnington, and Murugesapillai Mylvaganam. 2010. "New Aspects of the Regulation of Glycosphingolipid Receptor Function." Chemistry and Physics of Lipids 163: 27–35. https://doi.org/10.1016/j.chemphyslip.2009.09.001.

LiPuma, John J. 2010. "The Changing Microbial Epidemiology in Cystic Fibrosis." Clinical Microbiology Reviews 23 (2): 299–323. https://doi.org/10.1128/CMR.00068-09.

Melton-Celsa, Angela R. 2014. "Shiga Toxin (Stx) Classification, Structure, and Function Angela." Microbiol Spectr 2 (2): 1–21. https://doi.org/10.1128/microbiolspec.EHEC-0024-2013.Shiga.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. “Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju.” Nature Communications 7: 1–9. https://doi.org/10.1038/ncomms11257.

Moons, Pieter, David Faster, and Abram Aertsen. 2013. “Lysogenic Conversion and Phage Resistance Development in Phage Exposed Escherichia Coli Biofilms.” Viruses 5 (1): 150–61. https://doi.org/10.3390/v5010150.

Nacional Center for Biotechnology Information. 2016. “SRA Study Number SRP173673.” 2016. https://www.ncbi.nlm.nih.gov/sra/?term=SRP173673.

National Center for Biotecnology Information. 2016a. “Contamination in Sequence Databases.” 2016. https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/.

“The UniVec Database.” 2016. https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.

National Human Genome Research Institute, NIH. 2018. “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).” 2018. www.genome.gov/sequencingcostsdata.

Obrig, T G, R M Seaner, M Bentz, C A Lingwood, B Boyd, A Smith, and W Narrow. 2003. “Induction by Sphingomyelinase of Shiga Toxin Receptor and Shiga Toxin 2 Sensitivity in Human Microvascular Endothelial Cells” 71 (2): 845–49. https://doi.org/10.1128/IAI.71.2.845.

Obrig, Tom G. 2010. “Escherichia Coli Shiga Toxin Mechanisms of Action in Renal Disease.” Toxins 2 (12): 2769–94. https://doi.org/10.3390/toxins2122769.

Pielou, Evelyn Chrystalla. 1979. Biogeography. Wiley.

Pluskal, Tomáš, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. 2010. “MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data.” BMC Bioinformatics 11 (1): 395. https://doi.org/10.1186/1471-2105-11-395.

Quinn, Robert A., Yan Wei Lim, Tytus D. Mak, Katrine Whiteson, Mike Furlan, Douglas Conrad, Forest Rohwer, and Pieter Dorrestein. 2016. “Metabolomics of Pulmonary Exacerbations Reveals the Personalized Nature of Cystic Fibrosis Disease.” PeerJ 4: e2174. https://doi.org/10.7717/peerj.2174.

Quinn, Robert A, Jose A Navas-molina, Embriette R Hyde, Jin Song, Yoshiki Vázquez-baeza, Greg Humphrey, James Gaffney, et al. 2016. “From Sample to Multi-Omics Conclusions in under 48 Hours.” MSystems 1 (2). https://doi.org/10.1128/mSystems.00038-16.Editor.

Quinn, Robert A, Vanessa V Phelan, Katrine L Whiteson, Neha Garg, Barbara A Bailey, Yan Wei Lim, Douglas J Conrad, Pieter C Dorrestein, and Forest L Rohwer. 2015. "Microbial, Host and Xenobiotic Diversity in the Cystic Fibrosis Sputum Metabolome." The ISME Journal 095384 (6): 1–16. https://doi.org/10.1038/ismej.2015.207.

Quinn, Robert A, Katrine Whiteson, Yan-wei Lim, Peter Salamon, Barbara Bailey, Simone Mienardi, Savannah E Sanchez, Don Blake, Doug Conrad, and Forest Rohwer. 2014. "A Winogradsky-Based Culture System Shows an Association between Microbial Fermentation and Cystic Fibrosis Exacerbation" 9 (4): 1024–38. https://doi.org/10.1038/ismej.2014.234.

Saladin, Michèle, V. T B Cao, Thierry Lambert, Jean L. Donay, Jean Louis Herrmann, Zahia Ould-Hocine, Charlotte Verdet, Françoise Delisle, Alain Philippon, and Guillaume Arlet. 2002. "Diversity of CTX-M β-Lactamases and Their Promoter Regions from Enterobacteriaceae Isolated in Three Parisian Hospitals." FEMS Microbiology Letters 209 (2): 161–68. https://doi.org/10.1016/S0378-1097(02)00484-6.

Sandvig, Kirsten, Anne Berit Dyve Lingelem, Tore Skotland, and Jonas Bergan. 2012. "Shiga Toxins: Properties and Action on Cells." In The Comprehensive Sourcebook of Bacterial Protein Toxins, 4th ed., 1085–1107. Elsevier. https://doi.org/10.1016/j.toxicon.2012.07.016.

Schmidt, K., S. Mwaigwisya, L. C. Crossman, M. Doumith, D. Munroe, C. Pires, A. M. Khan, et al. 2017. "Identification of Bacterial Pathogens and Antimicrobial Resistance Directly from Clinical Urines by Nanopore-Based Metagenomic Sequencing." Journal of Antimicrobial Chemotherapy 72 (1): 104–14. https://doi.org/10.1093/jac/dkw397.

Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." Bioinformatics 27 (6): 863–64. https://doi.org/10.1093/bioinformatics/btr026.

Seitz, Aaron P., Heike Grassmé, Michael J. Edwards, Yael Pewzner-Jung, and Erich Gulbins. 2015. "Ceramide and Sphingosine in Pulmonary Infections." Biological Chemistry 396 (6–7): 611–20. https://doi.org/10.1515/hsz-2014-0285.

Shannon, C.E., and W. Weaver. 1949. The Mathematical Theory of Communication. Edited by The University of Illinois Press. Urbana.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction

Networks." Genome Research 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

"SMALT Manual." 2010. October, no. 1: 1–7.

Smith, Colin A, Grace O'Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. 2005. "METLIN: A Metabolite Mass Spectral Database." Therapeutic Drug Monitoring 27 (6): 747–51.

Sumner, Lloyd W, Alexander Amberg, Dave Barrett, Michael H Beale, Richard Beger, Clare A Daykin, Teresa W-M Fan, et al. 2007. "Proposed Minimum Reporting Standards for Chemical Analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)." Metabolomics : Official Journal of the Metabolomic Society 3 (3): 211–21. https://doi.org/10.1007/s11306-007-0082-2.

Surette, Michael G. 2014. "The Cystic Fibrosis Lung Microbiome." Annals of the American Thoracic Society 11 (SUPPL. 1): 61–65. https://doi.org/10.1513/AnnalsATS.201306-159MG.

Uchida, H, N Kiyokawa, T Taguchi, H Horie, J Fujimoto, and T Takeda. 1999. "Shiga Toxins Induce Apoptosis in Pulmonary Epithelium-Derived Cells." The Journal of Infectious Diseases 180 (6): 1902–11. https://doi.org/10.1086/315131.

Vasil, Michael L, Martin J Stonehouse, Adriana I Vasil, Sandra J Wadsworth, Howard Goldfine, Robert E Bolcome Iii, and Joanne Chan. 2009. "A Complex Extracellular Sphingomyelinase of Pseudomonas Aeruginosa Inhibits Angiogenesis by Selective Cytotoxicity to Endothelial Cells" 5 (5). https://doi.org/10.1371/journal.ppat.1000420.

Wandro, Stephen, Lisa Carmody, Tara Gallagher, John J. LiPuma, and Katrine Whiteson. 2017. "Making It Last: Storage Time and Temperature Have Differential Impacts on Metabolite Profiles of Airway Samples from Cystic Fibrosis Patients." MSystems 2 (6): e00100-17. https://doi.org/10.1128/mSystems.00100-17.

Wang, Mingxun, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, et al. 2016. "Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking." Nature Biotechnology 34 (8): 828–37. https://doi.org/10.1038/nbt.3597.

Watrous, Jeramie, Patrick Roach, Theodore Alexandrov, Brandi S Heath, and Jane Y Yang. 2012. "Mass Spectral Molecular Networking of Living Microbial Colonies" 109 (26): 1743–52. https://doi.org/10.1073/pnas.1203689109.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. "Improvements to PATRIC, the All-

Bacterial Bioinformatics Database and Analysis Resource Center." Nucleic Acids Research 45 (D1): D535–42. https://doi.org/10.1093/nar/gkw1017.

Whelan, Fiona J., Alya A. Heirali, Laura Rossi, Harvey R. Rabin, Michael D. Parkins, and Michael G. Surette. 2017. "Longitudinal Sampling of the Lung Microbiota in Individuals with Cystic Fibrosis." PLoS ONE 12 (3): 1–17. https://doi.org/10.1371/journal.pone.0172811.

Whiteson, Katrine L., Simone Meinardi, Yan Wei Lim, Robert Schmieder, Heather Maughan, Robert Quinn, Donald R. Blake, Douglas Conrad, and Forest Rohwer. 2014. "Breath Gas Metabolites and Bacterial Metagenomes from Cystic Fibrosis Airways Indicate Active PH Neutral 2,3-Butanedione Fermentation." ISME Journal 8 (6): 1247–58. https://doi.org/10.1038/ismej.2013.229.

Wiener, A. Liaw and M. 2002. "Classification and Regression by RandomForest." R News.

Willner, Dana, Matthew R. Haynes, Mike Furlan, Nicole Hanson, Breeann Kirby, Yan Wei Lim, Paul B. Rainey, et al. 2012. "Case Studies of the Spatial Heterogeneity of DNA Viruses in the Cystic Fibrosis Lung." American Journal of Respiratory Cell and Molecular Biology 46 (2): 127–31. https://doi.org/10.1165/rcmb.2011-0253OC.

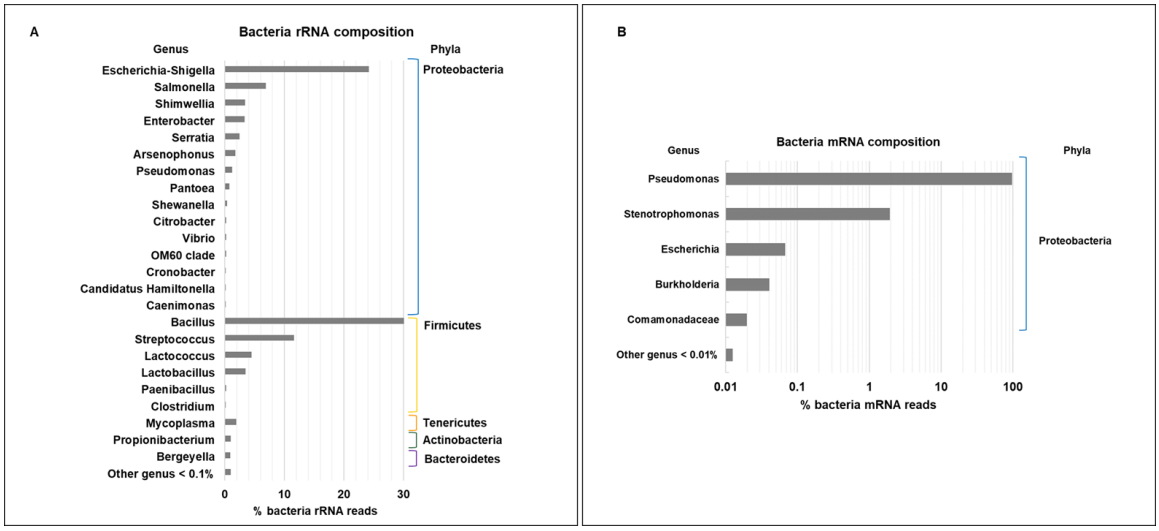
Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A Major Update to the DrugBank Database for 2018." Nucleic Acids Research 46 (D1): D1074–82. https://doi.org/10.1093/nar/gkx1037.

Woyke, Tanja, Alexander Sczyrba, Janey Lee, Christian Rinke, Damon Tighe, Scott Clingenpeel, Ramunas Stepanauskas, and Jan-fang Cheng. 2011. "Decontamination of MDA Reagents for Single Cell Whole Genome Amplification" 6 (10): 2–6. https://doi.org/10.1371/journal.pone.0026161.

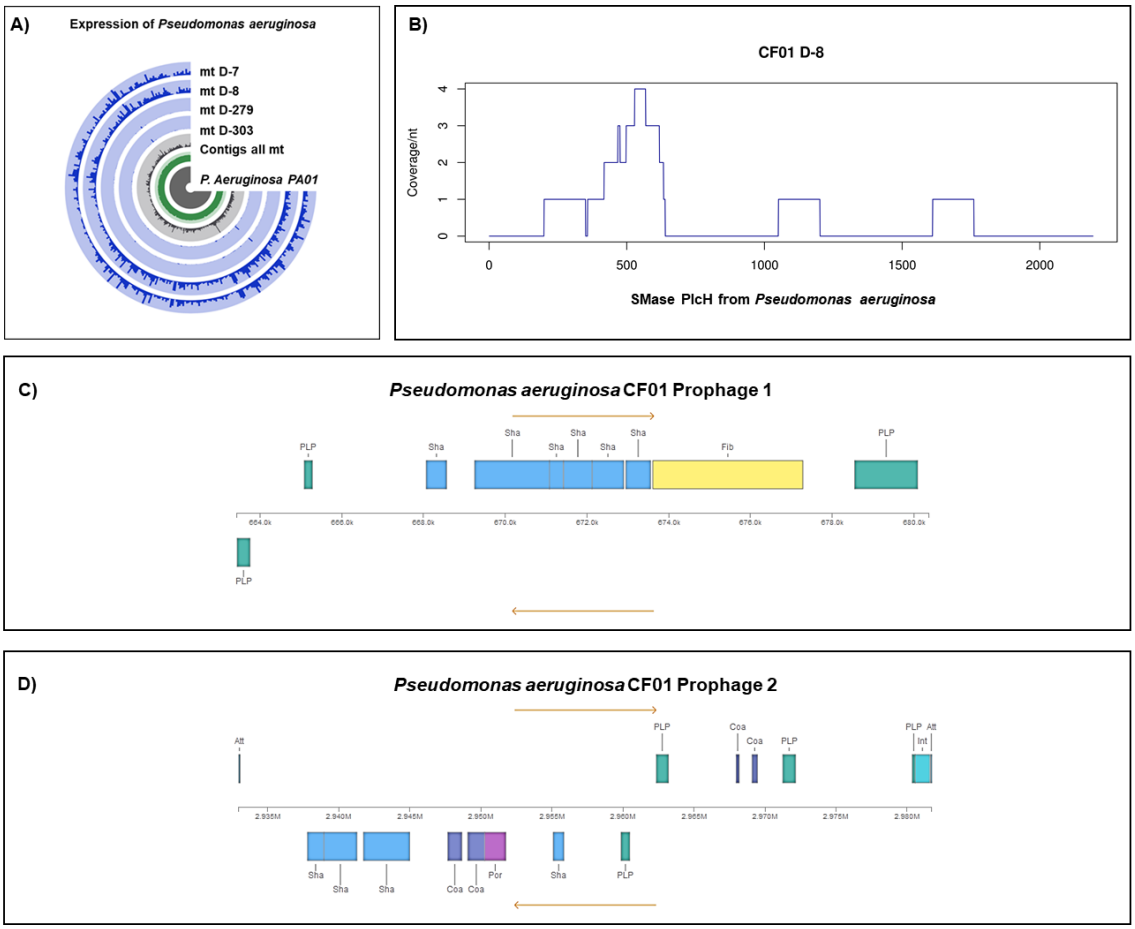
Zhao, Jiangchao, Patrick D Schloss, Linda M Kalikin, Lisa A Carmody, Bridget K Foster, Joseph F Petrosino, James D Cavalcoli, et al. 2012. "Decade-Long Bacterial Community Dynamics in Cystic Fibrosis Airways." Proceedings of the National Academy of Sciences of the United States of America 109 (15): 5809–14. https://doi.org/10.1073/pnas.1120577109.

Appendix for Chapter 3

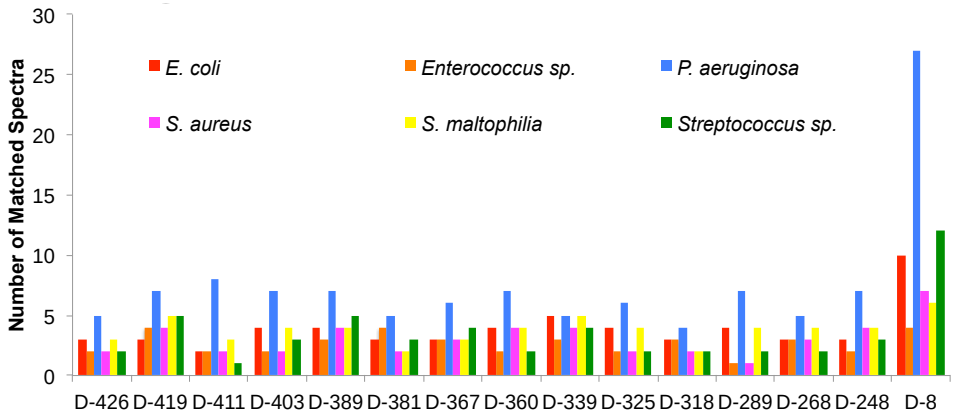
Supplemental figures



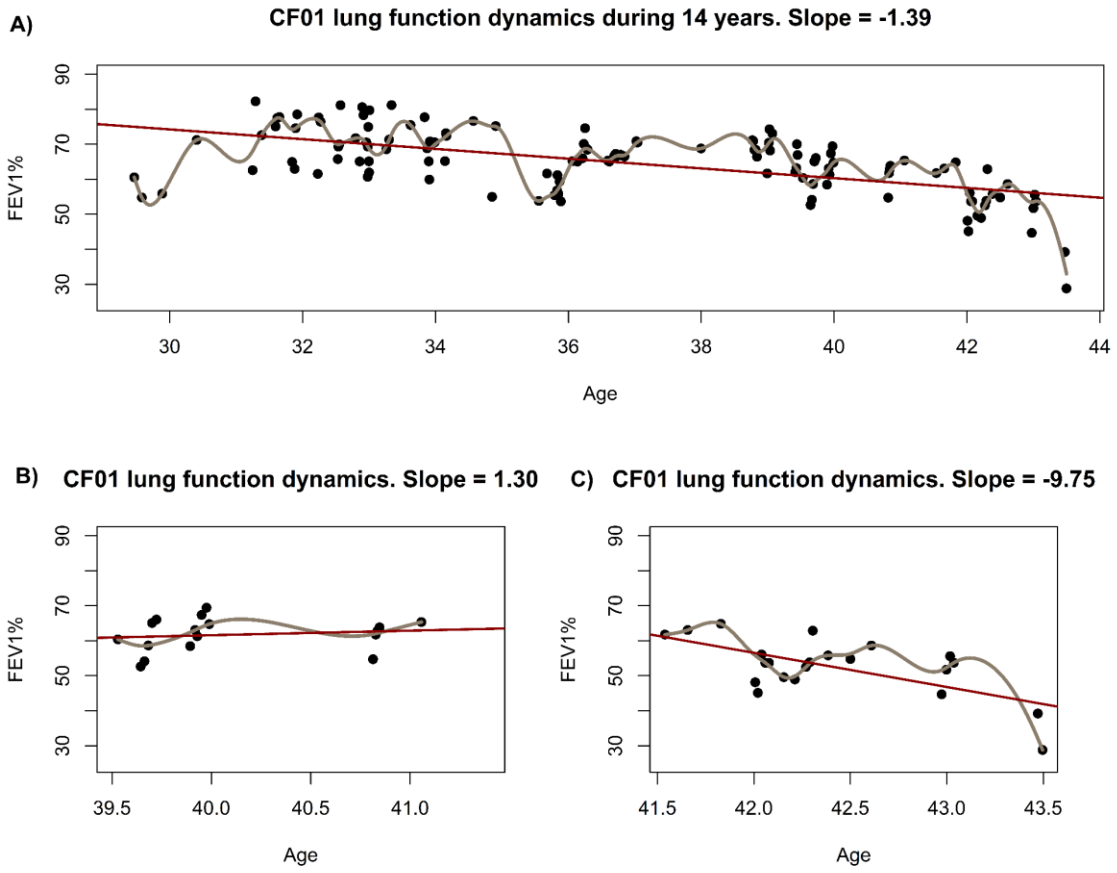
Supplemental Figure 3.1 Total bacterial RNA composition during a fatal exacerbation (sample D-8). A) Bacterial ribosomal RNA composition was assigned with BLASTn against the SILVA SSU database, with an E-value cutoff of 0.001. The best hit from 10,000 subsample replicates was used. Results are shown at the genus level. B) Bacterial non-ribosomal RNA composition at genus level assigned by BLASTn vs NT with an E-value cutoff of 0.001. The best hit was selected.



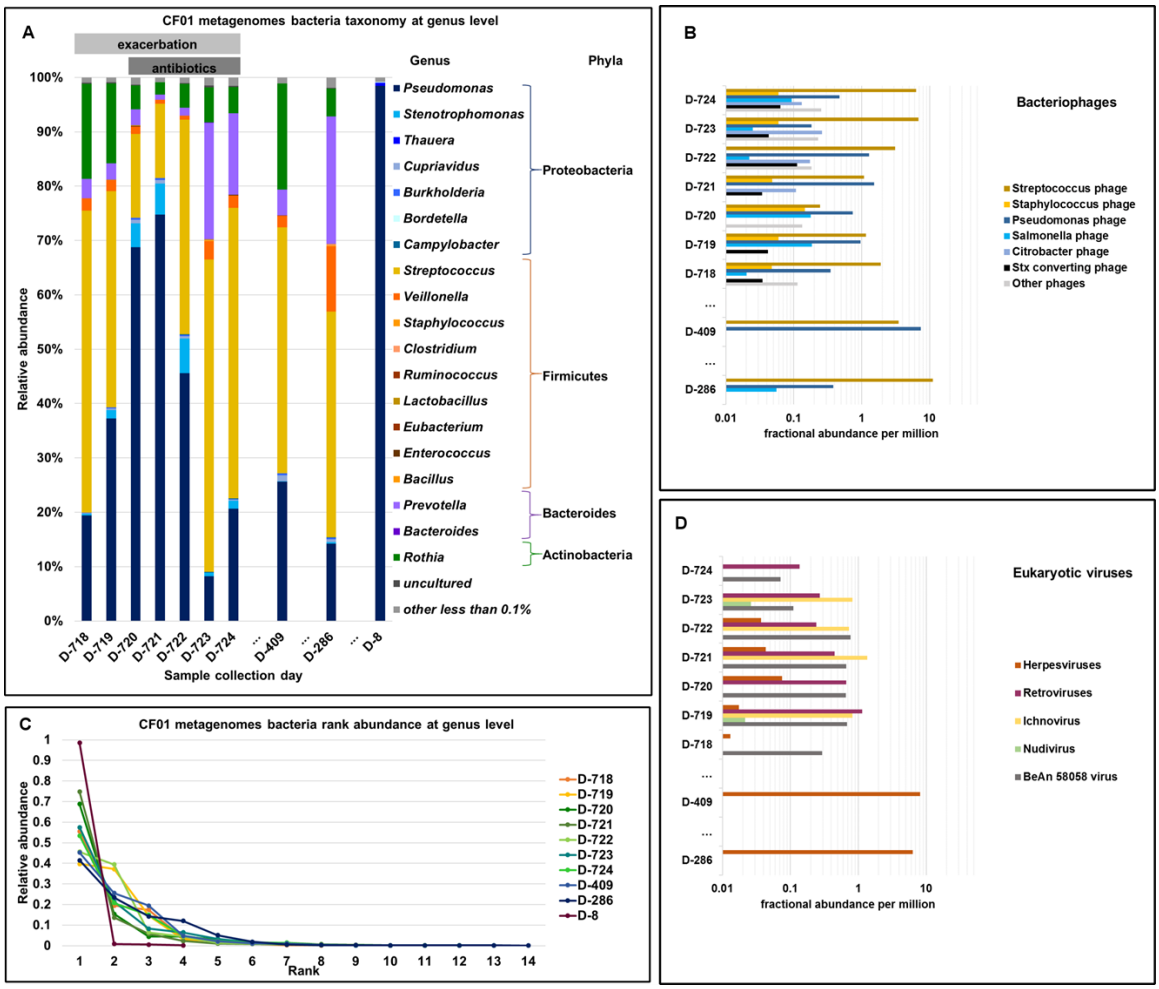
Supplemental Figure 3.2 A) *P. aeruginosa* gene expression during a stable period (D-279 & D-303) and fatal exacerbation (D-7 & D-8) based on fragment recruitment to the *P. aeruginosa* PAO1 reference genome. All microbial reads from metatranscriptomes D-303, D-279, D-8, and D-7 were individually mapped to the reference genome with SMALT using an identity cutoff 80%. Reads from samples D-8 and D-7 were *denovo* assembled into contigs. Contigs were mapped to the reference genome using SMALT with an identity cutoff 80%. B) *P. aeruginosa* SMase *plcH* coverage plot. Reads from metatranscriptome D-8 were mapped to the gene PlcH using SMALT at 80% identity cutoff. C) Predicted prophages from assembled genome *P. aeruginosa* CF01. Prophages were predicted using the online version of PHASTER. Protein annotations for partial prophage 1. PLP: phage like protein, Sha: tail protein, Fib: fiber protein. D) Predicted prophages from assembled genome *P. aeruginosa* CF01. Prophages were predicted using the online version of PHASTER Protein annotations for complete prophage 2. PLP: phage like protein, Sha: tail protein, Fib: fiber protein, Coa: coat protein, Por: portal protein, Att: attachment site.



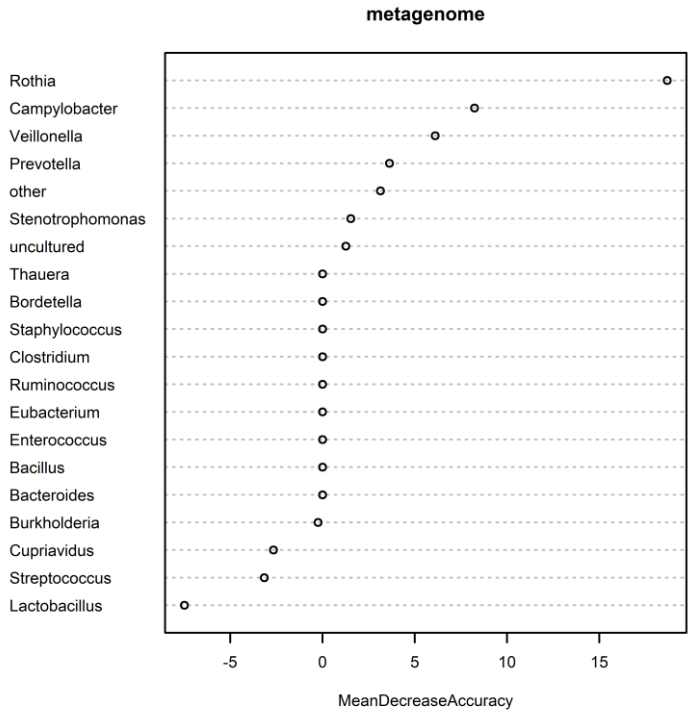
Supplemental Figure 3.3 Metabolomes from sample D-8 and their comparison to historical samples for CF01. Spectra from CF01 historical samples and fatal exacerbation sample D-8 were mapped to the individual spectra from known bacteria.



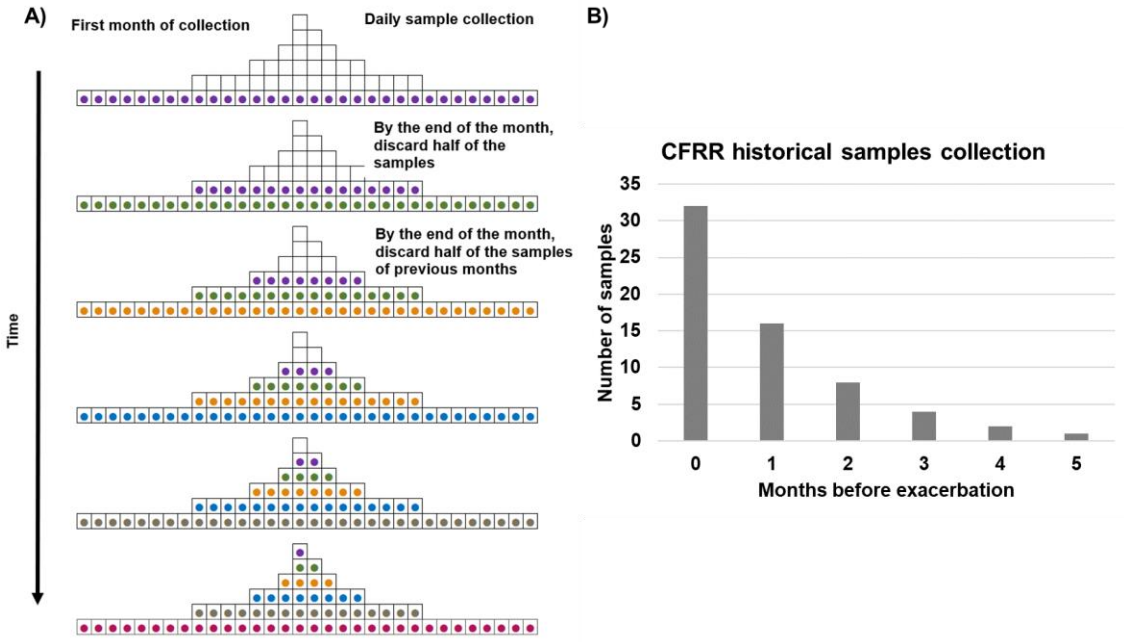
Supplemental Figure 3.4 A) Percentage of predicted FEV1 of patient CF01 for 14 years. B) Percentage of predicted FEV1 of patient CF01 for years 4 and 3 before death. C) Percentage of predicted FEV1 of patient CF01 for last two years of life. In all panels measurements obtained are represented by black dots. The grey line is the calculated spline as described by Conrad et al. The red line is the fitting of a linear model for the measurements shown in each panel. The slope of the linear model fitting in shown in the header of each panel as the slope.



Supplemental Figure 3.5 Metagenomic analysis was performed from sputum samples collected over a seven-day exacerbation period, during a subsequent stable period of 10 to 14 months and fatal exacerbation. A) Bacterial taxonomy was obtained at the genus level using BLASTn against NT. Relative abundances were calculated for genera whose abundances were greater than or equal to 0.1%. The phylum to which each genus belongs is indicated by a similar color gradient. B) The fractional abundances of phages obtained by mapping to ViralRefseq and FRAP normalization. C) Bacteria rank abundance plots for CF01 metagenomes described in panel A. Relative abundances are shown at genus level, genera with a relative abundance lower than 0.1% were not included in the plots D) The fractional abundances of eukaryotic viruses obtained by mapping to ViralRefseq and FRAP normalization.



Supplemental Figure 3.6 Variable importance plot using mean decrease accuracy for a supervised random forest with 5000 trees. Taxonomical relative abundance at genus level in metagenome samples of an exacerbation event two years before CF01's death. During the exacerbation event, two groups of samples were analyzed: antibiotic treatment and no antibiotic treatment.



Supplemental Figure 3.7 Sampling scheme for collection of "historical" sputum samples. A) Proposed at-home sample collection scheme where sputum samples are obtained daily. By the end of the first collection month, half of the samples are discarded (purple dots) and daily collection continues for the new month. By the end of the second month, half of all the previous months samples are discarded (purple and green dots) and so on. B) In an acute exacerbation event, the patient will bring the samples to the clinic and the CFRR methods will be applied to acute exacerbation and historical samples. With the proposed at home collection scheme, a higher sampling density will be obtained in the exacerbation month and less in the months before the exacerbation.

Supplemental tables

Supplemental Table 3.1A Brasfield scores of CF01 X-rays. X-rays were performed as part of the patient's regular clinical care.

Day	Air trapping	Linear markings	Nodular cystic lesions	Large lesions	General severity	Brasfield score
D-1	3	4	4	5	4	5
D-8	2	3	3	0	3	14
D-193	2	2	2	0	2	16

Supplemental Table 3.1B Hematology of CF01 during their last month of life. White blood cell counts are expressed in cells per microliter.

Day	White Blood Cells count	segmented neutrophils
D-0	24,000	96%
D-1	17,400	96%
D-2	12,400	95%
D-7	7,100	97%
D-23	10,900	95%

Supplemental Table 3.1C Bacteria and fungi cell culture results from the clinical microbiology laboratory for patient CF01 during their last two years of life.

Day	Status	Bacteria	Fungi
D-0	exacerbation	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia*	*Aspergillus terreus*
D-1	exacerbation	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia*	
D-7	exacerbation	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia*	Yeast
D-14	exacerbation	*Pseudomonas aeruginosa*	Yeast
D-24	exacerbation	*Pseudomonas aeruginosa* *Enterobacter cloacae*	Yeast
D-48	stable	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia* *Enterobacter cloacae*	Yeast
D-192	exacerbation	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia*	*Aspergillus terreus* Yeast
D-204	exacerbation	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia*	Yeast
D-373	stable	*Pseudomonas aeruginosa*	Yeast
D-414	stable	*Pseudomonas aeruginosa*	Yeast
D-540	stable	*Pseudomonas aeruginosa*	Yeast
D-547	stable	*Pseudomonas aeruginosa*	Yeast
D-674	stable	*Pseudomonas aeruginosa*	
D-719	exacerbation	*Pseudomonas aeruginosa* *Pseudomonas fluorescens*	Yeast *Aspergillus fumigatus*

Supplemental Table 3.1D Antibiotic received as treatment during the last two years of CF01's life. Class and mechanism of action were obtained from PubChem and DrugBank.

Month	Antibiotic	Class	Comments	Mechanism of action
M-0	Aztreonam	monobactam		Inhibit synthesis of bacteria cell wall, binds to and inactivates penicillin-binding-protein-3.
	Azithromycin	macrolide → azalide		Inhibit protein synthesis, reversible binding to 50S ribosomal subunit,
	- sulfa	sulfonamide		Interfere with folic acid synthesis, competition for the enzyme dihydropteorate synthetase
	- quinolone	quinolone		Inhibits DNA gyrase
	Colistin	cationic polypeptide	ER response	Solubilize cell membrane through a detergent like mechanism.
	Meropenem	beta-lactam → carbapenem	ER response	Inhibits cell wall synthesis, penetrates cell wall to reach penicillin-binding-protein targets.
M-1	Aztreonam	monobactam		Inhibit synthesis of bacteria cell wall, binds to and inactivates penicillin-binding-protein-3.
	Azithromycin	macrolide → azalide		Inhibit protein synthesis, reversible binding to 50S ribosomal subunit,
	- quinolone	quinolone		Inhibits DNA gyrase
	- sulfa	sulfonamide		Interfere with folic acid synthesis, competition for the enzyme dihydropteorate synthetase
M-6	Azithromycin	macrolide → azalide		Inhibit protein synthesis, reversible binding to 50S ribosomal subunit,
	- sulfa	sulfonamide		Interfere with folic acid synthesis, competition for the enzyme dihydropteorate synthetase
	- quinolone	quinolone		Inhibits DNA gyrase
	Meropenem	beta-lactam	ER response	Inhibits cell wall synthesis, penetrates cell wall to reach penicillin-binding-protein targets.
M-24	Doxycycline	tetracycline		Inhibit protein synthesis, reversible binding to 30S ribosomal subunit and possibly 50S.
	Ciprofloxacin	fluoroquinolone		Inhibits topoisomerase II (DNA gyrase) and topoisomerase IV

Supplemental Table 3.1E Metagenome and metatranscriptome sequencing overview for CF01 sputum samples. All libraries were sequenced on the Illumina platform. The Nextera library prep kit was used for all metagenomes whereas TruSeq was used for metatranscriptomes.

Day	Status	Library type	File name	SRA ID
D-724	Exacerbation	metagenome	polihed_CF01mgD724.fasta	SAMN10605062
D-723	Exacerbation	metagenome	polished_CF01mgD723.fasta	SAMN10605061
D-722	Exacerbation	metagenome	polished_CF01mgD722.fasta	SAMN10605060
D-721	Exacerbation	metagenome	polished_CF01mgD721.fasta	SAMN10605059
D-720	Exacerbation	metagenome	polished_CF01mgD720.fasta	SAMN10605058
D-719	Exacerbation	metagenome	polished_CF01mgD719.fasta	SAMN10605057
D-718	Exacerbation	metagenome	polished_CF01mgD718.fasta	SAMN10605056
D-409	Stable	metagenome	polished_CF01mgD409.fasta	SAMN10605055
D-286	Stable	metagenome	polished_CF01mgD286.fasta	SAMN10605054
D-8	Exacerbation	metagenome	polished_CF01mgD8.fasta	SAMN10605053
D-303	Stable	metatranscriptome	polished_CF01mtD303.fasta	SAMN10605052
D-279	Stable	metatranscriptome	polished_CF01mtD279.fasta	SAMN10605051
D-8	Exacerbation	metatranscriptome	polished_CF01mtD8.fasta	SAMN10605050
D-7	Exacerbation	metatranscriptome	polished_CF01mtD7.fasta	SAMN10605049

Supplemental Table 3.2A Comparison of molecules spectra between non-exacerbation samples (D-426 to D-248) and exacerbation sample D-8. P-values were calculated from a single tail normal distribution (pnorm function in R).

Molecule	p-value	z score
Globotriaosylceramide	~0	512960.92
Lactosylceramide	~0	44.60
Sphingomyelin	0.90	-1.28

Supplemental Table 3.2B Comparison of number of specific bacteria spectra between non-exacerbation samples (D-426 to D-248) and exacerbation sample D-8. P-values were calculated from a single tail normal distribution (pnorm function in R).

Bacteria spectra in D-8	CF strain ID	p-value	z score
Escherichia coli	VVP427	1.99e-17	8.41
Enterococcus sp.	VVP100	0.043	1.71
Pseudomonas aeruginosa	VVP172	3.38e-69	17.5
Staphylococcus aureus	VVP270	2.58e-05	4.04
Stenotrophomonas maltophilia	VVP327	0.005	2.51
Streptococcus sp.	VVP047	2.39e-14	7.53

Chapter 4 : Mobile genetic elements in Cystic Fibrosis exacerbations

Abstract

Phages and microbes colonize the respiratory airways of patients with Cystic Fibrosis causing persistent infections that impact lung function. Per treatment regimes, the microbial communities in Cystic Fibrosis respiratory airways is continuously exposed to antibiotics which confers enhanced antibiotic resistant to the population of bacterial pathogens present in the Cystic Fibrosis lung. In this work the microbiome of a group of patients whose lung function declined sharply and were non-responsive to the antibiotics treatment was studied through metagenomics. The lung microbiomes of Cystic Fibrosis patients in this study were characterized by enhanced bacterial growth with a reduction in community diversity and elevated lytic phage production. Four fatal exacerbations were dominated by *Pseudomonas aeruginosa* or *Achromobacter* spp. and four non-fatal exacerbations were dominated by *Stenotrophomonas maltophilia, Mycobacterium avium-intracellulare* or *Streptococcus salivarius*. The dominant bacteria in these exacerbations have between 3% and 6% of genome insertions, of which >60% of the coding sequences are phage-derived. Phages that encode toxins which damage the host tissue were identified, such as *Stenotrophomonas phage* phiSHP2 that encodes zonula occludens toxin. This work suggests phage activity in Cystic Fibrosis exacerbations as a likely mechanism of increased pathogenic virulence.

Introduction

Cystic Fibrosis (CF) is a recessive genetic disease in which malfunctioning of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) causes insufficient anion exchange across membranes and dehydration of mucosal surfaces. Mucus dysfunction in CF airways results in chronic polymicrobial infections (Laguna et al. 2016). Overtime, the vigorous innate and adaptive host immune responses to these chronic infections result in airway remodeling, gas exchange abnormalities and eventually respiratory failure. Airway clearance treatments, antibiotics, anti-inflammatory drugs, and new CFTR modulator therapies have extended median lifespan of CF patients, yet these polymicrobial lung infections persist as the major drivers of CF morbidity and mortality for the foreseeable future (Alexander et al. 2016).

Common bacteria that chronically colonize CF airways include *Pseudomonas aeruginosa, Staphylococcus aureus, Stenotrophomonas maltophilia, Achromobacter* spp., *Haemophilus influenzae,* and *Burkholderia cepacia* (Surette 2014; LiPuma 2010; Lim et al. 2013; Whiteson et al. 2014), but every CF individual has a unique microbial community that changes over time (Whelan et al. 2017; Zhao et al. 2012; Lim et al. 2014). This highlights the need to characterize the microbial communities temporally in each CF patient. Viruses (eukaryotic and bacteriophage) contribute to these unique microbial signatures among CF patients by acting as remodeling agents of bacterial communities (Reyes et al. 2015). How phage lytic-lysogenic life cycles provide top down control (James et al. 2015) of specific bacterial populations and drive bacterial population rank abundance and metagenomic composition is largely unexplored in the CF lung environment (Silveira and Rohwer 2016).

An accurate description of the dynamics between these viruses and their bacterial hosts is critical to understanding the drivers of the microbial community dynamics, the capacity for virulent metabolic processes, and the subsequent host immune response. Phage are known to transfer exotoxins (Wilson and Ho 2006; Krüger and Lucchesi 2015; Dobrindt et al. 2015), antibiotic resistance genes (Budzik et al. 2004), and virulence factors (Busby, Kristensen, and Koonin 2013) to bacterial populations which can confer enhanced virulence in host-pathogen systems. Detailed exploration of phage insertions in CF bacterial infections is needed to understand CF disease progression.

In this work we applied a Cystic Fibrosis Rapid Response strategy (Cobián-Güemes et al. 2019) to reveal genetic exchange between viral and bacterial populations during CF exacerbations. To accomplish this, a personalized multi-omics approach was used to characterize viromes (Willner et al. 2012), metagenomes and metatranscriptomes (Lim et al. 2013) in order to comprehensively monitor the microbial and viral dynamics during CF pulmonary exacerbations. The CF exacerbations of each patient were all dominated by a single pathogenic bacteria, resulting in low community diversity and were associated with host tissue damage via toxins. Using a tailored bioinformatics approach, we identified 1) previously unidentified pathogens in the CF lung, 2) toxin production and 3) unique mobile elements.

Results

Clinical characterization of Cystic Fibrosis pulmonary exacerbations.

To investigate the microbial community composition of CF acute exacerbations, eight patients in the UCSD adult Cystic Fibrosis clinic that suffered pulmonary exacerbations (CFPE) and required hospitalization within a one-year interval were included in this study (Table 4.1). Four patients suffered fatal exacerbations (CF01, CF094, CF116 and CF418) and four patients survived the exacerbation event (CF318, CF409, CF292 and CF146). Exacerbations were defined by the treating clinician and were characterized by a decrease in lung function (measured as FEV1), no response to antibiotics treatment, and general health decline.

The study population age ranged from 18 to 40 years. Exacerbations ranged from 8 to 63 days and a loss of lung function relative to each patient baseline ranged between 10% to 48%. The clinical laboratory reported the presence of *Pseudomonas aeruginosa* (CF01, CF094, CF146), *Stenotrophomonas maltophilia* (CF01, CF409, CF318), *Achromobacter* sp. (CF116), *Achromobacter xylosoxidans* (CF116, CF418), *Mycobacterium avium-intracellulare* (CF292, CF318), Multidrug Resistant *Staphylococcus aureus* (CF409), *Aspergillus terreus* (CF01), and yeast (not *Cryptococcus neoformans)* (CF409, CF318, CF292).

Table 4.1 Cystic Fibrosis patients clinical data during exacerbations (Ex.) that required hospitalization.

Ex. Outcome	Patient ID	Gender	Age	Ex. FEV1 loss	Ex. length (days)	Clinical microbiology during Ex.	Antibiotics used during Ex.
Fatal	CF01	Male	37	10%	8	●*Pseudomonas aeruginosa* ●*Stenotrophomonas maltophilia* ●*Aspergillus terreus*	aztreonam, azithromycin, colistin, meropenem
Fatal	CF094						
Fatal	CF116	Male	30	43%	27	●*Achromobacter* species ●*Achromobacter xylosoxidans*	ceftazidime-avibactam, doxycycline, sulfamethoxazole-trimethroprim, vancomycin, tigecycline, colistin, azithromycin, meropenem, imipenem-cilastatin, minocycline
Fatal	CF418						
Non-fatal	CF409	Female	18	40%	63	●*MRSA* ●*Stenotrophomonas maltophilia* ●*Aspergillus fumigatus* ●Yeast, not *Cryptococcus neoformans*	sulfamethoxazole-trimethroprim, ceftaroline, minocycline, linezolid, ceftazidime, tobramycin
Non-fatal	CF318	Male	27	29%	37	●*Stenotrophomonas maltophilia* ●*Mycobacterium avium-intracellulare* ●Yeast, not *Cryptococcus neoformans*	levofloxacin, sulfamethoxazole-trimethroprim, minocycline, meropenem
Non-fatal	CF292	Female	34	48%	53	●*Mycobacterium avium-intracellulare* ●Yeast, not *Cryptococcus neoformans*	aztreonam, linezolid, moxifloxacin, amikacin, azithromycin, ethambutol, rifampin
Non-fatal	CF146	Male	31	36%	14	●*Pseudomonas aeruginosa*	colistin, piperacillin-tazobactam

Cystic Fibrosis exacerbations display higher microbial loads and phage production.

Viral and microbial counts and virus to microbe ratios from the CF lungs were obtained for exacerbation samples and for a set of stable CF samples (n=16) (Figure 4.1). Exacerbation samples had an average bacterial abundance of 2.9 x 10^9 cells ml^{-1}, an average viral abundance of 2.5 x 10^{10} VLPs ml^{-1}, and an average virus to microbe ratio of 12.5. Stable samples were nearly an order of magnitude lower than exacerbation samples with an average bacterial abundance of 5.8 x 10^8 cells ml^{-1}, an average viral abundance of 3.1 x 10^9 VLPs ml^{-1}, and a virus to microbe ratio of 6.5. Exacerbation samples showed a significantly higher virus to microbe ratio than stable samples, suggesting enhanced bacteria growth and viral activity in CF exacerbations.

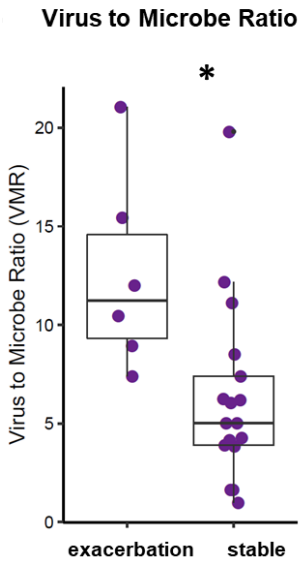


Figure 4.1 Virus to Microbe Ratio (VMR) in CF sputum samples quantified by epifluorescence microscopy. Exacerbation and stable samples VMR are significantly different (T-test p-value < 0.05, Wilcoxon test p-value 0.010 and w=88).

Microbial community composition of pulmonary exacerbations.

The microbial community composition of the described CFPEs was obtained from sputum or bronchioalveolar lavage samples and characterized molecularly using total DNA metagenomes or total RNA metatranscriptomes. The exacerbation microbial communities presented low diversity in both evenness and richness (Supplemental Figure 4.1) and were dominated by a single microbe (Figure 4.2). Diversity indexes were compared to a cohort of CF patients suffering mild exacerbations (Losada et al. 2016). Acute exacerbations presented here have significantly lower diversity than mild exacerbations. Species-level resolution was obtained from each dataset by comparing the microbial reads to bacterial reference genomes at a high identity (>96% identity over 100% of the read length). *P. aeruginosa* was the most active microbe in two fatal exacerbations, CF01 (96% relative abundance) and CF094 (98% relative abundance), as shown by transcriptomic analysis. *Achromobacter* spp. was the dominant genus in two fatal exacerbations. In CF116, *Achromobacter ruhlandii* relative abundance was 80% and in CF418 *Achromobacter xylosoxidans* relative abundance was 76%. In non-fatal exacerbations, the microbial community was dominated by *Stenotrophomonas maltophilia* (CF318 and CF409), *Mycobacterium intracellulare* (CF292) and *Streptococcus salivarius* (CF146) (Figure 4.2). *Propionibacterium acnes* was detected in all samples and is probably ubiquitous in CF respiratory samples. The yeast *Candida glabrata* was identified in CF292 (Supplemental Figure 4.2).

In exacerbations dominated by the genus *Achromobacter,* other members of the microbial community had a very low fractional abundance. In *Stenotrophomonas*-dominated exacerbations, the genera that followed in abundance were *Rothia, Streptococcus* and

Achromobacter. In Patient CF292, the genus *Mycobacterium* had the highest fractional abundance followed by the genera *Rothia* and *Streptococcus*. Patient CF146 microbial community was more diverse, *Streptococcus* was the genera with higher fractional abundance, followed by *Rothia*, *Neisseria* and *Prevotella.*

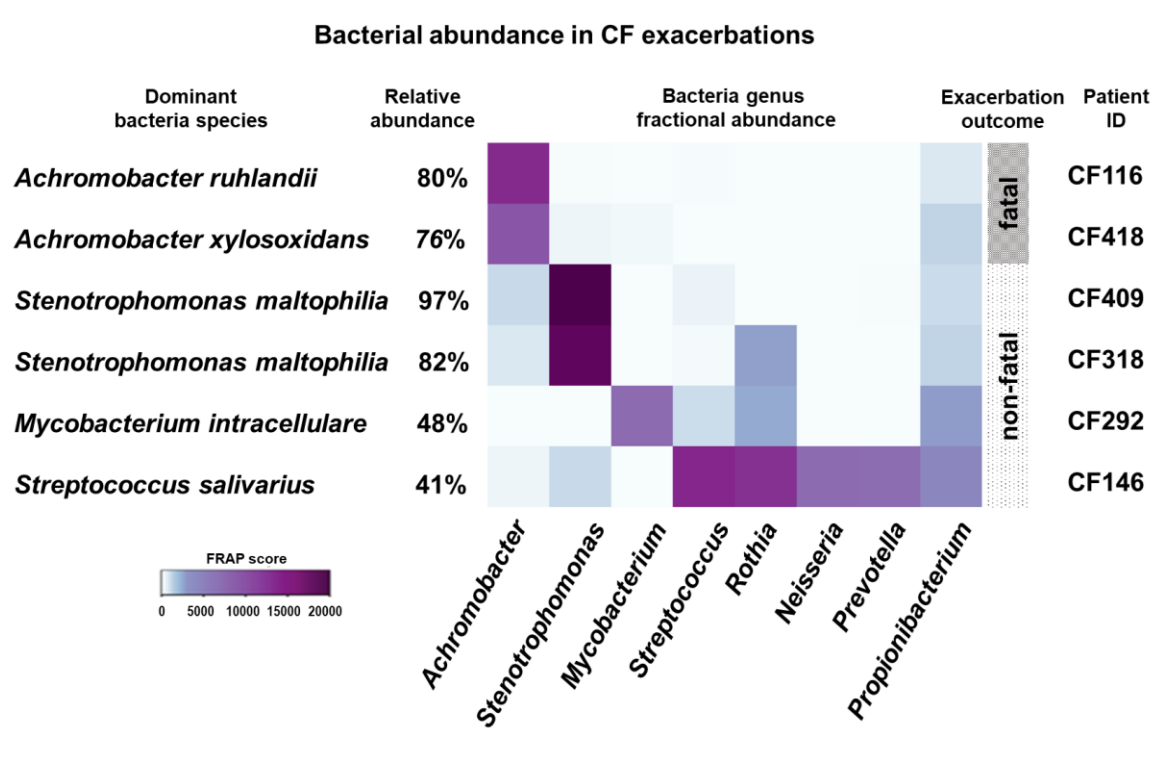


Figure 4.2 Bacterial abundance in CF exacerbations metagenomes. Polished reads were mapped to bacteria RefSeq database (n=66,000 genomes) using fragment recruitment assembly purification (FRAP) at 96% identity over 100% of the read. Fractional abundances from individual genomes were added by bacteria genera.

Temporal dynamics of microbial community composition

Longitudinal sampling of patient CF318 for a month before the acute exacerbation event showed that the microbial diversity (evenness) decreased from 0.88 to 0.28 (Supplemental Figure 4.3-A). The dominant genera in this microbial community was *Stenotrophomonas* (Supplemental Figure 4.16-B) during the sampling period, but the absolute abundance increased by an order of magnitude (from ~$2x10^8$ cells ml^{-1} to ~$1x10^9$ cells ml^{-1}). An increase in relative abundance of virulence factors was also observed during the exacerbation period (Supplemental Figure 4.3-C), which may be explained by the colonization of a different *Stenotrophomonas* species carrying these virulence factors.

Genomic insertions and deletions in CF pulmonary exacerbations.

Genome plasticity appeared common in the dominant bacteria present in CF exacerbation metagenomes, therefore insertions and deletions were quantified based on comparisons to the closest reference genome. Deletions that met a criteria of > 10,000 nucleotides were observed between 13 and 24 excised regions in the most abundant bacteria present in each patient metagenome (Supplemental Figures 4.4 to 4.9). These deleted coding sequences represented between 5% and 13% of the bacterial reference genomes. Annotation of these excised regions also showed that between 10% and 66% of the excised regions coding sequences were phage related (Table 4.2).

Table 4.2 Deletions in dominant genomes.

Patient ID	Reference genome	Deletions (nt)	% of genome deletions	Phage and hypothetical CDS in deletions (%)
CF116	*Achromobacter ruhlandii* (CP017433.1)	327,280	5.14	71.72
CF418	*Achromobacter xylosoxidans* strain FDAARGOS_150 (CP014028.1)	376,964	6.01	63.32
CF409	*Stenotrophomonas maltophilia* strain FDAARGOS_325 (CP022053.2)	410,006	8.45	79.91
CF318	*Stenotrophomonas maltophilia* strain FDAARGOS_325 (CP022053.2)	347,490	7.16	81.93
CF292	*Mycobacterium intracellulare (CP023149.1)*	859,906	15.32	58.68
CF146	*Streptococcus sp.* (CP014264.1)	285,568	13.11	60.11

Insertions were determined as regions >10,000 nucleotides present in metagenome assembled contigs but not in the closest reference genome (Supplemental Figures 4.10 to 4.12). Three bacterial genomes had enough coverage to obtain quality sequence assemblies in order to detect insertion regions (*A. ruhlandii*, and *S. maltophilia* strains from two different patients; Table 4.3). Between 3% and 6% of these assembled bacterial genomes were annotated as insertions.

Table 4.3 Insertions in dominant genomes.

ID	Microbial community dominant bacteria	Nucleotides in contigs > 1000 nt	Insertions (nt)	% of genome that are insertions	Phage and hypothetical CDS in insertions (%)
CF116	*Achromobacter ruhlandii*	6,636,892	283,932	4.2	77.1
CF318	*Stenotrophomonas maltophilia*	4,845,908	190,605	3.9	67.8
CF409	*Stenotrophomonas maltophilia*	4,917,770	311,160	6.3	68.4

Annotations of these inserted regions showed that between 67% and 77% of the coding sequences are phage-related. In *A. ruhlandii* assembled from CF116 metagenome, 12 insertions were identified, two of which were annotated with high confidence as phages (Figure 4.3-A). From 13 inserted regions identified in *S. maltophilia* assembled from CF318 metagenome, three of the regions encode phage-related proteins and two of them encode integrases (Figure 4.3-B). In *S. maltophilia* assembled from CF409, 14 inserted regions were identified, two of which are phages (Figure 4.3-C). *S. maltophilia* 409 insertion 14 is a 61.4 Kb region with phage replication and structural coding sequences (Figure 4.4-A). *S. maltophilia* 409 insertion 7 is a 23.3 Kb region inside a contig which is 51 Kb and is flanked by tRNAs (Figure 4.4-B). Both the insertion and the contig have coding sequences annotated as phage replication and structural proteins.

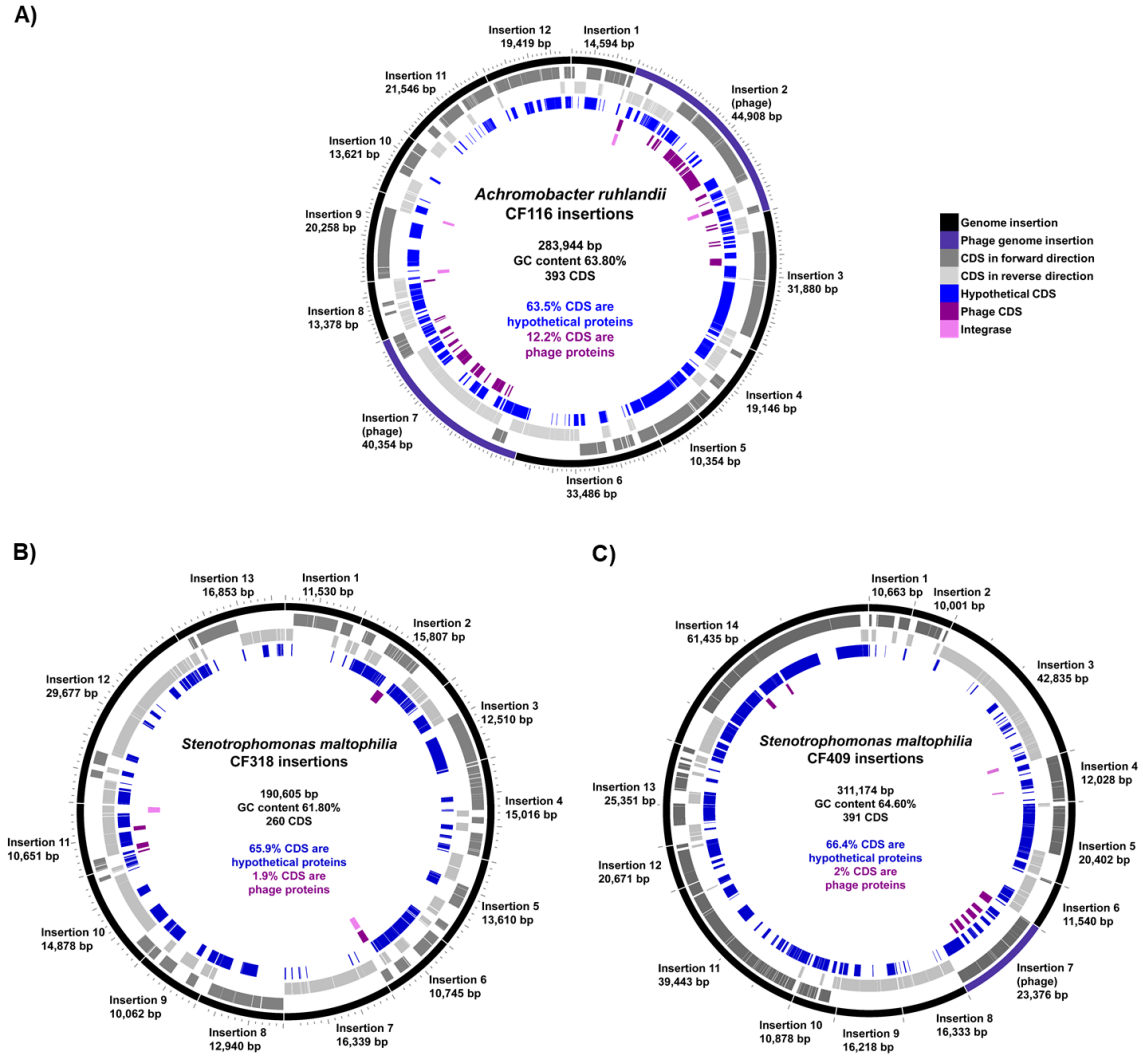


Figure 4.3 Insertions in CF exacerbation dominant genomes. Insertions are defined as regions >10,000 bp that are present in metagenome assembled contigs but are not present in the closest reference genome. A) *Achromobacter ruhlandii* CF116 B) *Stenotrophomonas maltophilia* CF318 C) *Stenotrophomonas maltophilia* CF409.

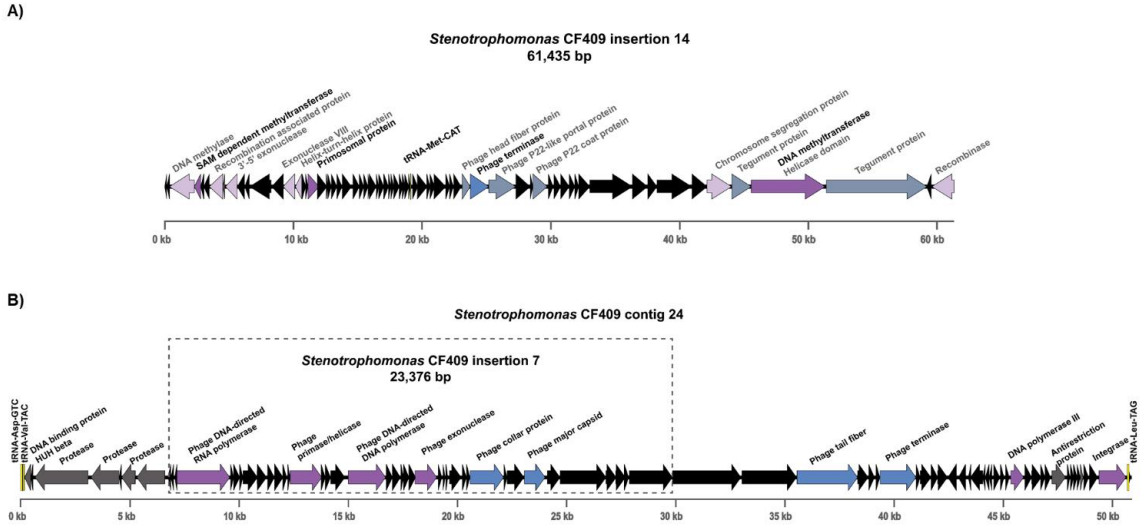


Figure 4.4 *Stenotrophomonas* insertions are phages. Annotations from PATRIC are in black. Annotation from the Conserved Domains Database are in grey. A) *S. maltophilia* CF409 insertion 14. B) *S. maltophilia* CF409 insertion 7 and the complete contig where the insertion was detected.

Phage activity in pulmonary exacerbations

The dominant microbes present in CF exacerbations are lysogens carrying known and uncharacterized prophages in their genomes. In exacerbations that led to patient mortality, an *Achromobacter phage* was detected in CF116 whose dominant bacteria was *A. xylosoxidans*. Lysogens in CFPEs that were not fatal were also observed, such as a *Streptococcus phage* detected in CF146 whose dominant bacteria was *S. salivarius*, and *Stenotrophomonas phages* were detected in CF409 whose dominant bacteria was *S. maltophilia*. Additional phages were detected in both fatal and non-fatal exacerbations such as, *Burkholderia phage, Pseudomonas phage, Staphylococcus phage, Salmonella phage* and nine other phages (Supplemental Figure 4.13-B).

Phages are known to carry toxins that can enhance replication and virulence of bacterial pathogens. An exotoxin carried by *Stenotrophomonas phage* phiSHP2 (Hagemann, Hasse, and Berg 2006) was identified during CF409 exacerbation. *Stenotrophomonas phage* phiSHP2 is a 5.8 Kb phage (Figure 4.5-A) that encodes nine proteins, only two of them with functional annotations: a replication protein and zonula occludens toxin. CF409 microbial community was dominated by *S. maltophilia* carrying a prophage that was induced during the exacerbation, as detected by the recovery of the viral genome in the total metagenome (Figure 4.5-B and C) and viral particle-enriched virome (Figure 4.5-D). This bacteriophage was carrying the exotoxin zonula occludens which disrupts tight junctions in epithelial cells (Di Pierro et al. 2001).

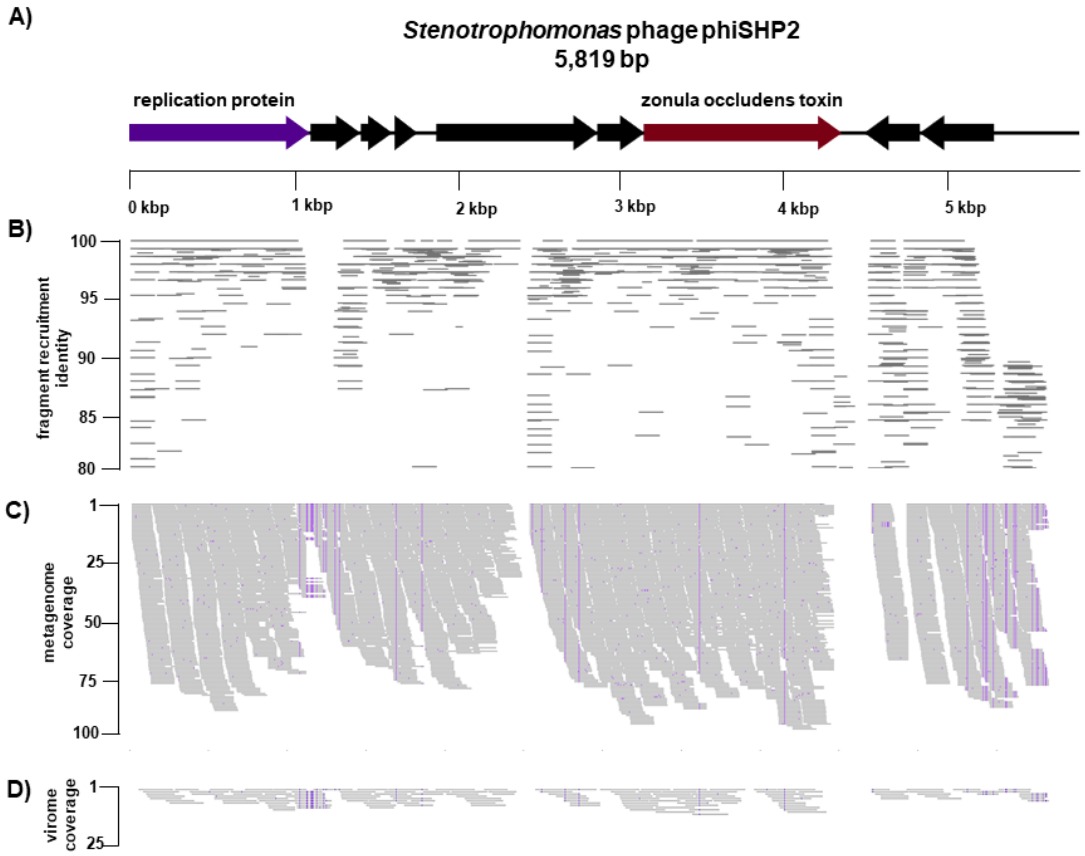


Figure 4.5 *Stenotrophomonas phage* SHP2 and zonula occludens toxin. CF409 exacerbation sample metagenome in which *Stenotrophomonas* relative abundance was 97%. A) *S. maltophilia* phage *phiSHP2* complete genome annotation, a replication protein and zonula occludens toxin are annotated, the remaining 7 ORFs are annotated as hypothetical proteins. B) Fragment recruitment plot. C) Coverage plot from metagenome. D) Coverage plot from virome.

Discussion

CF exacerbations and bacteria genomic insertions

The microbial communities of the CF exacerbations explored in this study were dominated by distinctive bacteria strains that had unique genomic insertions. Such insertions included prophages encoding toxins and other mobile elements carrying antibiotic resistance genes. These results illustrate that precise identification of not only strain level, but unique genomic material in the CF microbiome is essential to understand each CF exacerbation.

Achromobacter in CF exacerbations

Achromobacter is an emerging genera in CF exacerbations (Amoureux et al. 2013; Rudkjøbing et al. 2012; Ridderberg, Nielsen, and Nørskov-Lauritsen 2015). In this study, the lung microbial community of two CF fatal exacerbations was dominated by *A. ruhlandii* and *A. xylosoxidans*, respectively. Potential mechanisms contributing to the development of fatal exacerbations by *Achromobacter* spp. include direct attack to the host tissue via toxins, resistance to antibiotics (Bador et al. 2013), and the ability to perform aerobic and anaerobic respiration (Jakobsen et al. 2013). Hemolysins were detected in both fatal exacerbations (Supplemental Figure 4.13-C); these hemolytic proteins are known to be carried by *Achromobacter* and can directly damage host tissues (Benz 2015). In addition to hemolysins, *Achromobacter* can deliver other virulence factors to host tissues using type III secretion systems and attack competing bacteria with the protein colicin V (Swenson and Sadikot 2015; Jakobsen et al. 2013; Rodrigues et al. 2016). These factors are posited to contribute to *Achromobacter* pathogenicity in the CF lung, leading to fatal exacerbations.

Stenotrophomonas in CF exacerbations

Stenotrophomonas acquisition is related to a decrease in lung function in CF (Waters et al. 2013; Barsky et al. 2017). In this study, two exacerbations were dominated by *S. maltophilia*, one of which carried the *Stenotrophomonas phage* phiSHP2 encoding zonula occludens toxin (Figure 4.5). *S. phage* phiSHP2 was active in this exacerbation, since its genome was identified in viral particles. Zonula occludens toxin is an exotoxin that induces actin depolymerization, which provokes opening of tight junctions and disruption of epithelial integrity (Di Pierro et al. 2001; Fasano et al. 1991; Uzzau et al. 2001). This mechanism

possibly contributed to the disruption of lung epithelial cells (Ruan et al. 2014) in this CF exacerbation.

Unique mobile elements in CF exacerbations

Phages and genomic islands are the main components of a unique CF mobilome in every exacerbation presented in this study. The mobilome facilitates pathogen adaptation to the local environment (Dobrindt et al. 2015; Jeukens et al. 2017). The detailed comparative genomics methods presented here allowed for the identification of already characterized phages such as *Stenotrophomonas phage* phiSHP2, previously uncharacterized genomic islands, and new prophages identified in inserted regions of *Achromobacter* spp. and *Stenotrophomonas* spp.

Microbial ecology models of acute CF exacerbations.

The CF mobilome is an essential component to understand the pathogenesis of CF microbial communities. In this study, acute exacerbations were characterized by low microbial diversity and the dominance of a single microbe with a unique mobilome. Toxins that directly attack host tissues were also present in the described CF exacerbations. These exacerbations were distinct from CF exacerbations previously studied by our group (Lim et al. 2014; Quinn et al. 2014; Whiteson et al. 2014) and others (Moran Losada et al. 2016; Feigelman et al. 2017). Therefore, the following model for descriptions of the microbial ecology in CF acute exacerbations is proposed: pathogens dominance, in which acute exacerbations are characterized by low diversity, dominance of a single microbe with unique mobile elements including toxins and lysogenic phages. In this model, the CF mobilome plays an essential role to understand and treat CF acute exacerbations.

Materials and methods

Clinical data.

Sample collection procedures and access to clinical data were approved by the Institutional Review Board of University of California San Diego (HRPP 081510) and San Diego State University (IRB#1711018R).

Samples collection and pre-processing.

Sputum samples were collected by hospital personnel during the patient's stay at the hospital. Expectorated phlegm was collected in a sterile cup and stored at 4 ºC while transported to the research lab. In the research lab samples were homogenized with a syringe (no needle) and distributed into 500µl aliquots.

Viral and microbial enumeration.

Sputum samples were homogenized in SM buffer (1:6 dilution) and treated with DNAse (1000 U ml^{-1}) for 1h. Next, they were filtered sequentially through Whatman Nucleopore Track-Etched Membranes of 8 um and 2 um pore sizes to remove large particles and eukaryotic cells. The filtrate was fixed with paraformaldehyde (2%), stained with SYBR Gold (Life Technologies, USA) and filtered on a 0.02 um Anodisc membrane (Whatman) for epifluorescence microscopy. Viral and microbial particles were quantified based on size (viral-like-particles < 0.2 um, and bacterial cells > 0.2 um) in at least 10 images for each sample. The dilution factors were used to calculate viral particles or cells per ml.

Total DNA metagenomes.

Five hundred microliters of sputum were transferred to a cryovial. The tube was submerged in a dry-ice and ethanol bath for 5 minutes, then transferred to a water bath at 100ºC for 5 minutes, this process was repeated 3 times. The sample was transferred to the beads tube from Qiagen power soil DNA extraction kit and homogenized by shaking for 45 minutes. The rest of the Qiagen power soil DNA extraction kit protocol was used. Ten nanograms of DNA were used for Nextera library prep. Libraries were sequenced on Illumina MiSeq using 150 cycles as single end. This procedure is illustrated in Supplemental Figure 4.14

Total RNA metatranscriptomes.

Five hundred microliters to 2 ml of sputum were mixed with 4 mL of guanidinium thiocyanate (TRIzol, Invitrogen), homogenized by vortexing and stored at -80 ºC until further processing. Samples were defrosted and 0.2 volume of chloroform was added, the mixture was homogenized by vortexing and incubated for 20 minutes at 4 ºC. Samples were centrifuged at 13,800 x G for 20 minutes at 4 ºC and the aqueous layer was transferred to a new RNase free tube. Aqueous phase volume was measured and an equal volume of isopropanol and 2µl of glycogen (20mg ml^{-1}) were added, the mixture was incubated at 4 ºC for 20 minutes to precipitate nucleic acids. The sample was centrifuged at 13,800 x G for 20 minutes at 4 ºC and the supernatant was removed and discarded. The pellet was washed with 1 ml of 75% ice cold ethanol and centrifuged 5 minutes at 4 ºC, this procedure was repeated 2 times. The pellet was air dried for 15 minutes and resuspended in 50 µl of molecular grade water and incubated at 55 ºC for 5 minutes. DNase treatment was performed by adding 2 µl of

TURBO DNase and 5 µl of DNase buffer, the mixture was incubated at 37 ºC for 30 minutes, 0.2 volume of DNase inactivation reagent was added and incubated 5 minutes at room temperature. The sample was centrifuged at 10,000 x G for 1.5 minutes at 24 ºC and the aqueous phage was transferred to a new RNase free tube. RNA concentration was measured using Qubit (between 50 and 200 ng μl^{-1} were obtained in all samples). Total RNA sequencing libraries were prepared using Illumina TruSeq total RNA without ribosomal RNA depletion and no fragmentation. One µg of total RNA in 8.5 µl of molecular grade water was used as input for library prep. RNA was mixed with 8.5 µl of fragment prime finish mix and 5 µl of 5X first strand buffer and incubated at 65 ºC for 5 minutes, at 72 ºC for 5 minutes and 4 ºC for 5 minutes. Next 1 µl of superscript II reverse transcriptase and 8 µl of first strand mix ActD were added and the protocol followed without modifications, in the last PCR 15 cycles were used. When libraries were ready the volume was brought up to 50 µl and a right-side selection was performed using SPRIselect beads in a ratio of 0.9X. Libraries quality was evaluated using Bioanalyzer. Libraries were sequenced on Illumina MiSeq using 150 cycles as single end.

Virome

Sputum samples were homogenized in SM buffer (1:6 dilution) and treated with DNAse (1000 U ml^{-1}) for 1h. at 37 ºC Next, they were filtered sequentially through 8 um, 2 um and 0.45 um filters to remove large particles and eukaryotic cells. The filtrate was mixed with chloroform (10 %) and homogenized by hand for 5 min. The sample was centrifuged at 3,000 g for 5 min and the chloroform was removed. CsCl was added to the chloroform-treated sample at 1.15 g ml^{-1} density. The sample was was added onto a CsCl gradient comprised of 4

layers: 1.7 g ml^{-1}, 1.5 g ml^{-1}, 1.35 g ml^{-1} and 1.2 g ml^{-1} prepared with SM buffer. The sample was centrifuged at 4 ºC for 14 h at 35,000 rpm in a Beckman Coulter Ultracentrifuge using the rotor SW41 (151263 x g average and 210053 x g max). The 1.5 g ml^{-1} fraction was collected after the centrifugation using a needle and transferred to an epi tube. A 50 ul aliquot was checked for purity using epifluorescence microscopy, as described above. The remaining sample was subjected to DNA extraction using the formamide and CTAB protocol (Thurber et al 2009). The DNA was prepared for sequencing using the Accel-NGS 2S library prep kit (Swift Biosciences) using Covaris fragmentation (Covaris). Library quality was evaluated using Bioanalyzer and the libraries were sequenced on the Illumina MiSeq Platform.

Clinical isolates genome sequencing.

Stenotrophomonas spp. and *Achromobacter* spp. isolates were characterized at the clinical lab. Liquid cultures were grown overnight in LB media at 37ºC. Cells were pelleted and resuspended in 200µl of PBS, DNA extraction was performed using DNeasy blood and tissue kit (Qiagen). Four hundred nanograms of gDNA were used for Nanopore sequencing, libraries were prepared and sequenced in individual flow cells on the MinION instrument. Reads were base called using ALBACORE, error correction and contigs assembly were performed with CANU(Koren et al. 2017) with an estimated genome length of 5Mb. Contigs were annotated using RAST(Overbeek et al. 2014) server.

Mycobacterium spp. isolates were characterized at the clinical lab. A scrap of the bacteria growing on the solid media tube (1.5g) was homogenized in a cryovial with buffer from the Qiagen power soil DNA extraction kit. The tube was submerged in a dry-ice and ethanol bath for 5 minutes, then transferred to a water bath at 100ºC for 5 minutes, this

process was repeated 3 times. The sample was transferred to the beads tube from Qiagen power soil DNA extraction kit and homogenized by shaking for 45 minutes. The rest of the Qiagen power soil DNA extraction kit protocol was used, and DNA was obtained at a concentration of 20 ng μl^{-1} in 200 µl.

Metagenomes, metatranscriptomes and viromes data analysis.

Raw fastq files were filtered and dereplicated using PRINSEQ++ with minimum quality threshold 20, dereplication and entropy threshold 50. Cloning vector sequences were removed using SMALT with 80 % identity against the NCBI UniVec database. Human genome sequences were removed using SMALT with 80 % identity against the human reference genome GRCh38 (Supplemental Figure 4.15). The remaining reads are refered as polished reads. Polished reads were mapped to the NCBI RefSeq database of complete bacterial genomes using SMALT at 96 % identity. Hits were normalized using Fragment Recruitment Assembly Purification (FRAP, code available at https://github.com/yinacobian/frap). This procedure is illustrated in Supplemental Figure 4.16

Genomic insertions and deletions identification.

The closest reference genome for each dataset was identified as the genome that recruited more reads at 96% identity over 100% of the read. Regions >10,000 bp with no coverage in the reference genome were identified and extracted from the reference genome (code available at https://github.com/yinacobian/getholes). Insertions were identified by mapping *de novo* assembled polished reads (SPAdes –only-assembly) to the closest reference genome using MEDUSA. Regions >10,000 bp that were present in assembled contigs but not

in the reference genome were identified as insertions and extracted from the contigs (code available at https://github.com/yinacobian/getaddons).

Insertions and deletions annotations.

Fasta files of insertions and deletions were annotated using PATRIC (https://www.patricbrc.org/), Conserved Domains Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and PHANNS (https://edwards.sdsu.edu/phanns). Genome maps were generated with EasyFig.

Technical considerations for CF sputum metagenomics.

CF sputum is a complex mixture of host derived cells, free DNA and mucus (Manzenreiter et al. 2012; Martínez-alemán, Campos-garcía, and Palma-nicolas 2017) in which viral, bacterial and fungi cells are entangled (DePas et al. 2016). Methods to get rid of host cells and DNA or RNA result in biases in the microbial community composition sampling. To preserve the abundances of each microbe in the system, a procedure with minimal sample manipulation was implemented.

Raw sputum or BAL samples were homogenized, and total DNA extracted without host removal or further manipulation. This procedure result in a high abundance of human genome reads (between 52% to 95% of total reads) in the metagenomes (Supplemental Figure 4.15). A sequencing depth of at least 5 million reads per sample is needed to detect the microbial community reads in this samples. Since sequencing costs are decreasing, we consider this strategy viable in the future of CF clinical metagenomics efforts.

In the CF microbial community, the detection of *Mycobacterium* and fungi is very important and cannot be ignored. To make sure that DNA from this hard to lyse microorganisms was extracted from sputum samples, a modified DNA extraction protocol was developed. It consists of three rounds of submerging the sputum sample in a cold bath (ethanol and dry ice) for 5 minutes and in a hot bath (water bath at 100 ºC) for 5 minutes, after which a bead-beating step was incorporated and DNA extraction performed with power-soil DNA kit. We showed that this procedure opens *Mycobacterium* and *Candida* cells, which are both concerns in CF clinical care. Metagenomic data analysis was designed for precise identification at strain level resolution. Recruitment at 100% identity allowed the identification of closest reference strain, in addition insertions and deletions were detected in CF patients metagenomes. This addresses the individual nature of CF disease progression, where even the same bacteria species have a distinctive mobilome. Genome plasticity can be identified with metagenomes and careful genomes analysis considering uncharacterized mobile elements.

Acknowledgments

Spruance Foundation. Jenna Mielke. UCSD sequencing core. Dinsdale lab for access to equipment.

References

Alexander, Bruce Marshall, Elbert Kristofer Petren, Samar Rizvi, Aliza Fink, Josh Ostrenga, Ase Sewall, and Deena Loeffler. 2016. “Cystic Fibrosis Foundation Patient Registry.” Annual Data Report. Bethesda, Maryland. https://www.cff.org/Our-Research/CF-Patient-Registry/2015-Patient-Registry-Annual-Data-Report.pdf.

Amoureux, Lucie, Julien Bador, Eliane Siebor, Nathalie Taillefumier, Annlyse Fanton, and Catherine Neuwirth. 2013. “Epidemiology and Resistance of Achromobacter Xylosoxidans from Cystic Fibrosis Patients in Dijon, Burgundy: First French Data.” Journal of Cystic Fibrosis 12 (2): 170–76. https://doi.org/10.1016/j.jcf.2012.08.005.

Bador, Julien, Lucie Amoureux, Emmanuel Blanc, and Catherine Neuwirth. 2013. “Innate Aminoglycoside Resistance of Achromobacter Xylosoxidans Is.” Antimicrobial Agents and Chemotherapy 57 (1): 603–5. https://doi.org/10.1128/AAC.01243-12.

Barsky, Emily E., Kathryn A. Williams, Gregory P. Priebe, and Gregory S. Sawicki. 2017. “Incident Stenotrophomonas Maltophilia Infection and Lung Function Decline in Cystic Fibrosis.” Pediatric Pulmonology 52 (10): 1276–82. https://doi.org/10.1002/ppul.23781.

Benz, Roland. 2015. Basic Mechanism of Pore-Forming Toxins. The Comprehensive Sourcebook of Bacterial Protein Toxins. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-800188-2.00021-5.

Budzik, Jonathan M, William a Rosche, Arne Rietsch, and George a O Toole. 2004. “Isolation and Characterization of a Generalized Transducing Phage for Pseudomonas Aeruginosa Strains PAO1 and PA14 Isolation and Characterization of a Generalized Transducing Phage for Pseudomonas Aeruginosa Strains PAO1 and PA14.” Society 186 (10): 3270–73. https://doi.org/10.1128/JB.186.10.3270.

Busby, Ben, David M. Kristensen, and Eugene V. Koonin. 2013. “Contribution of Phage-Derived Genomic Islands to the Virulence of Facultative Bacterial Pathogens.” Environmental Microbiology 15 (2): 307–12. https://doi.org/10.1111/j.1462-2920.2012.02886.x.

Cobián-Güemes, Ana Georgina, Wei Lim, Robert A Quinn, Douglas J Conrad, Sean Benler, Heather Maughan, Rob Edwards, et al. 2019. “Cystic Fibrosis Rapid Response : Translating Multi-Omics Data into Clinically Relevant Information.” MBio 10 (2): 1–15.

DePas, William H., Ruth Starwalt-Lee, Lindsey Van Sambeek, Sripriya Ravindra Kumar, Viviana Gradinaru, and Dianne K. Newman. 2016. “Exposing the Three-

Dimensional Biogeography and Metabolic States of Pathogens in Cystic Fibrosis Sputum via Hydrogel Embedding, Clearing, and RRNA Labeling." MBio 7 (5): 1–11. https://doi.org/10.1128/mBio.00796-16.

Dobrindt, Ulrich, Sarah Tjaden, Sadrick Shah, and Jörg Hacker. 2015. Mobile Genetic Elements and Pathogenicity Islands Encoding Bacterial Toxins. The Comprehensive Sourcebook of Bacterial Protein Toxins. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-800188-2.00002-1.

Fasano, A., B. Baudry, D. W. Pumplin, S. S. Wasserman, B. D. Tall, J. M. Ketley, and J. B. Kaper. 1991. "Vibrio Cholerae Produces a Second Enterotoxin, Which Affects Intestinal Tight Junctions." Proceedings of the National Academy of Sciences 88 (12): 5242–46. https://doi.org/10.1073/pnas.88.12.5242.

Feigelman, Rounak, Christian R Kahlert, Florent Baty, Frank Rassouli, Rebekka L Kleiner, Philipp Kohler, Martin H Brutsche, and Christian von Mering. 2017. "Sputum DNA Sequencing in Cystic Fibrosis: Non-Invasive Access to the Lung Microbiome and to Pathogen Details." Microbiome 5 (1): 20. https://doi.org/10.1186/s40168-017-0234-1.

Hagemann, Martin, Dirk Hasse, and Gabriele Berg. 2006. "Detection of a Phage Genome Carrying a Zonula Occludens like Toxin Gene (Zot) in Clinical Isolates of Stenotrophomonas Maltophilia." Archives of Microbiology 185 (6): 449–58. https://doi.org/10.1007/s00203-006-0115-7.

Jakobsen, Tim Holm, Martin Asser Hansen, Peter Østrup Jensen, Lars Hansen, Leise Riber, Mette Kolpen, Christine Rønne Hansen, et al. 2013. "Complete Genome Sequence of the Cystic Fibrosis Pathogen Achromobacter Xylosoxidans NH44784-1996 Complies with Important Pathogenic Phenotypes" 8 (7): 8–11. https://doi.org/10.1371/journal.pone.0068484.

James, Chloe E., Emily V. Davies, Joanne L. Fothergill, Martin J. Walshaw, Colin M. Beale, Michael A. Brockhurst, and Craig Winstanley. 2015. "Lytic Activity by Temperate Phages of Pseudomonas Aeruginosa in Long-Term Cystic Fibrosis Chronic Lung Infections." ISME Journal 9 (6): 1391–98. https://doi.org/10.1038/ismej.2014.223.

Jeukens, Julie, Luca Freschi, Antony T. Vincent, Jean Guillaume Emond-Rheault, Irena Kukavica-Ibrulj, Steve J. Charette, and Roger C. Levesque. 2017. "A Pan-Genomic Approach to Understand the Basis of Host Adaptation in Achromobacter." Genome Biology and Evolution 9 (4): 1030–46. https://doi.org/10.1093/gbe/evx061.

Julio, Steven M., Douglas M. Heithoff, and Michael J. Mahan. 2000. "SsrA (TmRNA) Plays a Role in Salmonella Enterica Serovar Typhimurium Pathogenesis." Journal

of Bacteriology 182 (6): 1558–63. https://doi.org/10.1128/JB.182.6.1558-1563.2000.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu:Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation." Genome Research 27:1-15: 1–11. https://doi.org/10.1101/gr.215087.116.Freely.

Krüger, Alejandra, and Paula M.A. Lucchesi. 2015. "Shiga Toxins and Stx Phages: Highly Diverse Entities." Microbiology (United Kingdom) 161 (3): 1–12. https://doi.org/10.1099/mic.0.000003.

Laguna, Theresa A., Brandie D. Wagner, Cynthia B. Williams, Mark J. Stevens, Charles E. Robertson, Cole W. Welchlin, Catherine E. Moen, Edith T. Zemanick, and Jonathan K. Harris. 2016. "Airway Microbiota in Bronchoalveolar Lavage Fluid from Clinically Well Infants with Cystic Fibrosis." PLoS ONE 11 (12): 1–15. https://doi.org/10.1371/journal.pone.0167649.

Lim, Yan Wei, Jose S. Evangelista, Robert Schmieder, Barbara Bailey, Matthew Haynes, Mike Furlan, Heather Maughan, Robert Edwards, Forest Rohwer, and Douglas Conrad. 2014. "Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis." Journal of Clinical Microbiology 52 (2): 425–37. https://doi.org/10.1128/JCM.02204-13.

Lim, Yan Wei, Robert Schmieder, Matthew Haynes, Dana Willner, Mike Furlan, Merry Youle, Katelynn Abbott, et al. 2013. "Metagenomics and Metatranscriptomics: Windows on CF-Associated Viral and Microbial Communities." Journal of Cystic Fibrosis 12 (2): 154–64. https://doi.org/10.1016/j.jcf.2012.07.009.

LiPuma, John J. 2010. "The Changing Microbial Epidemiology in Cystic Fibrosis." Clinical Microbiology Reviews 23 (2): 299–323. https://doi.org/10.1128/CMR.00068-09.

Losada, P.M., P. Chouvarine, A. Schulz, S. Hedtfeld, S. Mielke, M. Dorda, L. Wiehlmann, and B. Tümmler. 2016. "The Cystic Fibrosis Lower Airways Microbial Metagenome." ERJ Open Research 2 (2): 00096–02015. https://doi.org/10.1183/23120541.00096-2015.

Manzenreiter, Reinhard, Ferry Kienberger, Veronica Marcos, Kurt Schilcher, Wolf D. Krautgartner, Astrid Obermayer, Marlene Huml, et al. 2012. "Ultrastructural Characterization of Cystic Fibrosis Sputum Using Atomic Force and Scanning Electron Microscopy." Journal of Cystic Fibrosis 11 (2): 84–92. https://doi.org/10.1016/j.jcf.2011.09.008.

Martínez-alemán, Saira R, Lizbeth Campos-garcía, and José P Palma-nicolas. 2017. "Understanding the Entanglement : Neutrophil Extracellular Traps (NETs) in Cystic Fibrosis" 7 (April): 1–7. https://doi.org/10.3389/fcimb.2017.00104.

Moore, Sean D., and Robert T. Sauer. 2007. "The TmRNA System for Translational Surveillance and Ribosome Rescue." Annual Review of Biochemistry 76 (1): 101–24. https://doi.org/10.1146/annurev.biochem.75.103004.142733.

Moran Losada, Patricia, Philippe Chouvarine, Marie Dorda, Silke Hedtfeld, Samira Mielke, Angela Schulz, Lutz Wiehlmann, and Burkhard Tümmler. 2016. "The Cystic Fibrosis Lower Airways Microbial Metagenome." ERJ Open Research 2 (2): 00096–02015. https://doi.org/10.1183/23120541.00096-2015.

Overbeek, Ross, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, et al. 2014. "The SEED and the Rapid Annotation of Microbial Genomes Using Subsystems Technology (RAST)." Nucleic Acids Research 42 (D1): 206–14. https://doi.org/10.1093/nar/gkt1226.

Pelludat, Cosima, Susanne Mirold, and Wolf Dietrich Hardt. 2003. "The SopEφ Phage Integrates into the SsrA Gene of Salmonella Enterica Serovar Typhimurium A36 and Is Closely Related to the Fels-2 Prophage." Journal of Bacteriology 185 (17): 5182–91. https://doi.org/10.1128/JB.185.17.5182-5191.2003.

Pierro, Mariarosaria Di, Ruliang Lu, Sergio Uzzau, Wenle Wang, Klara Margaretten, Carlo Pazzani, Francesco Maimone, and Alessio Fasano. 2001. "Zonula Occludens Toxin Structure-Function Analysis: Identification of the Fragment Biologically Active on Tight Junctions and of the Zonulin Receptor Binding Domain." Journal of Biological Chemistry 276 (22): 19160–65. https://doi.org/10.1074/jbc.M009674200.

Quinn, Robert A, Katrine Whiteson, Yan-wei Lim, Peter Salamon, Barbara Bailey, Simone Mienardi, Savannah E Sanchez, Don Blake, Doug Conrad, and Forest Rohwer. 2014. "A Winogradsky-Based Culture System Shows an Association between Microbial Fermentation and Cystic Fibrosis Exacerbation" 9 (4): 1024–38. https://doi.org/10.1038/ismej.2014.234.

Reyes, Alejandro, Laura V. Blanton, Song Cao, Guoyan Zhao, Mark Manary, Indi Trehan, Michelle I. Smith, et al. 2015. "Gut DNA Viromes of Malawian Twins Discordant for Severe Acute Malnutrition." Proceedings of the National Academy of Sciences 112 (38): 11941–46. https://doi.org/10.1073/pnas.1514285112.

Ridderberg, Winnie, Signe Maria Nielsen, and Niels Nørskov-Lauritsen. 2015. "Genetic Adaptation of Achromobacter Sp. during Persistence in the Lungs of Cystic Fibrosis Patients." PLoS ONE 10 (8): 1–14. https://doi.org/10.1371/journal.pone.0136790.

Rodrigues, Elenice Ra, Géssica A. Rocha, Alex G. Ferreira, Robson S. Leão, Rodolpho M. Albano, and Elizabeth A. Marques. 2016. “Draft Genome Sequences of Four Achromobacter Ruhlandii Strains Isolated from Cystic Fibrosis Patients.” Memorias Do Instituto Oswaldo Cruz 111 (12): 777–80. https://doi.org/10.1590/0074-02760160130.

Ruan, Ye Chun, Yan Wang, Nicolas da Silva, Bongki Kim, Rui Ying Diao, Eric Hill, Dennis Brown, Hsiao Chang Chan, and Sylvie Breton. 2014. “CFTR Interacts with ZO-1 to Regulate Tight Junction Assembly and Epithelial Differentiation through the ZONAB Pathway.” Journal of Cell Science 127 (20): 4396–4408. https://doi.org/10.1242/jcs.148098.

Rudkjøbing, Vibeke B., Trine R. Thomsen, Morten Alhede, Kasper N. Kragh, Per H. Nielsen, Ulla R. Johansen, Michael Givskov, Niels Høiby, and Thomas Bjarnsholt. 2012. “The Microorganisms in Chronically Infected End-Stage and Non-End-Stage Cystic Fibrosis Patients.” FEMS Immunology and Medical Microbiology 65 (2): 236–44. https://doi.org/10.1111/j.1574-695X.2011.00925.x.

Silveira, Cynthia B, and Forest Rohwer. 2016. “Piggyback-the-Winner in Host-Associated Microbial Communities.” Npj Biofilms and Microbiomes 2 (2): 10–13. https://doi.org/10.1038/npjbio.

Surette, Michael G. 2014. “The Cystic Fibrosis Lung Microbiome.” Annals of the American Thoracic Society 11 (SUPPL. 1): 61–65. https://doi.org/10.1513/AnnalsATS.201306-159MG.

Swenson, Colin E, and Ruxana T Sadikot. 2015. “Achromobacter Respiratory Infections.” Annals of the American Thoracic Society 12 (2): 252–58. https://doi.org/10.1513/AnnalsATS.201406-288FR.

Uzzau, Sergio, Ruliang Lu, Wenle Wang, Cara Fiore, and Alessio Fasano. 2001. “Purification and Preliminary Characterization of the Zonula Occludens Toxin Receptor from Human (CaCo2) and Murine (IEC6) Intestinal Cell Lines.” FEMS Microbiology Letters 194 (1): 1–5. https://doi.org/10.1016/S0378-1097(00)00478-X.

Wang, Xiaoxue, Younghoon Kim, Qun Ma, Seok Hoon Hong, Karina Pokusaeva, Joseph M. Sturino, and Thomas K. Wood. 2010. “Cryptic Prophages Help Bacteria Cope with Adverse Environments.” Nature Communications 1 (9). https://doi.org/10.1038/ncomms1146.

Waters, Valerie, Eshetu G Atenafu, Annie Lu, Yvonne Yau, Elizabeth Tullis, and Felix Ratjen. 2013. “Chronic Stenotrophomonas Maltophilia Infection and Mortality or Lung Transplantation in Cystic Fi Brosis Patients.” Journal of Cystic Fibrosis 12 (5): 482–86. https://doi.org/10.1016/j.jcf.2012.12.006.

Whelan, Fiona J., Alya A. Heirali, Laura Rossi, Harvey R. Rabin, Michael D. Parkins, and Michael G. Surette. 2017. “Longitudinal Sampling of the Lung Microbiota in Individuals with Cystic Fibrosis.” PLoS ONE 12 (3): 1–17. https://doi.org/10.1371/journal.pone.0172811.

Whiteson, Katrine L., Simone Meinardi, Yan Wei Lim, Robert Schmieder, Heather Maughan, Robert Quinn, Donald R. Blake, Douglas Conrad, and Forest Rohwer. 2014. “Breath Gas Metabolites and Bacterial Metagenomes from Cystic Fibrosis Airways Indicate Active PH Neutral 2,3-Butanedione Fermentation.” ISME Journal 8 (6): 1247–58. https://doi.org/10.1038/ismej.2013.229.

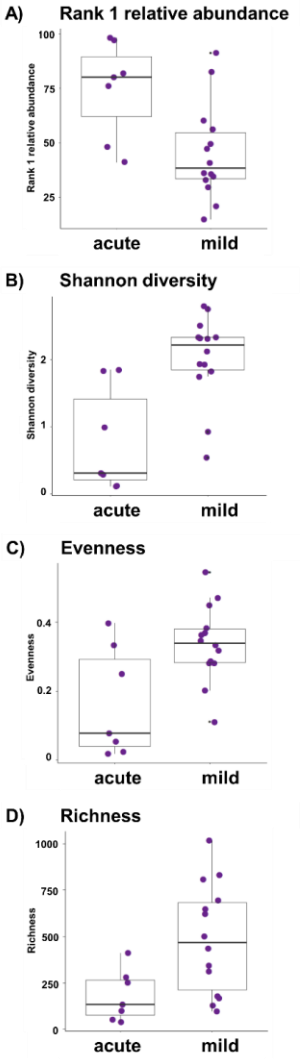
Willner, Dana, Matthew R. Haynes, Mike Furlan, Nicole Hanson, Breeann Kirby, Yan Wei Lim, Paul B. Rainey, et al. 2012. “Case Studies of the Spatial Heterogeneity of DNA Viruses in the Cystic Fibrosis Lung.” American Journal of Respiratory Cell and Molecular Biology 46 (2): 127–31. https://doi.org/10.1165/rcmb.2011-0253OC.

Wilson, Brenda A., and Mengfei Ho. 2006. Evolutionary Aspects of Toxin-Producing Bacteria. The Comprehensive Sourcebook of Bacterial Protein Toxins. Elsevier Ltd. https://doi.org/10.1016/B978-012088445-2/50007-X.

Zhao, Jiangchao, Patrick D Schloss, Linda M Kalikin, Lisa A Carmody, Bridget K Foster, Joseph F Petrosino, James D Cavalcoli, et al. 2012. “Decade-Long Bacterial Community Dynamics in Cystic Fibrosis Airways.” Proceedings of the National Academy of Sciences of the United States of America 109 (15): 5809–14. https://doi.org/10.1073/pnas.1120577109.

Appendix for Chapter 4

Supplemental figures



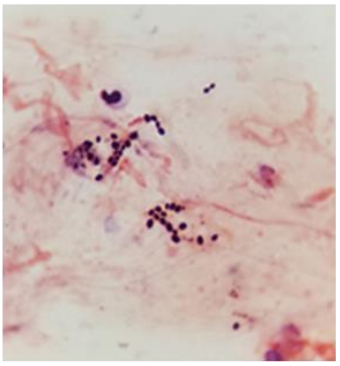
Supplemental Figure 4.1 Bacteria diversity in acute and mild CF exacerbations. Bacteria relative abundance was obtained from sputum metagenomes from acute exacerbations (this study) and mild exacerbations (Losada, 2016). A) Relative abundance of most abundant member of the microbial community. B) Shannon diversity (H). C) Evenness calculated as H/In(S). D) Richness (S) calculated as the total number of bacteria species. An unpaired two-samples Wilcoxon test between the acute and mild groups was performed for all measurements, in all cases the two populations are significantly different with a p-value < 0.01. Wilcoxon test results are: rank 1 relative abundance w=81, p-value 0.015, Shannon diversity w=10, p-value=0.002, richness w=17, p-value=0.015, evenness w=20, p-value 0.030.

B

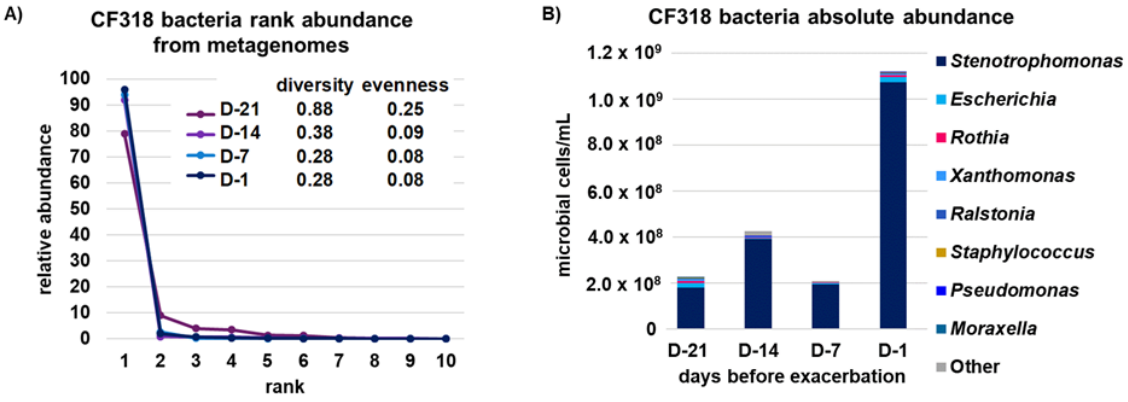
***Candida glabrata* genome coverage**

Chromosome Length	Reads	Mismatch %	Coverage
491,328	1,137	1.43	0.35
502,101	1,164	0.91	0.35
558,804	1,404	1.38	0.38
651,701	1,375	1.02	0.32
687,738	1,502	1.08	0.33
927,101	2,064	1.06	0.33
992,211	2,897	3.65	0.44
1,050,361	2,312	0.91	0.33
1,100,349	2,561	1.09	0.35
1,195,132	2,640	0.93	0.33
1,302,831	3,033	0.92	0.35
1,455,689	3,864	1.01	0.40
1,402,899	3,161	0.88	0.34

C **Sputum gram stain**



Supplemental Figure 4.2 CF292 exacerbation sample metagenome and gram-stain. B) *Candida glabrata* genome coverage. C) Gram-stain of sputum sample, *Candida glabrata* cells stained gram positive.



C)

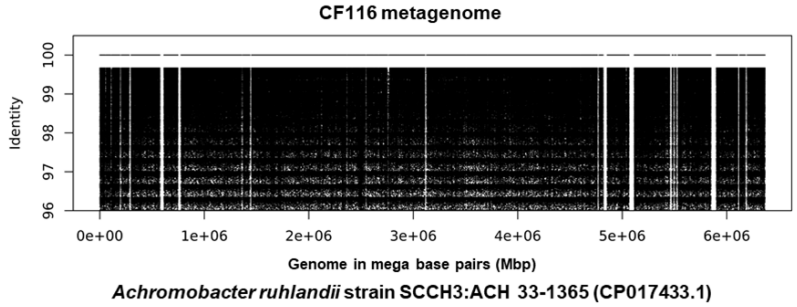
Sample	Tissue damage	Other	Iron acquisition	Alginate	Capsule	Motility
D-21	0.00	28.54	0.00	0.00	0.00	0.00
D-14	0.00	0.00	0.00	0.00	0.00	66.95
D-7	0.00	0.00	0.00	0.00	0.00	0.00
D-1	7.65	66.41	5.84	31.20	2.94	232.12

D)

Sample collection day	viral particles/mL	bacteria cells/mL	virus to microbe ratio
stable D-21	2.5×10^9	2.3×10^8	11.11
stable D-14	6.9×10^8	4.2×10^8	1.63
stable D-7	4.1×10^9	2.1×10^8	19.81
exacerbation D-1	2.4×10^{10}	1.1×10^9	21.07

Supplemental Figure 4.3 CF318 historical sampling. Weekly monitoring of hyper-variable CF patient CF318 for one month which includes three stable samples (D-21, D-14 and D-7) and one exacerbation sample (D-1). A) Microbial community structure from bacteria relative abundances obtained from metagenomes. B) Bacteria genus absolute abundances inferred from metagenomes relative abundances combined with microbial cells/mL counts. C) Virulence factors fractional abundances obtained from sputum metagenomes. D) Viral particles and bacteria cells quantification by epifluorescence microscopy.

A)

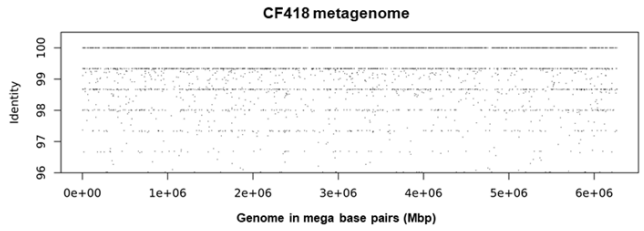


B)

A. ruhlandii (CP017433.1) excised region	length (bp)	number of cds	phage cds	mobile element cds	Integrase cds	transposase cds	hypothetical cds	start position	end position
excised region 1	10,807	11	0	0	0	0	8	193,417	204,224
excised region 2	15,374	24	0	0	0	1	21	282,608	297,982
excised region 3 : type III secretion related	36,815	44	0	0	0	0	24	576,443	613,258
excised region 4 : phage	**28,228**	**57**	**3**	**0**	**2**	**0**	**46**	**748,053**	**776,281**
excised region 5 : phage	**12,691**	**16**	**8**	**0**	**1**	**0**	**7**	**1,437,930**	**1,450,621**
excised region 6	13,719	3	0	0	0	0	3	3,111,991	3,125,710
excised region 7 : ABC transporter related	12,507	14	0	0	0	0	1	4,763,603	4,776,110
excised region 8 : phage/mobile element	**37,305**	**50**	**1**	**2**	**1**	**0**	**32**	**4,818,072**	**4,855,377**
excised region 9 : mobile element	**49,431**	**68**	**0**	**2**	**0**	**1**	**46**	**5,064,875**	**5,114,306**
excised region 10 : mobile element	**19,166**	**23**	**1**	**2**	**1**	**0**	**16**	**5,455,867**	**5,475,033**
excised region 11 : phage	**12,677**	**20**	**4**	**0**	**0**	**0**	**15**	**5,489,249**	**5,501,926**
excised region 12	13,681	22	0	0	0	0	22	5,517,693	5,531,374
excised region 13 : type IV secretion related	50,664	65	1	0	1	0	44	5,849,272	5,899,936
excised region 14 : carbohydrates metabolism	14,215	18	0	0	0	0	9	6,179,651	6,193,866

Supplemental Figure 4.4 Excised regions in *Achromobacter ruhlandii* from CF116 exacerbation metagenome.

A)



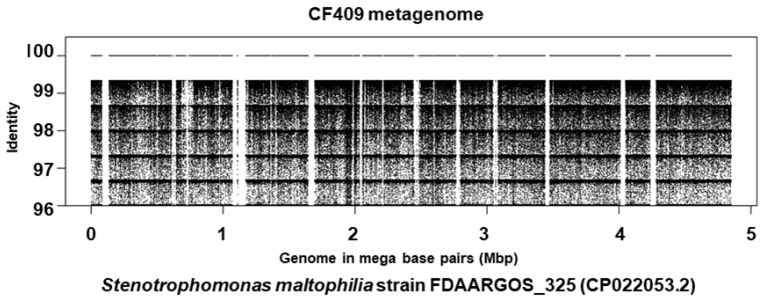
Achromobacter xylosoxidans strain FDAARGOS_150 (CP014028.1)

B)

A. xylosoxidans strain FDAARGOS_150 (CP014028.1) excised region	length (bp)	number of cds	phage cds	mobile element cds	Integrase cds	transposase cds	hypothetical cds	start position	end position
excised region 1	11,648	56	0	0	0	0	43	39,996	51,644
excised region 2	25,063	29	0	0	0	0	19	214,473	239,536
excised region 3	10,454	12	0	0	0	0	8	791,132	801,586
excised region 4	15,010	17	0	0	0	0	11	905,937	920,947
excised region 5 : CRISPR-Cas	14,450	57	0	0	0	0	9	1,202,120	1,216,570
excised region 6	10,341	16	0	0	0	0	11	1,383,998	1,394,339
excised region 7 : ABC transporter related	10,798	16	0	0	0	0	9	1,401,085	1,411,883
excised region 8 : benzoate metabolism related	15,792	17	0	0	0	0	9	1,498,140	1,513,932
excised region 9	10,191	12	0	0	0	0	7	1,838,415	1,848,606
excised region 10 : mobile element	**23,981**	**42**	**0**	**6**	**0**	**1**	**27**	**1,913,022**	**1,937,003**
excised region 11	10,000	18	0	0	0	0	7	2,221,007	2,231,007
excised region 12	19,271	21	0	0	0	0	15	2,565,989	2,585,260
excised region 13	37,225	39	0	0	0	0	19	2,642,134	2,679,359
excised region 14	23,922	24	0	0	0	0	21	2,916,796	2,940,718
excised region 15	11,134	12	0	0	0	0	6	2,948,589	2,959,723
excised region 16	10,005	11	0	0	0	0	5	3,173,391	3,183,396
excised region 17	14,435	19	0	0	0	0	12	4,036,783	4,051,218
excised region 18 : phage/mobile element	**11,200**	**18**	**1**	**1**	**0**	**0**	**15**	**4,743,125**	**4,754,325**
excised region 19 : phage	**16,630**	**24**	**2**	**0**	**0**	**0**	**17**	**4,754,476**	**4,771,106**
excised region 20 : phage	**19,675**	**17**	**2**	**0**	**0**	**0**	**15**	**4,778,461**	**4,798,136**
excised region 21	11,704	27	0	0	0	0	19	4,893,784	4,905,488
excised region 22	12,169	19	0	0	0	0	16	4,989,609	5,001,778
excised region 23 : mobile element	**17,229**	**31**	**0**	**1**	**0**	**0**	**21**	**5,916,539**	**5,933,768**
excised region 24 : mobile element	**14,637**	**24**	**0**	**2**	**0**	**0**	**20**	**6,230,042**	**6,244,679**

Supplemental Figure 4.5 Excised regions in *Achromobacter xylosoxidans* from CF418 exacerbation metagenome.

A)

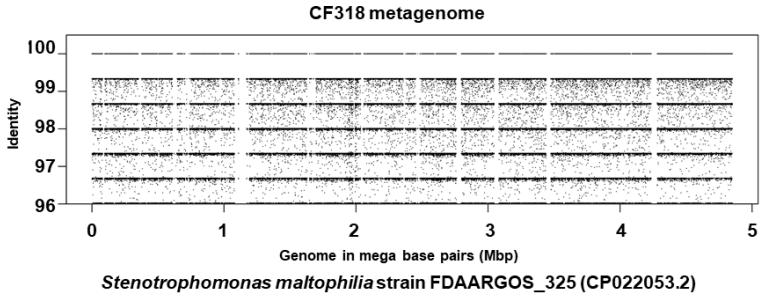


B)

S. maltophilia strain FDAARGOS_325 (CP022053.2) excised regions	length (bp)	number of cds	phage cds	mobile element cds	Integrase cds	transposase cds	hypothetical cds	start position	end position
excised region 1 : mobile element	**52,776**	**147**	**1**	**4**	**1**	**0**	**89**	**82,701**	**135,477**
excised region 2 : fatty acids metabolism	11,726	12	0	0	0	0	9	610,059	621,785
excised region 3 : phage	**37,491**	**58**	**12**	**0**	**0**	**0**	**41**	**1,070,729**	**1,108,220**
escised region 4 : phage	**58,955**	**98**	**7**	**0**	**1**	**0**	**88**	**1,111,375**	**1,170,330**
excised region 5 : phage/mobile element	**48,904**	**77**	**1**	**1**	**2**	**0**	**69**	**1,645,445**	**1,694,349**
excised region 6 : mobile element, efflux RND transporter	16,853	20	0	2	1	0	7	2,034,493	2,051,346
escised region 7	12,098	7	0	0	0	0	3	2,204,841	2,216,939
excised region 8 : carbohydrates metabolism	11,123	12	0	0	0	0	3	2,455,560	2,466,683
excised region 9 : carbohydrates metabolism	11,654	12	0	0	0	0	7	2,470,368	2,482,022
escised region 10 : mobile element	**16,358**	**21**	**0**	**2**	**0**	**0**	**14**	**2,765,717**	**2,782,075**
excised region 11	13,913	25	0	0	0	0	15	3,047,492	3,061,405
excised region 12 : mobile element	**33,821**	**64**	**0**	**3**	**0**	**1**	**57**	**3,441,184**	**3,475,005**
escised region 13 : phage	**39,271**	**62**	**11**	**0**	**0**	**0**	**47**	**4,010,367**	**4,049,638**
excised region 14 : phage	**45,063**	**67**	**8**	**1**	**0**	**0**	**56**	**4,236,173**	**4,281,236**

Supplemental Figure 4.6 Excised regions in *Stenotrophomonas maltophilia* from CF409 exacerbation metagenome.

A)

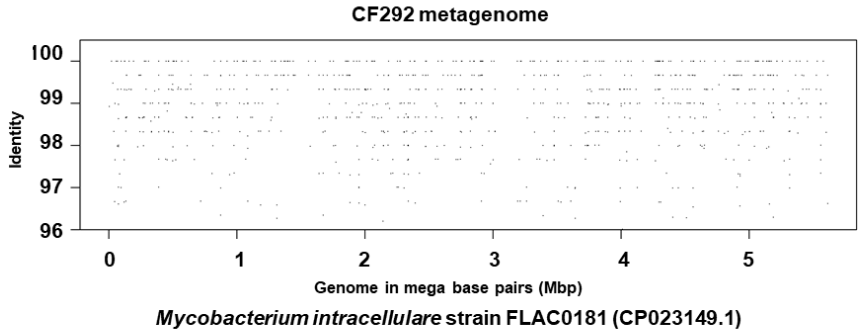


B)

S. maltophilia strain FDAARGOS_325 (CP022053.2) excised regions	length (bp)	number of cds	phage cds	mobile element cds	Integrase cds	transposase cds	hypothetical cds	start position	end position
excised region 1 : mobile element	**15,287**	**31**	**1**	**2**	**1**	**0**	**24**	**81,296**	**96,583**
excised region 2 : fatty acids metabolism	11,766	12	0	0	0	0	9	610,051	621,817
excised region 3 : mobile element	**11,340**	**21**	**1**	**2**	**0**	**0**	**14**	**622,551**	**633,891**
excised region 4 : phage	**28,642**	**48**	**9**	**0**	**0**	**0**	**35**	**1,079,679**	**1,108,321**
excised region 5 : phage	**58,967**	**98**	**7**	**0**	**1**	**0**	**88**	**1,111,363**	**1,170,330**
excised region 6 : phage	**10,936**	**18**	**1**	**0**	**1**	**0**	**17**	**1,180,547**	**1,191,483**
excised region 7 : efflux transport system	14,979	23	0	0	0	0	15	1,631,053	1,646,032
excised region 8 : mobile element	**10,925**	**19**	**1**	**1**	**1**	**0**	**17**	**1,666,151**	**1,677,076**
excised region 9 : efflux RND transporters	13,960	17	0	1	0	0	7	2,037,020	2,050,980
excised region 10 : carbohydrates metabolism	10,410	12	0	0	0	0	7	2,359,650	2,370,060
excised region 11 : carbohydrates metabolism	12,277	13	0	0	0	0	7	2,469,683	2,481,960
excised region 12 : mobile element	**33,187**	**50**	**0**	**2**	**1**	**1**	**31**	**2,763,658**	**2,796,845**
excised region 13 : mobile element	**35,849**	**55**	**0**	**1**	**0**	**0**	**38**	**3,046,679**	**3,082,528**
excised region 14 : mobile element	**33,843**	**64**	**0**	**3**	**0**	**1**	**57**	**3,441,169**	**3,475,012**
excised region 15 : phage	**45,122**	**67**	**8**	**1**	**0**	**0**	**55**	**4,236,173**	**4,281,295**

Supplemental Figure 4.7 Excised regions in *Stenotrophomonas maltophilia* from CF318 exacerbation metagenome.

A)

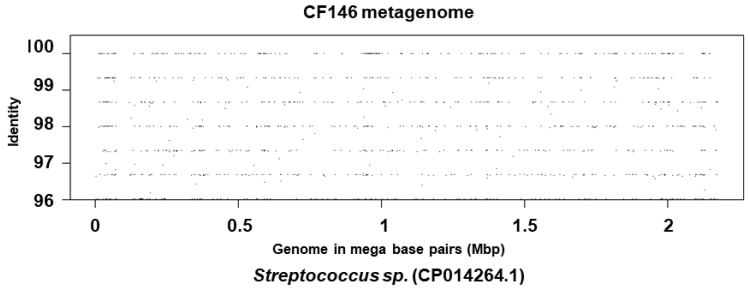


B)

M. intracellulare (CP023149.1) excised regions	length (bp)	number of cds	phage cds	mobile element cds	Integrase cds	transposase cds	hypothetical cds	start position	end position
excised region 1	42,370	59	0	0	0	0	27	390,232	432,602
excised region 2 : phage/mobile element	**49,208**	**86**	**1**	**2**	**2**	**0**	**55**	**668,111**	**717,319**
excised region 3	41,583	58	0	1	0	0	28	874,365	915,948
excised region 4	104,304	144	0	2	0	0	79	1,455,150	1,559,454
excised region 5	40,467	53	0	0	0	0	31	2,779,488	2,819,955
excised region 6 : mobile element	**173,810**	**266**	**1**	**6**	**1**	**2**	**176**	**3,016,024**	**3,189,834**
excised region 7	55,588	73	0	0	0	0	44	3,421,448	3,477,036
excised region 8 : mobile element	**61,463**	**106**	**0**	**4**	**0**	**0**	**74**	**3,648,475**	**3,709,938**
excised region 9	53,555	60	0	0	0	0	25	3,907,436	3,960,991
excised region 10	96,508	147	0	1	1	0	88	4,171,351	4,267,859
excised region 11	41,901	49	0	0	0	0	21	4,576,662	4,618,563
excised region 12	41,192	59	0	0	0	0	33	5,218,259	5,259,451
excised region 13	57,957	67	0	0	0	0	37	5,466,651	5,524,608

Supplemental Figure 4.8 Excised regions in *Mycobacterium intracellulare* from CF292 exacerbation metagenome.

A)



***Streptococcus sp.* (CP014264.1)**

B)

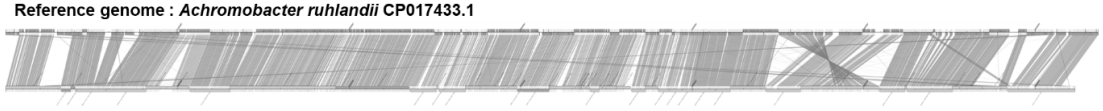
***Streptococcus* sp. (CP014264.1) excised regions**	**length (bp)**	**number of cds**	**phage cds**	**mobile element cds**	**Integrase cds**	**transposase cds**	**hypothetical cds**	**start position**	**end position**
excised region 1 : phage	**41,831**	**70**	**29**	**0**	**1**	**0**	**39**	**73,710**	**115,541**
excised region 2 : nucleic acids metabolism	12,477	13	0	0	0	0	4	138,932	151,409
excised region 3 : carbohydrates metabolism	15,044	20	0	0	0	0	8	393,619	408,663
excised region 4 : RNA metabolism	10,299	21	0	0	0	0	10	411,163	421,462
excised region 5 : carbohydrates metabolism	10,863	12	0	0	0	0	4	714,224	725,087
excised region 6 : restriction modification systems	22,255	22	0	0	0	0	11	726,924	749,179
excised region 7 : B12 transporters, ABC transporters	17,022	27	0	1	0	0	13	776,457	793,479
excised region 8 : nucleoic acids metabolism, hemolysin III	12,408	20	0	0	0	0	5	841,982	854,390
excised region 9 : carbohydrates metabolism	21,639	2	0	0	0	0	0	1,077,668	1,099,307
excised region 10 : ABC transporters	10,036	17	1	0	0	0	8	1,193,255	1,203,291
excised region 11	10,270	13	0	0	0	0	7	1,207,985	1,218,255
excised region 12	10,382	14	0	0	0	0	6	1,316,323	1,326,705
excised region 13	13,157	23	0	0	0	0	20	1,423,851	1,437,008
excised region 14 : ABC transporters	14,941	27	0	0	0	0	12	1,467,779	1,482,720
excised region 15	14,544	14	0	0	0	0	6	1,510,645	1,525,189
excised region 16	22,433	25	0	0	0	0	21	1,752,304	1,774,737
excised region 17	13,316	4	0	0	0	0	3	1,823,203	1,836,519
excised region 18 : carbohydrates metabolism	12,651	7	0	0	0	0	4	2,020,919	2,033,570
excised region 19 : phage	**13,712**	**32**	**6**	**0**	**2**	**0**	**19**	**2,062,563**	**2,076,275**

Supplemental Figure 4.9 Excised regions in *Streptococcus sp.* from CF146 exacerbation metagenome.

A)

CF116 metagenome, *Achromobacter* contigs with insertions

Reference genome : *Achromobacter ruhlandii* CP017433.1



Assembled contigs : *Achromobacter ruhlandii* CF116

B)

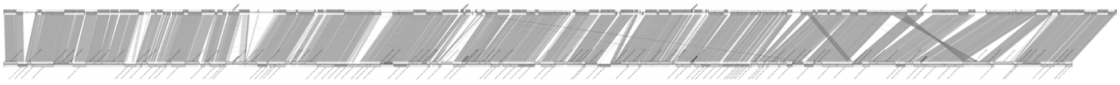
Inserted regions in *Achromobacter ruhlandii* CF116

Achromobacter ruhlandii CF116 contig insert regions	length (bp)	number of cds	phage cds	mobile element cds	integrase cds	transposase cds	hypothetical cds	start position in contig	end position in contig	strain in contig
Insert region 1 : TCA cycle	14,593	18	0	0	0	0	6	17,066	31,659	+
Insert region 2 : phage	**44,907**	**73**	**21**	**0**	**2**	**0**	**51**	**75,058**	**119,965**	**+**
Insert region 3 : phage	**31,879**	**48**	**6**	**0**	**0**	**0**	**39**	**1**	**31,880**	**+**
Insert region 4	19,145	30	0	0	0	0	23	89,857	109,002	+
Insert region 5	10,353	2	0	0	0	0	2	102,155	112,508	+
Insert region 6 : ABC transporters	33,485	37	0	0	0	0	15	38,890	72,375	+
Insert region 7 : phage	**40,353**	**68**	**21**	**0**	**0**	**0**	**45**	**41,967**	**82,320**	**+**
Insert region 8 : phage	**13,377**	**29**	**4**	**0**	**0**	**0**	**24**	**10,796**	**24,173**	**+**
Insert region 9 : phage	**20,257**	**15**	**1**	**0**	**2**	**0**	**10**	**139,426**	**159,683**	**+**
Insert region 10 : LPS export, ABC transporters	13,620	17	0	0	0	0	3	16,155	29,775	+
Insert region 11 : ABC transporters	21,545	34	0	0	0	0	18	93,287	114,832	+
Insert region 12 : carbohydrates metabolism	19,418	23	0	0	0	0	15	247,035	266,453	+

Supplemental Figure 4.10 Insertions in contigs from CF116 metagenome A) Synteny between the closest reference genome and contigs assembled from CF116 metagenome. B) Addon regions identified in assembled contigs.

A) **CF409 metagenome, *Stenotrophomonas* contigs with insertions**

Reference genome : *Stenotrophomonas*



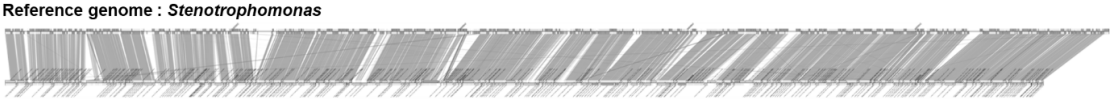
Assembled contigs : *Stenotrophomonas maltophilia* CF409

B) **Inserted regions in *Stenotrophomonas maltophilia* CF409**

Stenotrophomonas maltophilia CF409 contig insert regions	length (bp)	number of cds	phage cds	mobile element cds	integrase cds	transposase cds	hypothetical cds	start position in contig	end position in contig	strain in contig
Insert region 1 : beta-lactamases	10,662	11	0	0	0	0	4	1	10,663	+
Insert region 2 : efflux pumps	10,000	9	0	0	0	0	1	19,677	29,677	+
Insert region 3 : conjugative elements, efflux pumps	42,834	62	0	0	1	0	31	13,928	56,762	+
Insert region 4 : DNA metabolism	12,027	20	0	0	1	0	13	13,830	25,857	+
Insert region 5	20,401	26	0	0	0	0	25	1	20,402	+
Insert region 6	11,539	13	0	0	0	0	6	25,799	37,338	+
Insert region 7 : phage	**23,375**	**39**	**6**	**0**	**0**	**0**	**32**	**6,834**	**30,209**	**+**
Insert region 8 : ribosomal proteins	16,332	20	0	0	0	0	8	16,691	33,023	+
Insert region 9 : carbohydrates metabolism	16,217	14	0	0	0	0	6	1	16,218	+
Insert region 10	10,877	17	0	0	0	0	14	43,058	53,935	+
Insert region 11	39,442	52	0	0	0	0	35	1	39,443	+
Insert region 12	10,670	11	0	0	0	0	7	48,158	58,828	+
Insert region 13 : beta-lactamases	25,350	28	0	0	0	0	16	9,614	34,964	+
Insert region 14 : phage	**61,434**	**71**	**2**	**0**	**0**	**0**	**63**	**5,127**	**66,561**	**+**

Supplemental Figure 4.11 Insertions in contigs from CF409 metagenome A) Synteny between the closest reference genome and contigs assembled from CF409 metagenome. B) Addon regions identified in assembled contigs.

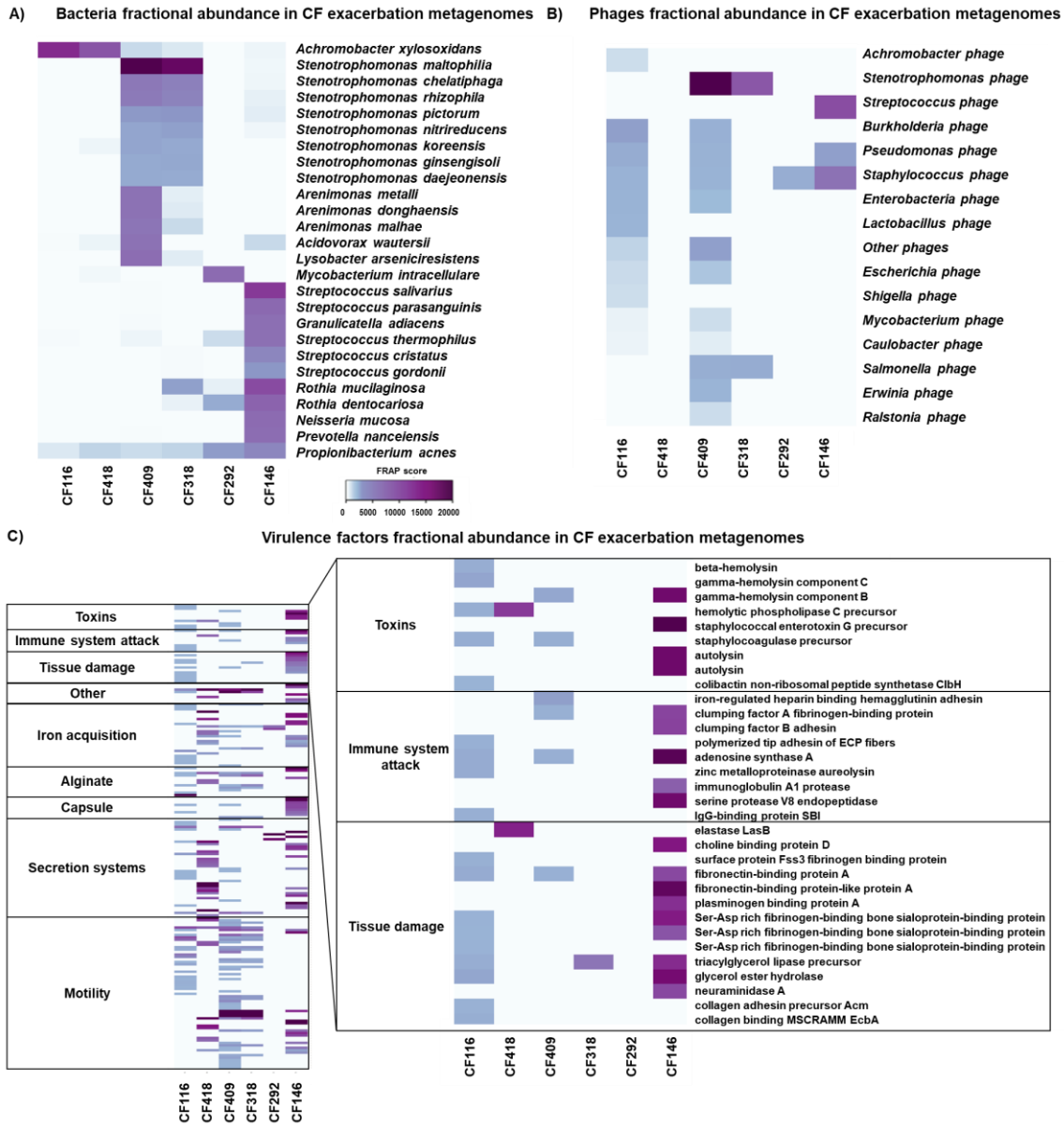
A) **CF318 metagenome, *Stenotrophomonas* contigs with insertions**



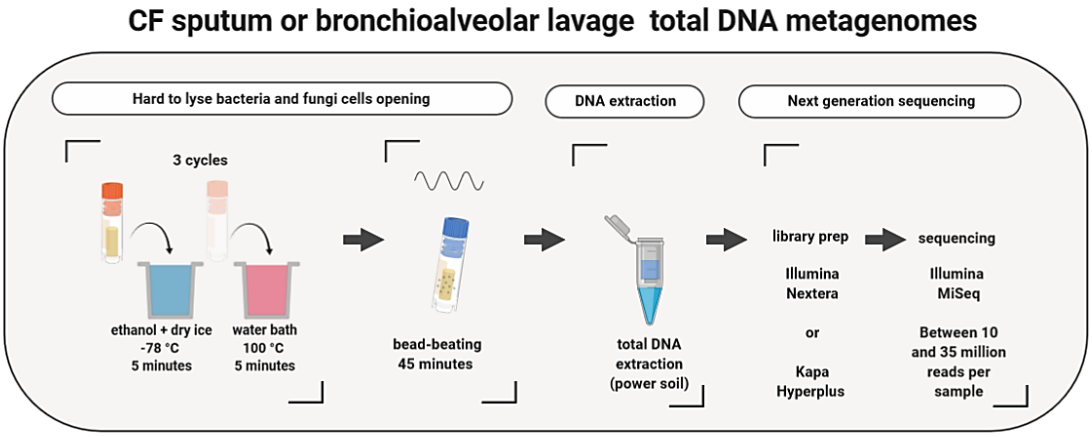
B) **Inserted regions in *Stenotrophomonas maltophilia* CF318**

Stenotrophomonas maltophilia CF318B contig insert regions	length (bp)	number of cds	phage cds	mobile element cds	integrase cds	transposase cds	hypothetical cds	start position in contig	end position in contig	strain in contig
Insert region 1 : hydrolase, ligase, TonB receptor	11,529	11	0	0	0	0	3	16,814	28,343	+
Insert region 2 : phage/mobile element	**15,806**	**38**	**1**	**1**	**0**	**0**	**36**	**6,081**	**21,887**	**+**
Insert region 3	12,509	9	0	0	0	0	7	20,677	33,186	+
Insert region 4 : transcriptional regulation	15,015	25	0	0	0	0	13	9,119	24,134	+
Insert region 5	13,609	17	0	0	0	0	11	12,696	26,305	+
Insert region 6 : phage	**10,744**	**15**	**1**	**0**	**1**	**0**	**14**	**56,859**	**67,603**	**+**
Insert region 7 : transcription, ribosomal proteins	16,338	20	0	0	0	0	8	16,877	33,215	+
Insert region 8 : carbohydrates metabolism	12,939	13	0	0	0	0	7	4,379	17,318	+
Insert region 9 : carbohydrates metabolism	10,061	12	0	0	0	0	5	25,265	35,326	+
Insert region 10	14,877	15	0	0	0	0	10	6,895	21,772	+
Insert region 11 : phage	**10,650**	**18**	**3**	**0**	**1**	**0**	**14**	**6,392**	**17,042**	**+**
Insert region 12 : efflux RND trasporters	29,676	41	0	0	0	0	29	1	29677	+
Insert region 13 : efflux RND transporters	16,852	27	0	0	0	0	15	1	16,853	+

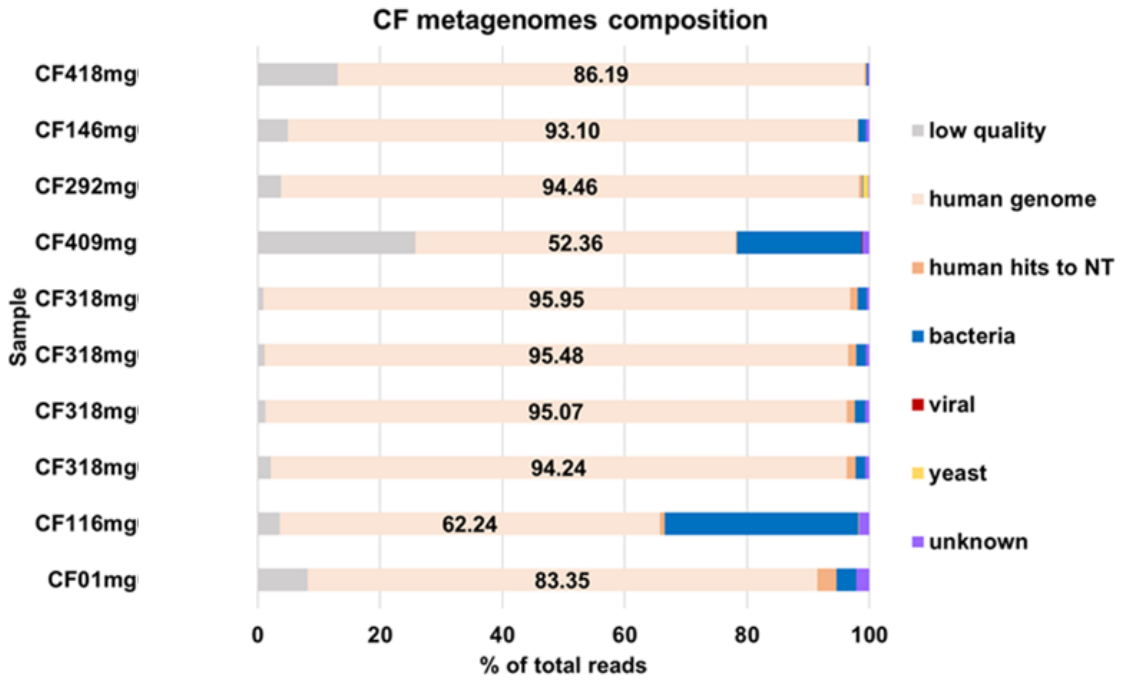
Supplemental Figure 4.12 Insertions in contigs from CF318 metagenome A) Synteny between the closest reference genome and contigs assembled from CF318 metagenome. B) Addon regions identified in assembled contigs.



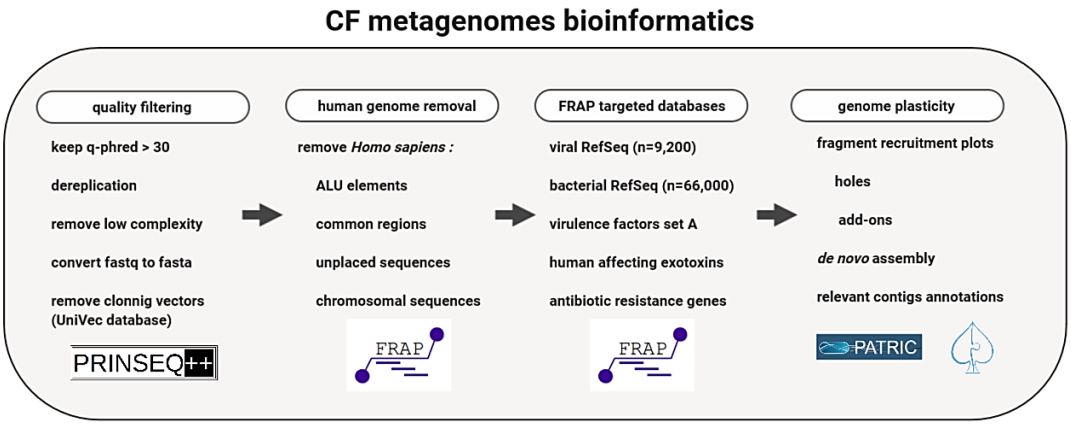
Supplemental Figure 4.13 Bacteria and bacteriophages fractional abundance in metagenomes from CF acute exacerbations. A) Bacteria refseq database (n=66,000 genomes) B) Bacteriophages refseq (n=4,500 genomes). Polished metagenomes were compared to each database using fragment recruitment assembly purification (FRAP) at 96% identity over 100% of the read. C) Virulence factors. Polished metagenomes were compared to virulence factors database Set A using fragment recruitment assembly purification (FRAP) at 96% identity over 100% of the read.



Supplemental Figure 4.14 CF sputum bronchioalveolar lavage total DNA metagenomes prodecure.



Supplemental Figure 4.15 CF metagenomes composition. Raw reads were quality filtered using prinseq++ (q-phred > 30), compared to the human reference genome using smalt (y = 0.5). Polished sequences were compared to NT using BLASTn with an e-value cutoff of 0.1 with output taxonomy for each hit, the best hit was kept.



Supplemental Figure 4.16 CF metagenomes bioinformatic methods.

Chapter 5 : Compounding Achromophages for therapeutic applications

Abstract

Achromobacter spp. colonization in Cystic Fibrosis respiratory airways is an increasing concern. Two adult Cystic Fibrosis patients colonized by *Achromobacter xylosoxidans* CF418 or *Achromobacter ruhlandii* CF116 experienced fatal exacerbations. *Achromobacter* species are naturally resistant to several antibiotics, therefore the use of phage therapy for the control of *Achromobacter* is proposed in this study. Twelve lytic phages were isolated and characterized at a morphological and genomic level. They are presented as the *Achromobacter* phages Kumiai collection. The proposed methods for large-scale production of phages and removal of endotoxins resulted in a phage concentrate of $1x10^9$ plaque forming units per mililiter with an endotoxin concentration of 65 endotoxin units per mililiter, which is below the Food and Drugs Administration recommended maximum threshold for human administration. The infectivity of all phages in the Kumiai collection was tested in 23 *Achromobacter* clinical isolates. Eighteen out of the 23 tested *Achromobacter* isolates were lysed by at least one phage. Six distinctive *Achromobacter* phage genome clusters were identified based on a comprehensive phylogenetic analysis of all publicly available *Achromobacter* phage genomes. A cryptic prophage was induced in *Achromobacter xylosoxidans* CF418 when infected with lytic phages. This prophage genome was also characterized and is presented as *Achromobacter phage* CF418-P1. Thus lytic-lysogenic phage interactions require further exploration in the context of phage therapy interventions.

This study provides a framework for the isolation and characterization of phages to kill *Achromobacter* species in order to manage Cystic Fibrosis pulmonary infections.

Introduction

Achromobacter spp. were identified as the dominant member of the microbial community in sputum samples from two Cystic Fibrosis (CF) patients that suffered acute exacerbations. *Achromobacter* spp. are Proteobacteria of the order Burkholderiales, they can use anaerobic metabolism in the presence of nitrate or nitrite and can use denitrification for respiration. *Achromobacter* spp. are long term colonizers of the CF lungs (Ridderberg, Nielsen, and Nørskov-Lauritsen 2015) and of increasing concern. *Achromobacter xylosoxidans* isolated from CF patients have pathogenic characteristics such as the presence of toxins and virulence factors (Jakobsen et al. 2013). In addition to infections in the CF lungs, *Achromobacter* spp. have been reported to cause urinary tract infections (Tena et al. 2008), endocarditis (Tokuyasu et al. 2012), meningitis (Manckoundia et al. 2011), and ocular infections (Park, Song, and Koh 2012).

Bacteriophages (phages) have been used as therapeutics to modify the microbiome and aid in the clearance of bacterial infections (Gordillo-Altamirano and Barr 2019, Kortright et al. 2019). A CF patient was treated with therapeutic phages targeting *Achromobacter* spp. (Hoyle et al. 2018). The use of phages as therapeutics has been empirical and the phages applied to patients are not always characterized (Kutateladze and Adamia 2008), which poses a safety concern.

There are 24 publicly available *Achromobacter* phages (Achromophages) genomes in the literature (Table 5.1). The first sequenced Achromophages (Wittmann, Dreiseikelmann, Rohde, Meier-Kolthoff, et al. 2014) are members of the N4-like phages, *Achromobacter phage* JWAlpha and *Achromobacter phage* JWDelta, both of them are *Podoviridae.* Several

Achromophages have been isolated, but their genomes are not sequenced (Wittmann, Dreiseikelmann, Rohde, et al. 2014).

The need for lytic phages targeting *Achromobacter* is addressed by the isolation and characterization of twelve lytic phages. Their host range was tested in 23 *Achromobacter* clinical isolates from CF patients. Phage preparation methods to remove endotoxins were successful in providing high titer phages with low endotoxin levels. *Achromobacter xylosoxidans* CF418 carries a temperate phage that was induced when the strain was infected with lytic phages. These dynamics deserve further attention. Considerations when isolating and characterizing phages for therapeutic applications are addressed, such as the presence of prophages in the host strain, presence of toxins in the isolated phages, and lack of functional annotation on most phage proteins.

Results

Cystic fibrosis and Achromobacter

Achromobacter spp. were identified as the dominant members of the microbial community in two CF fatal exacerbations. Patient CF116 respiratory tract microbial community was dominated by *A. ruhlandii* with a relative abundance of 98.5% based on a sputum metagenome. The clinical laboratory reported the presence of *Achromobacter* sp. and *A. xylosoxidans*. Patient CF116 was treated with the following antibiotics over a month during the acute exacerbation: ceftazidime-avibactam, doxycycline, sulfamethoxazole-trimethoprim, vancomycin, and tigecycline. A second scheme during the acute exacerbation was composed of colistin, azithromycin, meropenem, imipenem-cilastatin, azithromycin, and minocycline.

Patient CF116 *Achromobacter* spp. infection was not resolved and the patient died three months after the acute exacerbation.

Patient CF418 respiratory tract microbial community was dominated by *A. xylosoxidans* with a relative abundance of 76.7% based on a bronchioalveolar lavage metagenome obtained during an acute exacerbation. The clinical laboratory reported a rhinovirus infection one week before the acute exacerbation and chronical presence of *Achromobacter* sp. and *Pseudomonas aeruginosa*. Patient CF418 was on cardiopulmonary bypass and waiting for a lung transplant. Patient CF418 died during the acute exacerbation.

These unresolved *Achromobacter* spp. infections in two fatal exacerbations motivated the implementation of a phage therapy strategy to kill the bacteria.

Achromobacter *clinical isolates*

Achromobacter spp. clinical isolates were obtained during acute exacerbations of patients CF116 and CF418. These isolates were characterized by 16S rDNA amplicon Sanger sequencing and whole genomes were obtained using Nanopore and Illumina sequencing. Strains are further referred as *Achromobacter ruhlandii* CF116, and *Achromobacter xylosoxidans* CF418.

Twenty additional *Achromobacter* clinical isolates from CF patients were obtained from the San Diego CF clinic. A reference *Achromobacter xylosoxidans* strain C54, HM-235 from a non-CF individual was obtained from the BEI collection (BEI, 2019).

Achromophage isolation and characterization

Availability of genomically characterized Achromophages is limited. Twenty four Achromophage genomes (Wittmann, Dreiseikelmann, Rohde, Meier-Kolthoff, et al. 2014; Rohde, Nimtz, and Wittmann 2017; Ma et al. 2016; Li, Yin, et al. 2016; Li, Zhao, et al. 2016) were available in public databases (Table 5.1).

Two genomically characterized Achromophages were obtained from the DSMZ-German collection of microorganisms (Leibniz Institute, 2019). These are *Achromobacter phage* JWalpha (DSM 26830) and *Achromobacter phage* JWDelta (DSM 26829). Their propagation was not succesfull in any of the *Achromobacter* spp. strains in our collection (n=23). A previous study reported the use of Achromophages in a CF patient as a phage therapy clinical intervention (Hoyle et al. 2018). Access to these phages was requested to the research group, but it was not granted. This motivated a phage hunt strategy to construct a publicly available Achromophage library whose genomes are characterized.

Table 5.1 Achromobacter lytic phages in the literature (n=24).

Phage name	Genome length (bp)	Accession number	Clade	Morphology (virus family)	Reference	Propagation strain
Achromobacter phage vB_AXyS_19-32_Axy04	73,834	MK962626	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy06	45,830	MK962627	JWX		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy09	43,287	MK962628	Axy09		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy10	73,898	MK962629	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC

Table 5.1 Achromobacter lytic phages in the literature (n=24). (Continued)

Phage name	Genome length (bp)	Accession number	Clade	Morphology (virus family)	Reference	Propagation strain
Achromobacter phage vB_AXyS_19-32_Axy11	73,413	MK962630	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy12	74,096	MK962631	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy13	70,103	MK962632	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy14	46,703	MK962633	JWX		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy16	46,178	MK962634	JWX		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy18	45,500	MK962635	phiAxp1		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy19	46,036	MK962636	phiAxp1		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy20	46,352	MK962637	phiAxp1		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy21	43,049	MK962638	Axy09		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AXyS_19-32_Axy22	71,710	MK962639	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AxyS_19-32_Axy23	43,773	MK962640	Axy09		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage vB_AxyS_19-32_Axy24	74,744	MK962641	JWAlpha		Unpublished (Pourcel C, et al., 2019)	*Achromobacter xylosoxidans* I2BC
Achromobacter phage phiAxp-1	45,045	NC_029033.1	phiAxp1	Siphoviridae	(Li E, et al., 2016)	*Achromobacter xylosoxidans* A22732

Table 5.1 Achromobacter lytic phages in the literature (n=24). (Continued)

Phage name	Genome length (bp)	Accession number	Clade	Morphology (virus family)	Reference	Propagation strain
Achromobacter phage phiAxp-2	62,220	NC_029106.1	phiAxp2	Siphoviridae	(Li E, et al., 2016)	*Achromobacter xylosoxidans* A22732
Achromobacter phage phiAxp-3	72,825	NC_028908.2	JWAlpha	Podoviridae	(Ma Y, et al., 2016)	*Achromobacter xylosoxidans* A22732
Achromobacter phage 83-24	48,216	NC_028834.1	JWX	Siphoviridae	(Rohde M, et al., 2017)	*Achromobacter xylosoxidans* HER 83-190
Achromobacter phage JWX	49,714	NC_028768.1	JWX	Siphoviridae	(Rohde M, et al., 2017)	*Achromobacter xylosoxidans* LMG 3465
Achromobacter phage JWF	81,541	NC_029075.1	JWF	Siphoviridae	(Rohde M, et al., 2017)	*Achromobacter xylosoxidans* CCUG 48136
Achromobacter phage JWDelta	73,659	KF787094.1	JWAlpha	Podoviridae	(Wittmann J, et al., 2014)	*Achromobacter xylosoxidans* DSM 11852
Achromobacter phage JWAlpha	72,329	NC_023556.1	JWAlpha	Podoviridae	(Wittmann J, et al., 2014)	*Achromobacter xylosoxidans* DSM 11852

Twelve lytic phages were isolated from environmental water sources including lakes, ponds, fountains, and influents of wastewater treatment plants in San Diego, CA (Table 5.2). The set of 12 phages is referred to as the *Achromobacter* phages Kumiai collection. Seven Achromophages were isolated on *A. ruhlandii* CF116 using LB media. It was not possible to isolate phages in *A. xylosoxidans* CF418 using LB media. Therefore the media was changed to BHIS and five Achromophages were isolated in *A. xylosoxidans* CF418. A previously uncharacterized prophage was induced from *A. xylosoxidans* CF418 when the strain was infected with lytic phages. The 12 new Achromophage names etymologies are from the Kumiai language spoken by native San Diegans (also known as Kumeyaay).

Table 5.2 Achromobacter bacteriophages isolated in this study. Etymologies are from the Kumiai language. All isolation sources are from San Diego, CA. Coding sequences, tRNAs and repeat regions were annotated from PATRIC.

Phage name	Genome length (bp)	GC content	CDS	tRNA	repeat regions	Clade	Host	Phage isolation source	Phage name etymology
Achromobacter phage nyashin	45,982	56.31	68	0	2	phiAxp1	*A.ruhlandii* CF418	Influent water	*nyashin* – sun
Achromobacter phage shaaii	45,029	56.11	63	0	0	phiAxp1	*A.ruhlandii* CF418	Influent water	*shaaii* - buzzard
Achromobacter phage nyaak	46,478	55.77	66	1	2	JWX	*A. ruhlandii* CF116	Influent water	*nyaak* - north
Achromobacter phage kewaak	46,215	56.19	66	1	0	JWX	*A. ruhlandii* CF116	SDSU fishpond	*kewaak* - south
Achromobacter phage wiik	50,543	55.75	83	1	37	JWX	*A. ruhlandii* CF116	Lake Murray	*wiik* - east
Achromobacter phage tuull	47,460	55.79	92	1	15	JWX	*A. ruhlandii* CF116	Influent water site 2	*tuull* - west
Achromobacter phage maay	46,086	56.31	62	1	2	JWX	*A. ruhlandii* CF116	Influent water site 1	*maay* - clouds or sky
Achromobacter phage xasilly	46,478	55.77	65	1	2	JWX	*A. ruhlandii* CF116	Influent water site 3	*xasilly* - sea or waves
Achromobacter phage ahaak	46,435	56.17	64	1	0	JWX	*A.ruhlandii* CF418	Influent water	*ahaak* – raven
Achromobacter phage emuu	46,012	55.86	62	1	2	JWX	*A.ruhlandii* CF418	Influent water	*emuu* - mountain sheep
Achromobacter phage ewii	43,305	55.51	64	1	0	JWX	*A.ruhlandii* CF418	Influent water	*ewii* - snake
Achromobacter phage kwar	33,215	55.59	73	0	6	JWX	*A. ruhlandii* CF116	Influent water site 4	*kwar* - red
Achromobacter prophage CF418-P1	58,030	65.63	73	1	2		*A.ruhlandii* CF418		

Achromophage genomes were sequenced using Illumina (Supplemental Table 5.1) and Nanopore (Supplemental Table 5.2) platforms. Genome lengths were between 33,215 bp and 50,543 bp. The GC content was between 55% and 56%. The number of coding sequences on

each phage was between 62 and 92. A tRNA was identified in 10 phages. Achromophage genome sizes for 10 out of 12 phages were corroborated by Pulse Field Gel Electrophoresis (PFGE), all the genomes were close to the 48.5 Kbp marker (Supplemental Figure 5.1). Achromophage morphology was determined by Transmission Electron Microscopy (TEM); siphoviridae and podoviridae morphologies were observed (Supplemental Figure 5.2).

Achromophage comparative genomics

Protein level whole genome comparisons to the Phage Proteomic Tree (Rohwer and Edwards 2002) were performed among previously published Achromophages (n=24, Table 5.1) and Achromophages form this study (n=12, Table 5.2) using VIPTree (Nishimura et al. 2017) (Figure 5.3-A). Isolated Achromophages clustered in 2 clades (Figure 5.1-B). Clade JWX (from *A. phage* JWX, NC_028768.1) included *Achromobacter phage tuull, Achromobacter phage emuu, Achromobacter phage ahaak, Achromobacter phage kewaak, Achromobacter phage maay, Achromobacter phage kwar, Achromobacter phage nyaak, Achromobacter phage xasilly, Achromobacter phage wiik,* and *Achromobacter phage ewii*. Clade phiAxp-1 (from *A. phage* phiAxp-1, NC_029033.1) included *Achromobacter phage shaaii* and *Achromobacter phage nyashin*. *Achromobacter phage* 83-24 (NC_028834.1) clustered in the JWX clade. *Achromobacter phage* phiAxp-2 (NC_029106.1) did not clustered with other Achromophages, it clustered with *Burkholderia* and *Xylella* phages. A third cluster identified as JWDelta (after A. *phage JWDelta*, KF87094.1) was formed by *A. phage JWDelta*, *A. phage JWAlpha* and *Achromobacter phage phiAxp-3*. *Achromobacter phage JWF* (NC_029075.1) did not clustered with other Achromophages and was more distantly related

to the rest of Achromophages, it clustered close to the Archaea virus *Natrinema virus SNJ1* and four *Holoarcula* viruses.

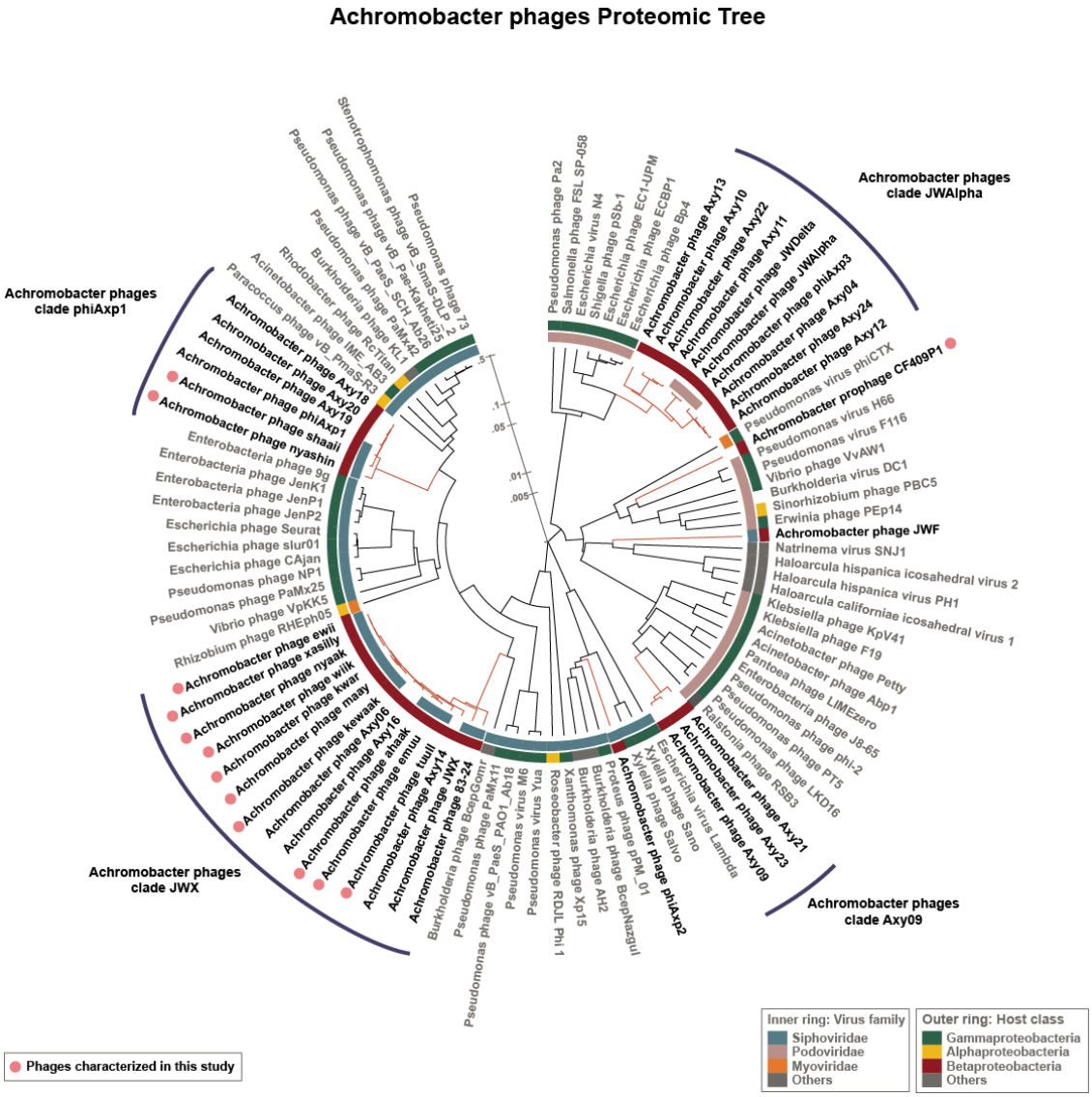


Figure 5.1 Achromobacter phages on The Phage Proteomic Tree.

Small variations were observed between phages in the Clade JWX (Figure 5.2). *A. phage emuu*, *A. phage ahaak*, and *A. phage kewaak* showed variation in a 3.5 Kbp region

containing 6 CDS. *A. phage ahaak* and *A. phage kewaak* showed variation in a tail fiber protein. *A. phage kwar* was closely related to *A. phage maay* and A. *phage nyaak*. Deletions were observed in *A. phage kwar*. *A. phage nyaak*, and *A. phage xasilly* differed only by 50 nucleotides in a CDS that is 118 aminoacids long and contains a ribon-helix-helix domain. *A. phage xasilly*, *A. phage wiik*, and *A. phage ewii* were closely related and rearrangements on their genomes were observed in the synteny plots, a 4 Kbp insertion was observed in *A. phage wiik*.

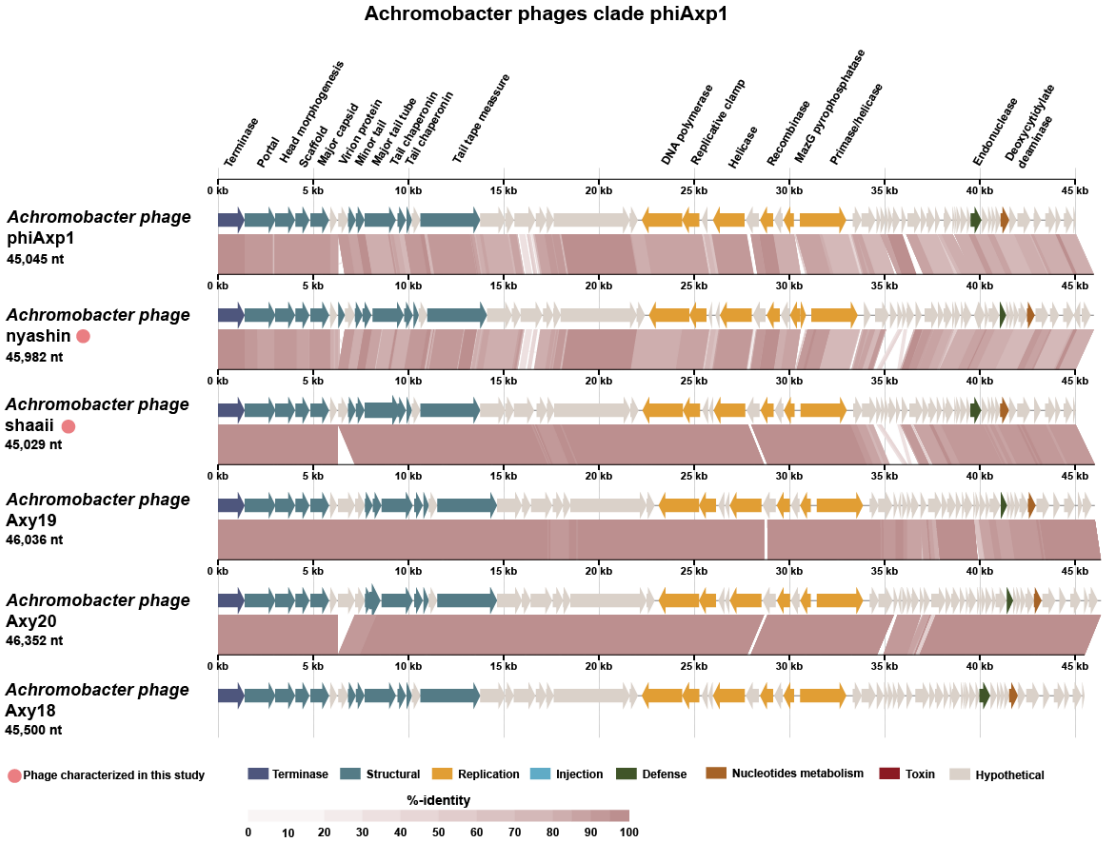


Figure 5.2 Achromobacter phages clade phiAxp1.

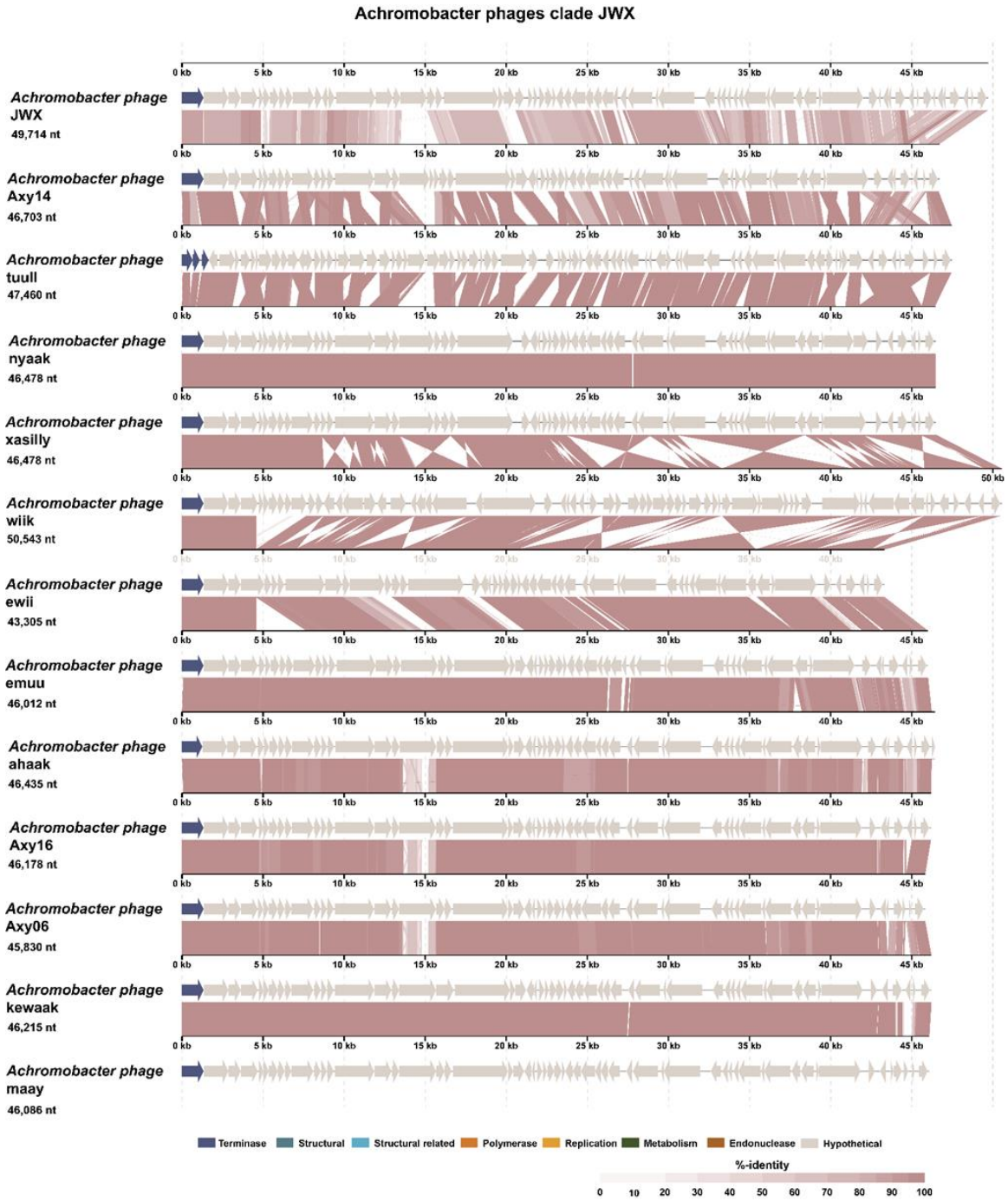


Figure 5.3 Achromobacter phages clade JWX, genomes comparisons.

In clade phiAxp1 (Figure 5.3) *A. phage phiAxp-1,* and *A. phage shaaii* were similar, with variation in a 3.2 Kbp region that contained 8 CDS, this variation region was also observed in *A. phage nyashin*.

Achromophage genome annotation

Achromophages genomes from the Kumiai collection encode between 62 and 92 CDS per genome. Genome annotations using protein comparisons based on k-mers (Brettin et al. 2015) resulted in mostly hypothetical proteins, which is a common challenge in phage annotations. Further annotations using conserved domains (CDD), hidden Markov models (HMMER Search), and artificial neural networks (ANNs for structural proteins), followed by expert curation, were used to annotate the phage genomes. Using this holistic approach, the number of hypothetical proteins was reduced. A terminase was identified in all analyzed phage genomes. The small and large terminases form the packaging machinery (Sun et al. 2012) which is responsible for feeding the DNA into the phage capsids. A transcriptional regulator was identified in all analyzed phage genomes, its original annotation was done in the ACLAME server (ACLAME, 2019). A tRNA was identified on each genome of the clade JWX, no tRNA was identified in the phages of clade phiAxp1.

Toxins annotations in Achromophages

A concern in the use of phages for therapeutic application is the production of toxins. Achromophages from the Kumiai collection have no identified toxins or virulence factors when annotated using PATRIC resources, which perform protein level comparisons to the

virulence factors database (~30,000 virulence factors reported by Chen et al. 2016), VICTORS (5,296 virulence factors reported by S. Sayers et al. 2019), and a PATRIC curated virulence factors database (1,293 virulence factors reported by Wattam et al. 2017). Conserved domains database annotations identified a potential toxin in one phage. A 39 aa long CDS in *A. phage tuull* was annotated by HMMER Search (EBI, 2019) as a hemolysin-type calcium binding region, 12 out of 19 aa were recognized as part of the toxin motif.

Achromophage lifestyle determination

The use of temperate phages for phage therapy is not desirable since the integration of phages in bacteria genomes may have unexpected results. No integrases were detected in any of the Achromobacter phages of the Kumiai collection. Lifestyle characterization using PHACTS (McNair, Bailey, and Edwards 2012) classified *A. phage shaaii* and *A. phage nyashin* as lytic, the lifestyle classification for the remaining 10 phages was inconclusive, potentially due to the lack of well characterized Achromophages for comparisons (Supplemental Table 5.4).

Prophage CF418-P1 *induction and characterization*

Multiple phage genomes were identified in the genome sequencing of phage lysates LB5, LB7 and LB8 (Supplemental Figure 5.3). A new prophage *CF418-P1*, was identified (Figure 5.4-A). It was induced when *A. xylosoxidans* CF418 was infected by Achromophages. *Prophage CF418-P1* was induced with the following pairs of phages: *A. phage shaii* and A. *phage ahaak* were present in lysate LB5; *A. phage nyashin,* and *A. phage ewii* were present in phage lysate LB7; and *A. phage nyahin* and *A. phage emuu* were present in lysate LB8. Reads

from lysates LB6, LB5, LB7, and LB8 were mapped to the reference genome *Achromobacter xylosoxidans* NCTC10807 and showed recruitment in the prophage region (Supplemental Figure 5.4).

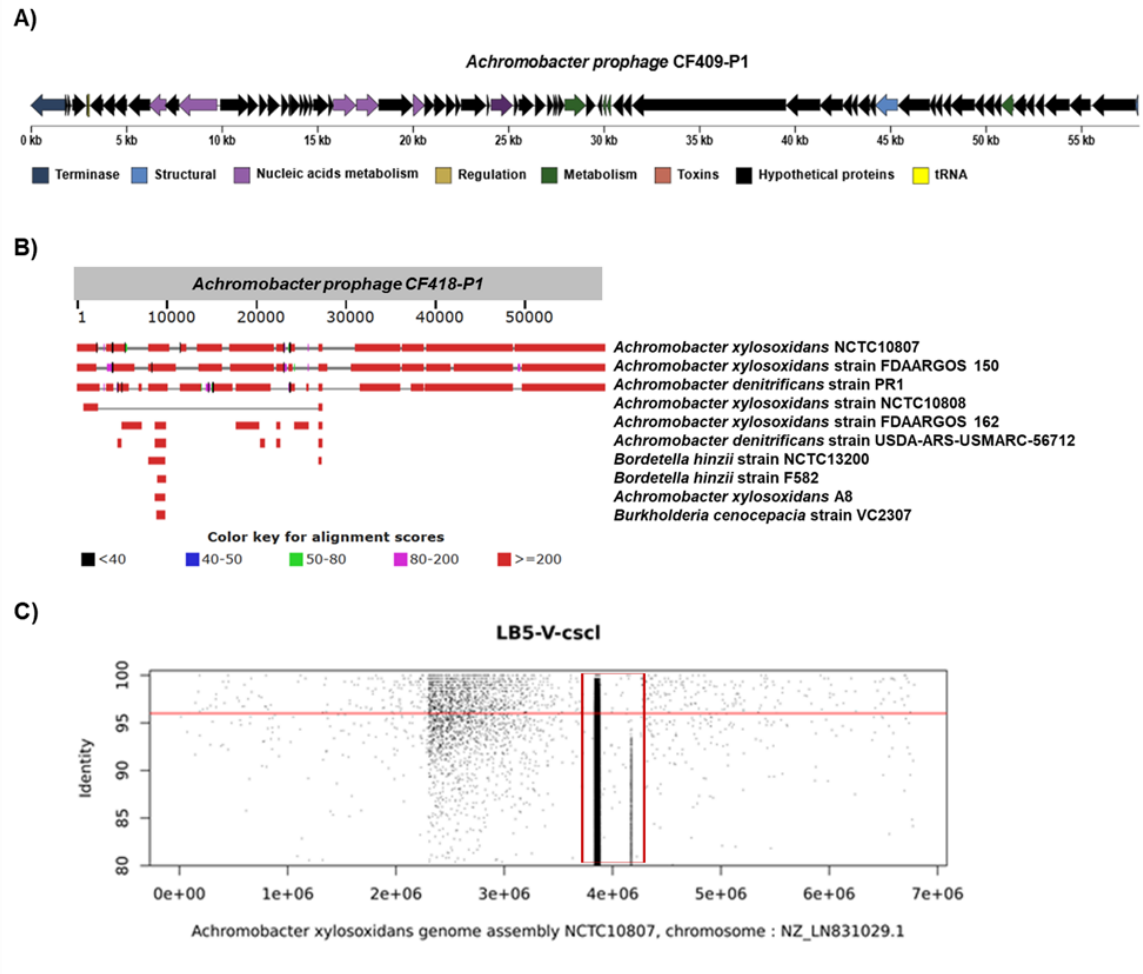


Figure 5.4 Achromobacter prophage induced when infected with other lytic phages. A) Prophage in *A. xylosoxidans* CF418.The prophage genome was assembled from samples LB5, LB7 and LB8 using SPades (Bankevich et al. 2012). Prophage annotation using PATRIC, CDD, HMMER and ANNs. B) Prophage *in A. xylosoxidans* CF418 in bacteria genomes. C) Prophage *in A. xylosoxidans* CF418. Fragment recruitment plots of phage lysate against *Achromobacter xylosoxidans* NCTC10807 reference genome. Prophage region recruited reads around 5.8 Mbp (highlighted in red square).

Prophage CF418-P1 showed identity to *A. xylosoxidans* NCTC10807 (96.85% identity and 73% query coverage), *Achromobacter xylosoxidans* strain FDAARGOS_150 (95.70% identity and 74% query coverage), and *Achromobacter denitrificans* strain PR1 (82.77% identity and 55% query coverage) (Figure 5.4-B). *Prophage CF418-P1* is not closely related to the Achromophages from the Kumiai collection, it formed a separate branch in the Phage Proteomic Tree and shared 6 short regions with low identity to *Pseudomonas virus* H66 and *Pseudomonas virus* F116 (Supplemental Figure 5.5).

Prophage CF418-P1 genome length is 58,030 bp, with a GC content of 65.7%, it encodes an integrase and has a temperate lifestyle. Nucleic acid metabolism and structural proteins were identified, as well as an endopeptidase and a peptidoglycan hydrolase. Most likely, *Prophage CF418-P1* packages pieces of the host genome in its viral capsids, being able to perform generalized transduction. The fragment recruitment plots to the host genome (Figure 5.4-C) suggested that.

Host range determination for Achromophages

The host range of Achromophages from the Kumiai collection was tested in 22 *Achromobacter* isolates from 13 patients with CF in the San Diego clinic, and the reference strain *A. xylosoxidans* HM-235 (Figure 5.5). Eighteen out of the 23 tested *Achromobacter* isolates were lysed by at least one phage from the Kumiai collection (n=12). Additional phages were isolated to cover a broader host range, these five phages (Supplemental Table 5.5) are *Achromobacter phage* SE2*, Achromobacter phage* M1, *Achromobacter phage* ENA1, *Achromobacter phage* M2, *and Achromobacter phage* MW2. Twenty out of the 23

Achromobacter strains were infected and lysed by at least one of the isolated lytic phages. Three *Achromobacter* clinical strains were not infected by any of the isolated phages, we hypothesize that these strains carry more prophages which prevent further infections and pose a challenge for lytic phages isolation. *A. phage* SE2 presented the broadest host range, it infects 13 out of 23 tested strains. *A. phage nyashin* showed a generalist behavior since it infected 10 out of 23 tested strains, and *A. phage* MW2 showed a specialist behavior since it only infected 2 out of 23 tested strains.

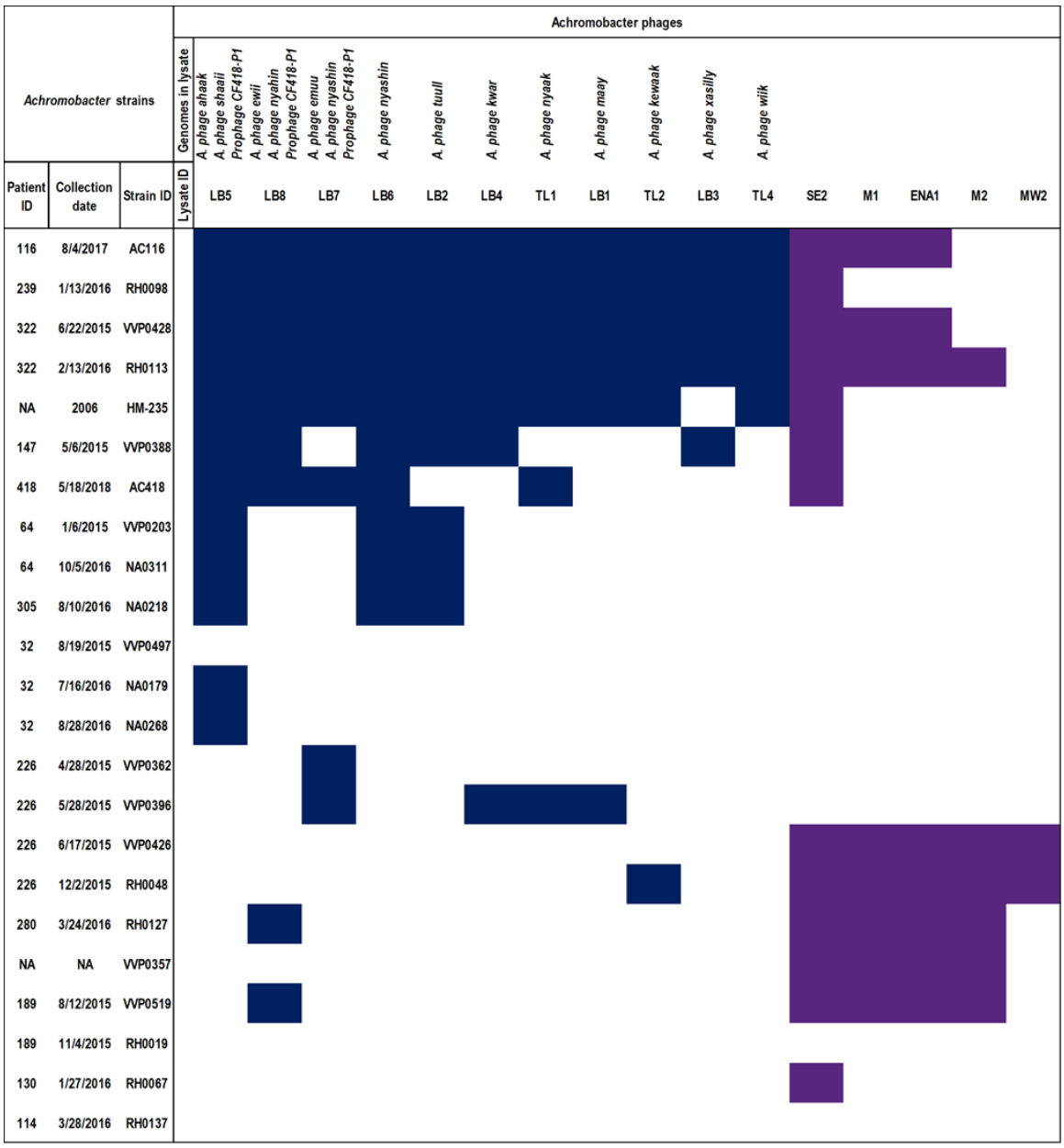


Figure 5.5 *Achromobacter* phages host range test. Twentytwo *Achromobacter* strains isolated from CF patients sputum and characterized by the clinical lab were used for host range test. *Achromobacter* reference strain (HM-235) was also used for host range test. Host range test was tested by spot test on a lawn of the host bacteria.

Achromobacter phages lysates preparation for therapeutic applications

Phage production for therapeutic application requires a high titer phage lysate with a low concentration of lipopolysaccharides (also known as LPS, lipoglycans or endotoxins). Thus, the production method needs to preserve the phage titer and ideally concentrate it, be fast, and avoid the use of toxic compounds. This method was illustrated using *A. phage nyaak* (Figure 5.6), the final preparation is a 0.15 M NaCl solution with a phage concentration of $1x10^{11}$ PFU ml^{-1} and a lipopolysaccharides concentration of 6500 EU ml^{-1}.

The proposed production method for Achromophages (Figure 5.6) started with one liter of overnight phage culture in LB media with a titer of $1x10^{9}$ PFU ml^{-1}. Bacterial cell debris was reduced by centrifugation and decantation, followed by the elimination of remaining bacteria cells by filtration. It is important to note that remaining bacterial cells were not disrupted by chloroform treatment to avoid an increase in LPS in the solution. This solution was concentrated to a volume of 55 ml using tangential flow filtration. It had a phage titer of $1x10^{10}$ PFU ml^{-1}. Phages are more stable in SM buffer than in LB, thus a buffer exchange and further concentration was performed via ultrafiltration using a 100 kDa regenerated cellulose membrane, this step resulted in a 10.6 ml solution with a phage titer of $1x10^{11}$ PFU ml^{-1}. Lipopolysaccharides were removed by 1-Octanol (Szermer-Olearnik and Boratyński 2015; Morrison and Leive 1976). The residual 1-Octanol was eliminated from the solution via dialysis, this resulted in 13 ml of a 0.15 M NaCl solution with a phage titer of $1x10^{11}$ PFU ml^{-1}. Lipopolysaccharides in the phage stock solution were quantified as endotoxin units (EU) using recombinant factor C based fluorescence detection (EndoZyme II, Biomérieux).

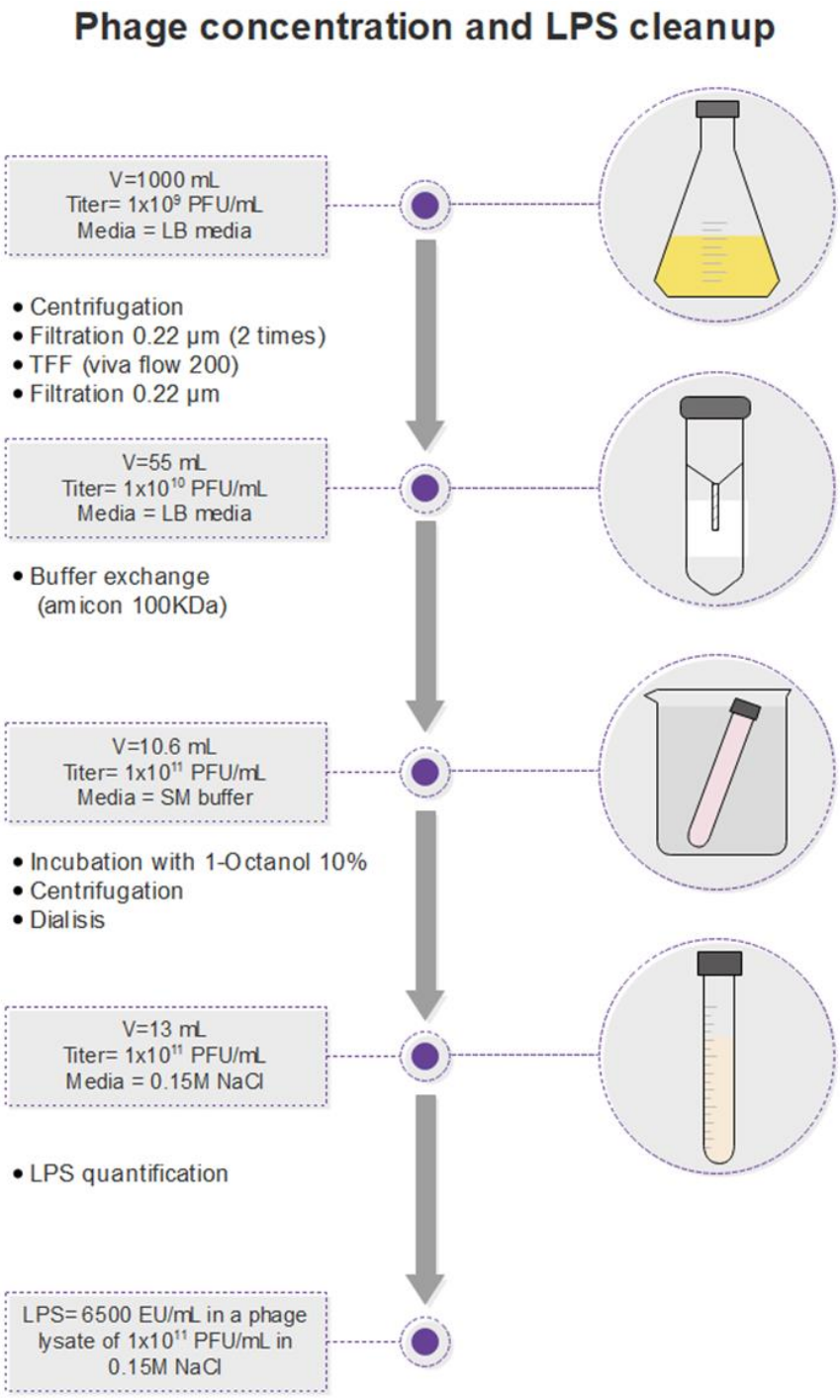


Figure 5.6 *Achromobacter phage nyaak* high titer lysate production and endotoxin removal procedure.

Discussion

Achromophages hunting

Phages capable of killing *Achromobacter* spp. were isolated from water samples collected around San Diego, CA. *A. ruhlandii* CF116 phages were isolated in the first two weeks of the search, however *A. xylosoxidans* CF418 phages were isolated in six months. The prophage present in *A. xylosoxidans* CF418 could be providing superinfection exclusion agains invading phages, therefore the isolation of phages against this strain was challenging. In the *Achromobacter* spp. collection (n=23) used in this study, three bacteria strains couldn't be infected by the tested lytic phages. Publicly available phage collections and cooperation among research groups are relevant resource for the advancement of the phage therapy field.

Achromophages genomics

Two distinctive clades were identified in the Achromophages Kumiai collection, this information may be usefull to design phage cocktails, since theoretically phages that are distantly related have a distinctive host range.

Toxins characterization in phage genomes.

Characterization of phage toxins based on genome sequencing is challenging (Philipson et al. 2018). A toxin motif was identified in *A. phage tuull,* therefore is not a suitable candidate for therapeutic use. No known toxins were identified in the rest of the Kumiai collection phages. Since most of phage proteins are annotated as hypotheticals, there is still a concern for the presence of uncharacterized toxins in phages intented for therapeutic

use. An alternative to evaluate their safety is to test the phage cytotoxicity in eukaryotic cell lines (Shan et al. 2018; Porayath et al. 2018) and their tolerance in animal models.

Cryptic prophages induction

During the isolation of lytyc phages, a cryptic prophage was induced from *A. xylosoxidans* CF418, wether this phage induction is promoting the expression of toxins is not known. Generalized transduction was observed in *Prophage CF418-P1,* since it packages pieces of the host genome in its viral capsids. This phenomenon has been observed in other Achromophages, such as *Achromobacter phage* α (Woods and Thomson 1975). Cryptic *Achromobacter* prophages induction mediated by superinfected by a related phage has been reported before (Thomson and Woods 1974).

Endotoxin removal from phage lysates for therapeutic applications.

The proposed method for endotoxin removal from phage lysates was succesfull in the number of phages recovered and the low amount of endotoxin in the prepatation. Phages for therapeutic applications are usually applied as a $1x10^9$ PFU ml^{-1} solution. The phage stock presented in this work, when diluted to a concentration of $1x10^9$ PFU ml^{-1} has a final concentration of 65 EU ml^{-1}. The FDA recommended maximum amount of endotoxin units (EU) in an intravenous solution is 5 EU per kilogram of body weight per hour. Up to 6 ml of the phage stock of $1x10^9$ PFU ml^{-1} which contains 65 EU ml^{-1} can be applied to a patient (average weight of 80 kg) every hour. Dosage of phage preparations is empirical and efforts towards understanding phage pharmacokinetics are needed (Malik et al. 2017).

A concern in the removal of endotoxin is the residual amount of 1-Octanol. It is important to note that 1-Octanil can be metabolized in the human body by class IV alcohol dehydrogenases (Danielsson et al. 1994; Satre, Žgombić-Knight, and Duester 1994; Jelski et al. 2006), which is an advantage of using 1-octanol for LPS removal, although other methods are available such as Hexafluoro isopropanol (McCord, Muddiman, and Khaledi 2017).

Beyond phage hunting

Alternative solutions to phage isolation are the engineering of phages to expand their host range (Lemire, Yehl, and Lu 2018); the use of phage tails to for therapeutic applications (Scholl 2017); or phage derived endolysins (Young and Gill 2015).

Materials and methods

Cystic fibrosis metagenomes

Informed consent was obtained from patients CF116 and CF418. This study was approved by UCSD (HRPP 081510) and San Diego State University (IRB#1711018R). For patient CF116, a sputum sample was collected in a sterile cup after sputum induction. For patient CF418 a bronchioalveolar lavage sample (BAL) was collected. From BAL or sputum samples, 500 µl were transferred to a cryovial. The tube was submerged in a dry-ice and ethanol bath for 5 minutes, then transferred to a water bath at 100 ºC for 5 minutes, this process was repeated 3 times. The sample was transferred to the beads tube from Qiagen power soil DNA extraction kit (catalog number 12888-100) and homogenized by shaking for 45 minutes. The rest of the Qiagen power soil DNA extraction kit protocol was followed. Ten

nanograms of DNA were used for Nextera library prep. Libraries were sequenced on the Illumina MiSeq platform.

Achromobacter strains

Achromobacter clinical isolates from patients CF116 and CF418 were phenotypically characterized by the UCSD clinical laboratory and grown in Remel blood agar. Both clinical isolates can grow in tryptone yeast extract glucose medium (TYG), supplemented brain heart infusion broth (BHIS) and lysogeny broth (LB). BHIS was supplemented with the following per liter: hemin 5 mg, menadione 1 mg, yeast extract 5 g, L-cysteine HCl 50 mg, $MgSO_4$ 120 mg, and $CaCl_2$ 50 mg. Each clinical isolate was cultured in liquid LB media at 37 ºC for 16 hours, cells were pelleted by centrifugation and resuspended in molecular grade water. Total DNA was extracted using Qiagen blood and tissue kit (Cat. No. 69504). 16S PCR was performed on each isolate using 27F and 1492R primers and ~1,500 bp amplicons sequenced by Sanger. Sanger sequences were compared to all NCBI using online megablast. The closest hit was to *A. xylosoxidans* for both isolates. In liquid culture, capsule formation was observed, characteristic of many pathogenic *Achromobacter*. A total of 400 nanograms of DNA were used for whole genome sequencing on the Nanopore platform. Whole genome analysis of both isolated determined the strains *Achromobacter xylosoxidans* CF418 and *Achromobacter ruhlandii* CF116.

Phage hunting

Aqueous samples (lake water, pond water, and fountain water) were collected and filtered with a 0.22 µm filter, and stored at 4 °C. Samples of influent from wastewater

treatment plants were stored at 4 ºC, then a 50 ml aliquot was centrifuged at 4000 RPM for 10 minutes to pellet debris and the supernatant was filtered with a 0.45 µM filter. Chloroform was added to 5% v/v for long term storage at 4 ºC.

Phage isolation for *A. ruhlandii* CF116 was performed based on PhagesDB (Russell and Hatfull 2017) protocols with the addition of 10 mM $MgSO_4$ and 5 mM $CaCl_2$ to bacterial cultures and top agar. Plates were incubated at 37 ºC overnight and examined for phage plaques. Individual plaques were streaked onto a top agar plate for phages purification, this procedure was repeated 3 times. After plaque purification, 3.5 ml of phage lysates were prepared by transferring a purified plaque into a growing bacteria culture, which was then incubated overnight at 37 ˚C. To prepare a stock of phage lysate (50 ml), previous 3.5 ml of phage lysate was used (Supplemental Table 5.3)

A. xylosoxidans CF418 from frozen glycerol stocks was streaked onto BHIS plates and incubated at 37 °C for 24 hours in an anaerobic chamber. Individual colonies were then cultured at 37 °C for 24 hours in 3-5 ml BHIS broth. Phages were isolated from influent samples from 4 sewage treatment sites, ponds, fountains, and a lake. Four ml of BHIS top agar, 200 µl host overnight culture, and processed influent (0.1-1 ml) were combined and poured as top agar over BHIS plates. Individual plaques were then be selected by streak-isolation with a toothpick to new top agar plate. The phages were passaged at least 3 times until only one phenotype was visible after last passage with individual plaques present. The phages library was preserved at 4 ºC and in glycerol stocks.

Phages Transmission Electron Microscopy (TEM)

Phages were stained for transmission electron microscopy. Glow-discharged 300-mesh copper grids coated with carbon and formvar were overlaid with drops (30 µl) of purified phage samples for 3 minutes. Salts were removed from the buffer by rinsing the grids 3 times with drops of water (20 µl). Next, the grids were negatively stained with uranyl acetate (0.5 %) for 15 seconds, dried, and examined using a FEI Tecnai T12 TEM (FEI, Hillsboro, OR) at the SDSU Electron Microscopy Facility, operating at 120 kV. Micrographs were taken with an AMT HX41 side mounted digital camera (Advanced Microscopy Technique, Woburn, MA) (Supplementa Figure 5.2).

Phages genome size determination by Pulse Field Gel Electrophoresis (PFGE)

Two hundred and fifty µL of pure phage resuspended in SM buffer were added to an equal volume of 1.6 % low melting (LM) agarose prepared in molecular grade 0.02 µm filtered water. Phage concentration in the starting suspensions are shown in Supplemental Figure 1-C. Before mixing, the LM agarose was placed in 50 ºC water bath for 20 minutes to avoid heat damage to the phage particles. The mix was immediately distributed in individual 75 µL wells of plug molds and allowed to solidify for 20 min at 4 ºC. A small suction bulb was used to pump the plugs out of the molds and place them in TE (10 mM Tris-HCl, 0.1 mM EDTA, pH 7.5) (always 2 ml of solution/3 plugs). Using flat bottom tubes avoid breaking the plugs during the procedure. Free DNA contamination was treated with DNase I by incubating the plugs in a solution containing 1 µg ml^{-1} of DNase and 1X DNase buffer in TE. Incubations were kept at 37 ºC for 1 hour. The liquid was removed, the plugs were transferred to a new tube containing ESP (0.5 M EDTA, pH 9, 1% N-laurylsarcosine, 1 mg ml^{-1} proteinase K), and

then incubated at 50 ºC overnight (ON). In order to inactivate the proteinase K, the plugs were transferred to a new tube containing PMSF solution (1 mM PMSF, 20 mM Tris-HCl, pH 8, 50 mM EDTA). The incubation was carried out for 1 hour at room temperature (RT) on a tube rocker under gentle agitation. The plugs were washed six times with TE, but in the first wash the plugs were transferred to a new tube and left ON at RT under gentle agitation. The five following washes were performed for 30 minutes each and no tube exchange was needed. Using low EDTA TE (10 mM tris and 0.1 mM EDTA), six extra washes of 30 minutes each were performed at the same conditions of the previous ones. After this step, the plugs were maintained in low EDTA TE at 4 ºC until the Pulsed-field agarose gel was prepared. The 0.22 filtered 0.5x TBE was kept in the PFGE machine until the temperature reached 14 ºC. After this step, the plugs were cut in half and placed in the wells of the 1% PFGE Agarose (Bio-rad) in the same filtered TBE. The wells were closed with melted agarose used to make the gel that was kept in a 50 ºC water bath. The gel was left at room temperature for 5 minutes until the agarose polymerized and loaded. The electrophoresis conditions were automatically set by the instrument using the option “auto algorithm” and adding the range of the standard marker (in this case, from 15 Kb to 300 Kb). The gradient selected was 6 V/cm, the time 23:52 h, the included angle 120º, the initial switch time of 1.19 s and the final switch times of 26.29s. The MidRange PFG marker (New England Biolabs) and T4 phage were used as size standards (Supplemental Figure 1-B).

Phages host range determination

Host range of isolated phages was tested in a collection of 22 *Achromobacter* strains isolated from sputum of cystic fibrosis patients at the UCSD CF clinic, and then in the

reference strain *Achromobacter xylosoxidans* HM-235. Host range was tested by spot test of 10 µl of phage lysate in top agar of a lawn of the bacteria. Lysis was evaluated after 16 hours of incubation at 37 ºC.

Phages DNA isolation for sequencing

Fifty mililiters of phage lysate were produced without chloroform treatment to minimize the amount of free bacterial DNA. Phage DNA isolation protocol(Gill, n.d.) (Supplemental Figure 2) consists of phage lysate filtration through a 0.22 µm filter, followed by DNAse and RNAse treatment, PEG precipitation, DNAse and RNAse treatment, proteinase K treatment, and viral particles opening through resin from Promega Wizard DNA clean-up system (Cat. No. A7280). DNA was resuspended in molecular grade water (Supplemental Figure 5.6).

Phages Illumina sequencing

Ten nanograms of phage DNA were used for library preparation using Swift ACCEL-NGS 1S PLUS (Cat. No. 10024) with 16 cycles of PCR amplification and an additional bead cleanup step (AMPure XP, Beckman-Coulter, Cat. No. A63881) with a proportion of 0.85X at the end of library prep to remove sequencing adapters. Libraries were pooled and sequenced in the Illumina platform MiSeq as pair end 300. The number of reads obtained per phage was between 400 and 5 million (Supplemental Table 5.1).

Phages Nanopore sequencing

Three to 400 ng of phage DNA were used for Nanopore sequencing using barcoding and flow cell R9, between 200 to 5000 reads were obtained per phage (Supplemental Table 5.2).

Phages genome assembly

Pair end reads were quality filtered using prinseq++ (Cantu, Sadural, and Edwards 2019) (-lc_entropy=0.5 -trim_qual_right=15 -trim_qual_left=15 -trim_qual_type mean -trim_qual_rule lt -trim_qual_window 2 -min_len 30 -min_qual_mean 20). A subsample with replacement of 50,000 and 100,000 reads per phage was obtained and used for denovo assembly with SPades(Bankevich et al. 2012) (--only-assembler). Attempts to assemble using all reads resulted in fragmented phage genomes. Assembly graphs (.fastg files) were inspected using BANDAGE(Wick et al. 2015), for some phages contigs were merged to obtain the complete genome, in most of the cases a complete genome was obtained in a single contig. Phage lysates propagated in *Achromobacter* CF418 had more than one phage in the assembled contigs, a common phage was identified in lysates LB5, LB7 and LB8 (Supplemental Figure 2) and further identified as a temperate phage in *Achromobacter* CF418, this is *Achromobacter prophage* CF418-P1. Phage genomes were sorted by the terminase gene using circulaline.

Phages genome annotation

Phage genomes were annotated in PATRIC (Wattam et al. 2017) using optimized gene calling for phages, which uses PHANOTATE (McNair et al. 2019). Using this approach most

of the protein annotations were hypothetical. To improve annotations, the phage protein sequences were annotated using the conserved domains database (CDD) (Sayers et al. 2009), HMMER search (EBI, 2019), and ANNs for structural proteins(Cantú, 2019). Expert manual curation of each phage annotation was performed and genome maps were obtained using EasyFig (Sullivan, Petty, and Beatson 2011).

Phages comparative genomics

Available *Achromobacter phage* genomes (n=24) and phage genomes isolated in this study (n=13) were compared to the Phage Proteomic Tree (Rohwer and Edwards 2002) using VIPTree (Nishimura et al. 2017).

High titer phage production and endotoxin removal and quantification

One L of phage lysate ($1x10^9$ PFU ml^{-1}) was produced in LB media supplemented with 10 mM $MgSO_4$ and 5 mM $CaCl_2$. Lysate was centrifuged to remove bacteria debris, filtered twice through a 0.22 µm filtering cup. Concentration was performed in a tangential flow filter (viva flow 200) to 55 ml, followed by filtration through a 0.22 µm filtering cup, phage titer after this step was $1x10^{10}$ PFU ml^{-1}. A buffer exchange from LB to SM was performed in an Amicon filter (100 kDa), concentration volume was 10.6 ml and phage titer $1x10^{11}$ PFU ml^{-1}. Endotoxins (lipopolysaccharides) were removed with incubation of 10% 1-Octanol, then it was removed by centrifugation followed by dialysis, volume after dialysis was 13 ml and phage titer of $1x10^{10}$ PFU ml^{-1} in a solution of 0.15 M NaCl. Lipopolysaccharides were quantified using a colorimetric essay (Biomerieux) and a total of 6500 EU ml^{-1} were present in a 0.15 M NaCl solution with a phage titer of $1x10^{10}$ PFU ml^{-1}.

This procedure is illustrated in Figure 5.6. Phage stock was stored at 4 ºC. This procedure was adapter from previous reports(Bonilla et al. 2016; Szermer-Olearnik and Boratyński 2015).

Acknowledgments

We are grateful to Dr. Dwayne Roach for providing access to lab equipment. We are grateful to Alejandro Vega for providing reagents. Thanks to Dr. Greg Peters and Jacob Vander Griend for fruitful discussions during this research.

Chapter 5 is in preparation for submission. Ana Georgina Cobián Güemes, Tram Le, Maria Isabel Rojas, Lorena Dominguez, Jessica Claire Octavio, Lance Boling, Helena Villela, Shr-Hau Hung, Lili Han, Vito Adrian Cantú, Rob Edwards, Anca Segal, Douglas Conrad and Forest Rohwer; 2019. The dissertation author was the primary investigator and author of this paper.

References

ACLAME. "A CLAssification of Mobile Genetic Elements." Accesed December 2019. http://aclame.ulb.ac.be/.

Amoureux, Lucie, Julien Bador, Eliane Siebor, Nathalie Taillefumier, Annlyse Fanton, and Catherine Neuwirth. 2013. "Epidemiology and Resistance of Achromobacter Xylosoxidans from Cystic Fibrosis Patients in Dijon, Burgundy: First French Data." Journal of Cystic Fibrosis 12 (2): 170–76. https://doi.org/10.1016/j.jcf.2012.08.005.

Bador, Julien, Lucie Amoureux, Emmanuel Blanc, and Catherine Neuwirth. 2013. "Innate Aminoglycoside Resistance of *Achromobacter xylosoxidans* Is." Antimicrobial Agents and Chemotherapy 57 (1): 603–5. https://doi.org/10.1128/AAC.01243-12.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." Journal of Computational Biology 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

BEI. “The Following Reagent Was Obtained through BEI Resources, NIAID, NIH as Part of the Human Microbiome Project: *Achromobacter xylosoxidans*, Strain C54, HM-235.” Accesed December 2019. https://www.beiresources.org/Catalog/bacteria/HM-235.aspx

Bonilla, N, M I Rojas, G Netto Flores Cruz, S H Hung, F Rohwer, and J J Barr. 2016. “Phage on Tap-a Quick and Efficient Protocol for the Preparation of Bacteriophage Laboratory Stocks.” PeerJ 4: e2261. https://doi.org/10.7717/peerj.2261.

Brettin, Thomas, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Gary J. Olsen, Robert Olson, et al. 2015. “RASTtk: A Modular and Extensible Implementation of the RAST Algorithm for Building Custom Annotation Pipelines and Annotating Batches of Genomes.” Scientific Reports 5. https://doi.org/10.1038/srep08365.

Busby, Ben, David M. Kristensen, and Eugene V. Koonin. 2013. “Contribution of Phage-Derived Genomic Islands to the Virulence of Facultative Bacterial Pathogens.” Environmental Microbiology 15 (2): 307–12. https://doi.org/10.1111/j.1462-2920.2012.02886.x.

Cantu, Adrian, Jeffrey Sadural, and Robert Edwards. 2019. “PRINSEQ++, a Multi-Threaded Tool for Fast and Efficient Quality Control and Preprocessing of Sequencing Datasets.” PeerJ Preprints, 43–45. https://doi.org/10.7287/peerj.preprints.27553.

Cantú, Vito Adrian. 2019. “Phanns.” https://edwards.sdsu.edu/phannies/upload.

Chen, Lihong, Dandan Zheng, Bo Liu, Jian Yang, and Qi Jin. 2016. “VFDB 2016: Hierarchical and Refined Dataset for Big Data Analysis - 10 Years On.” Nucleic Acids Research 44 (D1): D694–97. https://doi.org/10.1093/nar/gkv1239.

Danielsson, Olle, Silvia Atrian, Teresa Luque, Lars Hjelmqvist, Roser Gonzalez-Duarte, and Hans Jornvall. 1994. “Fundamental Molecular Differences Between Alcohol Dehydrogenase Classes.” PNAS 91 (11): 4980–84.

EBI. “HMMER Search.” Accesed December 2019. https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan.

Gill, Jason J. n.d. “Gill: Phage Genomic DNA Extraction.” https://openwetware.org/wiki/Gill:Phage_genomic_DNA_extraction.

Gordillo Altamirano, Fernandi, and Jeremy Barr. 2019. “Phage Therapy in the Postantibiotic Era.” Clinical Microbiology Reviews 32 (2): 1–25.

Hoyle, N, P Zhvaniya, N Balarjishvili, D Bolkvadze, L Nadareishvili, D Nizharadze, J Wittmann, C Rohde, and M Kutateladze. 2018. “Phage Therapy against

Achromobacter Xylosoxidans Lung Infection in a Patient with Cystic Fibrosis: A Case Report." Research in Microbiology. https://doi.org/10.1016/j.resmic.2018.05.001.

Huh, Haein, Shirley Wong, Jesse St. Jean, and Roderick Slavcev. 2019. "Bacteriophage Interactions with Mammalian Tissue: Therapeutic Applications." Advanced Drug Delivery Reviews. https://doi.org/10.1016/j.addr.2019.01.003.

Jakobsen, Tim Holm, Martin Asser Hansen, Peter Østrup Jensen, Lars Hansen, Leise Riber, Mette Kolpen, Christine Rønne Hansen, et al. 2013. "Complete Genome Sequence of the Cystic Fibrosis Pathogen Achromobacter Xylosoxidans NH44784-1996 Complies with Important Pathogenic Phenotypes" 8 (7): 8–11. https://doi.org/10.1371/journal.pone.0068484.

Jelski, Wojciech, Ã Lech Chrostek, Wlodzimierz Markiewicz, and Maciej Szmitkowski. 2006. "Activity of Alcohol Dehydrogenase (ADH) Isoenzymes and Aldehyde Dehydrogenase (ALDH) in the Sera of Patients With Breast Cancer." Cancer 108 (December 2005): 105–8. https://doi.org/10.1002/jcla.

Kortright, Kaitlyn E, Benjamin K Chan, Jonathan L Koff, and Paul E Turner. 2019. "Phage Therapy : A Renewed Approach to Combat Antibiotic-Resistant Bacteria." Cell Host and Microbe 25 (2): 219–32. https://doi.org/10.1016/j.chom.2019.01.014.

Kutateladze, M., and R. Adamia. 2008. "Phage Therapy Experience at the Eliava Institute." Medecine et Maladies Infectieuses 38 (8): 426–30. https://doi.org/10.1016/j.medmal.2008.06.023.

Leibniz Institute. "DSMZ." Accesed December 2019. https://www.dsmz.de/.

Lemire, Sebastien, Kevin M. Yehl, and Timothy K. Lu. 2018. "Phage-Based Applications in Synthetic Biology." Annual Review of Virology 5 (1): annurev-virology-092917-043544. https://doi.org/10.1146/annurev-virology-092917-043544.

Li, Erna, Zhe Yin, Yanyan Ma, Huan Li, Weishi Lin, Xiao Wei, and Ruixiang Zhao. 2016. "Identification and Molecular Characterization of Bacteriophage PhiAxp-2 of Achromobacter Xylosoxidans." Nature Publishing Group, no. September: 1–11. https://doi.org/10.1038/srep34300.

Li, Erna, Jiangtao Zhao, Yanyan Ma, Xiao Wei, Huan Li, Weishi Lin, Xuesong Wang, et al. 2016. "Characterization of a Novel Achromobacter Xylosoxidans Specific Siphoviruse: PhiAxp-1." Scientific Reports 6: 21943. https://doi.org/10.1038/srep21943.

Little, Mark, Maria Isabel Rojas, and Forest Rohwer. 2020. "Bacteriophage Can Drive Virulence in Marine Pathogens." In Marine Disease Ecology. Oxford University Press.

Ma, Yanyan, Erna Li, Zhizhen Qi, Huan Li, Xiao Wei, Weishi Lin, Ruixiang Zhao, et al. 2016. “Isolation and Molecular Characterisation of Achromobacter Phage PhiAxp-3, an N4-like Bacteriophage.” Scientific Reports 6 (1): 24776. https://doi.org/10.1038/srep24776.

Malik, Danish J., Ilya J. Sokolov, Gurinder K. Vinner, Francesco Mancuso, Salvatore Cinquerrui, Goran T. Vladisavljevic, Martha R.J. Clokie, Natalie J. Garton, Andrew G.F. Stapley, and Anna Kirpichnikova. 2017. “Formulation, Stabilisation and Encapsulation of Bacteriophage for Phage Therapy.” Advances in Colloid and Interface Science 249 (May): 100–133. https://doi.org/10.1016/j.cis.2017.05.014.

Manckoundia, Patrick, Emmanuel Mazen, Alexis Saloff Coste, Sophie Somana, Sophie Marilier, Jean Marie Duez, Agnès Camus, Laura Popitean, Julien Bador, and Pierre Pfitzenmeyer. 2011. “A Case of Meningitis Due to Achromobacter Xylosoxidans Denitrificans 60 Years after a Cranial Trauma.” Medical Science Monitor 17 (6): 2010–12. https://doi.org/10.12659/MSM.881796.

McCord, James P., David C. Muddiman, and Morteza G. Khaledi. 2017. “Perfluorinated Alcohol Induced Coacervates as Extraction Media for Proteomic Analysis.” Journal of Chromatography A 1523: 293–99. https://doi.org/10.1016/j.chroma.2017.06.025.

McNair, Katelyn, Barbara A. Bailey, and Robert A. Edwards. 2012. “PHACTS, a Computational Approach to Classifying the Lifestyle of Phages.” Bioinformatics 28 (5): 614–18. https://doi.org/10.1093/bioinformatics/bts014.

McNair, Katelyn, Carol Zhou, Elizabeth A Dinsdale, Brian Souza, and Robert A Edwards. 2019. “PHANOTATE: A Novel Approach to Gene Identification in Phage Genomes.” Bioinformatics, no. April: 1–6. https://doi.org/10.1093/bioinformatics/btz265.

Morrison, David, and Loretta Leive. 1976. “Fractions of Lipopolysaccharide from Escherichia Coli O111:B4 Prepared by Two Extractopn Procedures.” The Journal of Biological Chemistry 2750 (8): 2911–19.

Nishimura, Yosuke, Takashi Yoshida, Megumi Kuronishi, Hideya Uehara, Hiroyuki Ogata, and Susumu Goto. 2017. “Genome Analysis ViPTree : The Viral Proteomic Tree Server” 33 (March): 2379–80. https://doi.org/10.1093/bioinformatics/btx157.

Ofir, Gal, and Rotem Sorek. 2018. “Review Contemporary Phage Biology : From Classic Models to New Insights.” Cell 172 (6): 1260–70. https://doi.org/10.1016/j.cell.2017.10.045.

Park, Jung Hyun, Nang Hee Song, and Jae Woong Koh. 2012. “Achromobacter Xylosoxidans Keratitis after Contact Lens Usage.” Korean Journal of Ophthalmology : KJO 26 (1): 49–53. https://doi.org/10.3341/kjo.2012.26.1.49.

Philipson, Casandra W, Logan J Voegtly, Matthew R Lueder, Kyle A Long, Gregory K Rice, Kenneth G Frey, Biswajit Biswas, Regina Z Cer, Theron Hamilton, and Kimberly A Bishop-lilly. 2018. “Characterizing Phage Genomes for Therapeutic Applications,” 1–20. https://doi.org/10.3390/v10040188.

Porayath, Chandni, Amrita Salim, Archana Palillam Veedu, Pradeesh Babu, Bipin Nair, Ajith Madhavan, and Sanjay Pal. 2018. “Characterization of the Bacteriophages Binding to Human Matrix Molecules.” International Journal of Biological Macromolecules 110: 608–15. https://doi.org/10.1016/j.ijbiomac.2017.12.052.

Ridderberg, Winnie, Signe Maria Nielsen, and Niels Nørskov-Lauritsen. 2015. “Genetic Adaptation of Achromobacter Sp. during Persistence in the Lungs of Cystic Fibrosis Patients.” PLoS ONE 10 (8): 1–14. https://doi.org/10.1371/journal.pone.0136790.

Rohde, Manfred, Manfred Nimtz, and Johannes Wittmann. 2017. “Characterization and Genome Comparisons of Three Achromobacter Phages of the Family Siphoviridae,” 2191–2201. https://doi.org/10.1007/s00705-017-3347-8.

Rohwer, Forest, and Rob Edwards. 2002. “The Phage Proteomic Tree : A Genome-Based Taxonomy for Phage” 184 (16): 4529–35. https://doi.org/10.1128/JB.184.16.4529.

Russell, Daniel A, and Graham F Hatfull. 2017. “PhagesDB : The Actinobacteriophage Database” 33 (December 2016): 784–86. https://doi.org/10.1093/bioinformatics/btw711.

Satre, Michael A., Mirna Žgombić-Knight, and Gregg Duester. 1994. “The Complete Structure of Human Class IV Alcohol Dehydrogenase (Retinol Dehydrogenase) Determined from the ADH7 Gene.” Journal of Biological Chemistry 269 (22): 15606–12.

Sayers, Eric W., Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, et al. 2009. “Database Resources of the National Center for Biotechnology Information.” Nucleic Acids Research 38 (SUPPL.1): 5–16. https://doi.org/10.1093/nar/gkp967.

Sayers, Samantha, Li Li, Edison Ong, Shunzhou Deng, Guanghua Fu, Yu Lin, Brian Yang, et al. 2019. “Victors: A Web-Based Knowledge Base of Virulence Factors in Human and Animal Pathogens.” Nucleic Acids Research 47 (D1): D693–700. https://doi.org/10.1093/nar/gky999.

Scholl, Dean. 2017. “Phage Tail – Like Bacteriocins.” Annual Review of Virology, no. 4: 453–67.

Shan, Jinyu, Ananthi Ramachandran, Anisha M Thanki, Fatima B I Vukusic, and Jakub Barylski. 2018. “Bacteriophages Are More Virulent to Bacteria with Human Cells

than They Are in Bacterial Culture ; Insights from HT-29 Cells." Scientific Reports, no. July 2017: 1–8. https://doi.org/10.1038/s41598-018-23418-y.

Sullivan, Mitchell J, Nicola K Petty, and Scott A Beatson. 2011. "Easyfig : A Genome Comparison Visualizer" Bioinformatics 27 (7): 1009–10.

Sun, S., S. Gao, K. Kondabagil, Y. Xiang, M. G. Rossmann, and V. B. Rao. 2012. "Structure and Function of the Small Terminase Component of the DNA Packaging Machine in T4-like Bacteriophages." Proceedings of the National Academy of Sciences 109 (3): 817–22. https://doi.org/10.1073/pnas.1110224109.

Szermer-Olearnik, Bozena, and Janusz Boratyński. 2015. "Removal of Endotoxins from Bacteriophage Preparations by Extraction with Organic Solvents." PLoS ONE 10 (3): 1–10. https://doi.org/10.1371/journal.pone.0122672.

Tena, Daniel, Alejandro González-Praetorius, Mercedes Pérez-Balsalobre, Oliva Sancho, and Julia Bisquert. 2008. "Urinary Tract Infection Due to Achromobacter Xylosoxidans: Report of 9 Cases." Scandinavian Journal of Infectious Diseases 40 (2): 84–87. https://doi.org/10.1080/00365540701558714.

Thomson, J. A., and D. R. Woods. 1974. "Bacteriophages and Cryptic Lysogeny in Achromobacter." Journal of General Virology 22 (1): 153–57.

Tokuyasu, Hirokazu, Takehito Fukushima, Hirofumi Nakazaki, and Eiji Shimizu. 2012. "Infective Endocarditis Caused by Achromobacter Xylosoxidans: A Case Report and Review of the Literature." Internal Medicine 51 (9): 1133–38. https://doi.org/10.2169/internalmedicine.51.6930.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. "Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center." Nucleic Acids Research 45 (D1): D535–42. https://doi.org/10.1093/nar/gkw1017.

Wick, Ryan R, Mark B Schultz, Justin Zobel, and Kathryn E Holt. 2015. "Genome Analysis Bandage : Interactive Visualization of de Novo Genome Assemblies" 31 (June): 3350–52. https://doi.org/10.1093/bioinformatics/btv383.

Wittmann, Johannes, Brigitte Dreiseikelmann, Christine Rohde, Manfred Rohde, and Johannes Sikorski. 2014. "Isolation and Characterization of Numerous Novel Phages Targeting Diverse Strains of the Ubiquitous and Opportunistic Pathogen Achromobacter Xylosoxidans." PLoS ONE 9 (1). https://doi.org/10.1371/journal.pone.0086935.

Wittmann, Johannes, Brigitte Dreiseikelmann, Manfred Rohde, Jan P Meier-Kolthoff, Boyke Bunk, and Christine Rohde. 2014. "First Genome Sequences of

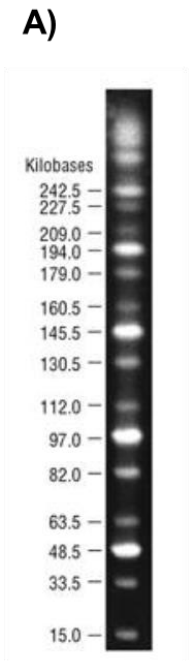
Achromobacter Phages Reveal New Members of the N4 Family." Virology Journal 11: 14. https://doi.org/10.1186/1743-422X-11-14.

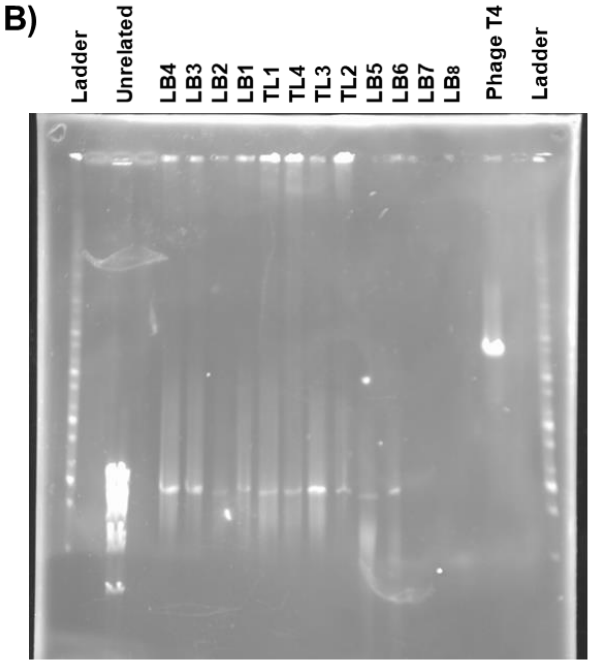
Woods, D. R., and J. A. Thomson. 1975. "Unstable Generalized Transduction in Achromobacter." Journal of General Microbiology 88 (1): 86–92. https://doi.org/10.1099/00221287-88-1-86.

Young, Ry, and Jason J Gill. 2015. "Phage Therapy Redux--What Is to Be Done?" Science (New York, N.Y.) 350 (6265): 1163–64. https://doi.org/10.1126/science.aad6791.

Appendix for Chapter 5

Supplemental figures

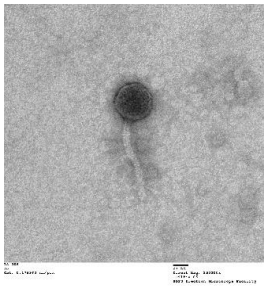




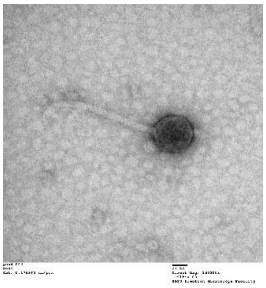
C)

Phage ID	Phage Concentration Pfu/mL
LB8	2.05×10^{4}
LB7	2.23×10^{5}
LB6	5.30×10^{9}
LB5	3.70×10^{9}
TL2	2.45×10^{9}
TL3	2.45×10^{9}
TL4	4.30×10^{9}
TL1	2.35×10^{8}
LB1	1.00×10^{12}
LB2	2.25×10^{9}
LB3	2.25×10^{10}
LB4	2.25×10^{10}

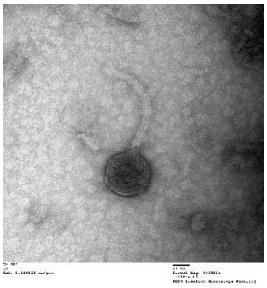
Supplemental Figure 5.1 Phages genome size by Pulse Field Gel Electrophoresis. A) Molecular size ladder B) Pulse Filed Gel Electrophoresis. C) Phages concentration used for Pulse Field Gel Electrophoresis plugs.



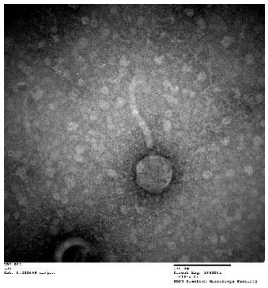
Achromobacter phage nyaak



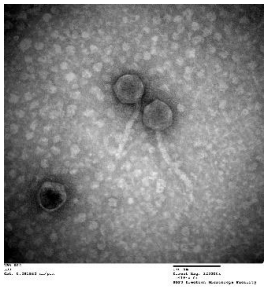
Achromobacter phage kewaak



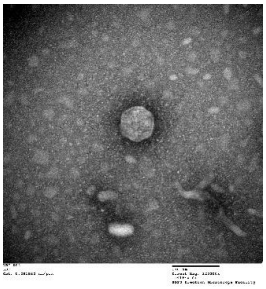
Achromobacter phage wiik



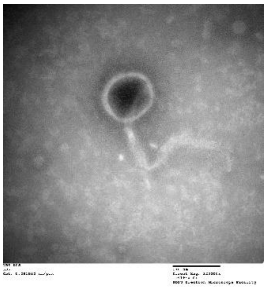
Achromobacter phage tuull



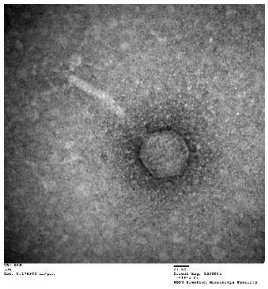
Achromobacter phage xasilly



Achromobacter phage emuu

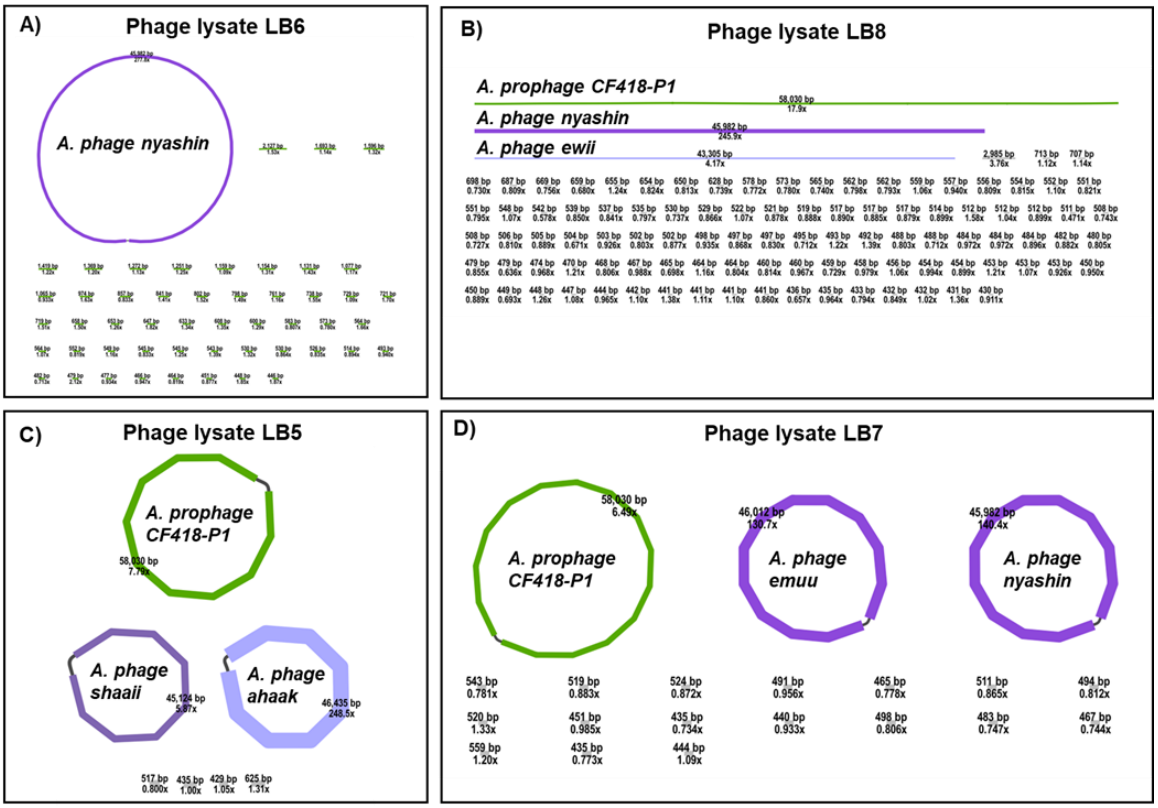


Achromobacter phage nyashin

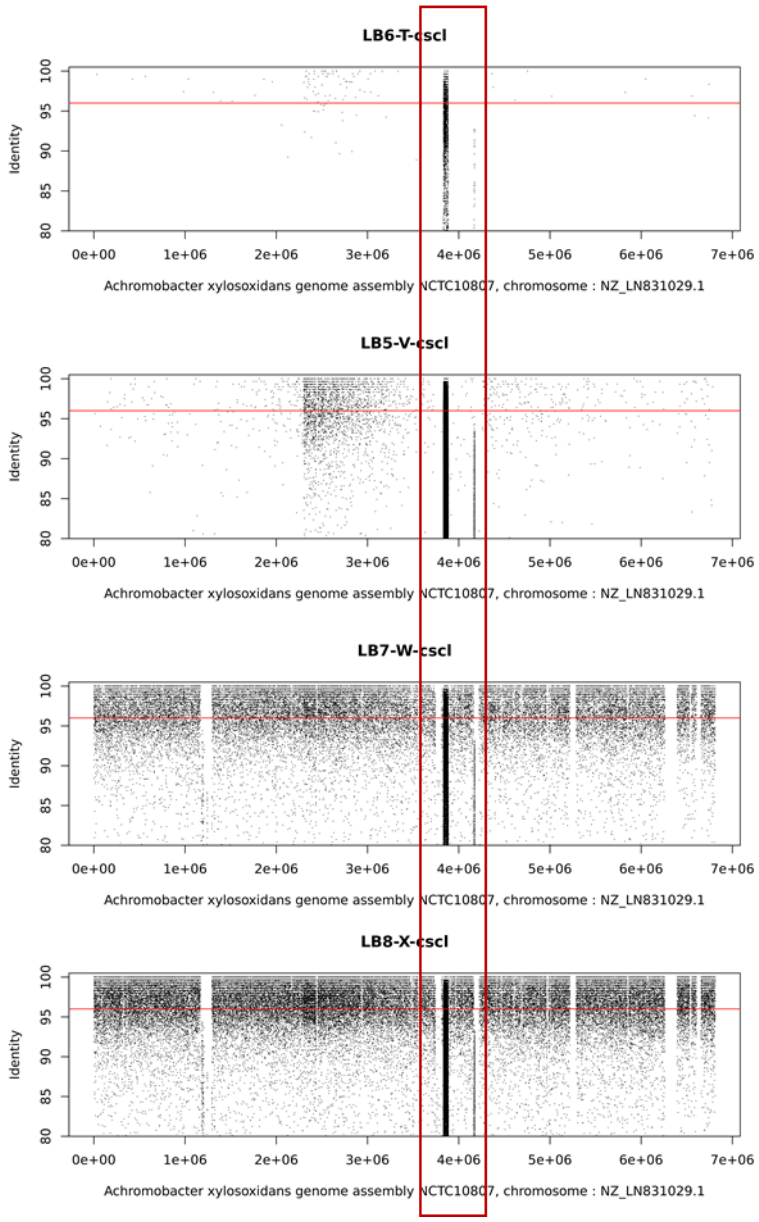


Achromobacter phage kwar

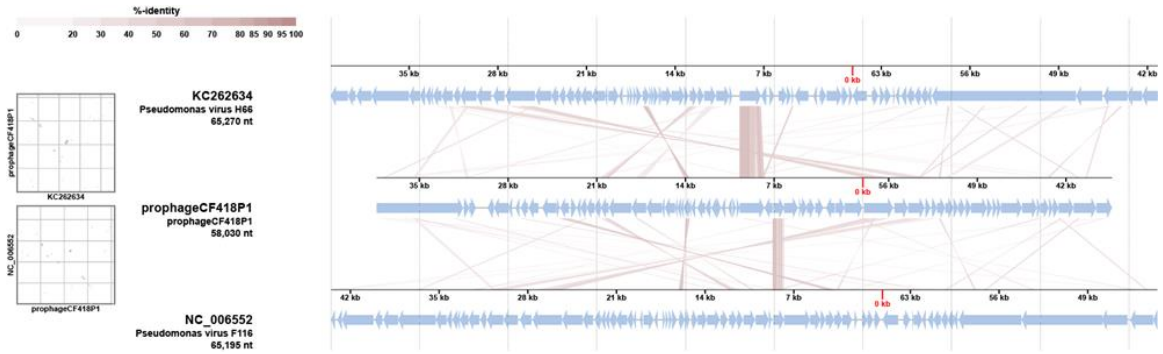
Supplemental Figure 5.2 Achromobacter phages transmission electron microscopy. *Achromobacter phages ewii* and *Achromobacter phage* maay were not imaged. *Achromobacter phage ahaak* and *Achromobacter phage shaii* were in a mixed lysate, imaging is not shown. Names etymology from the kumeyaai language spoken by native tribes of San Diego, CA.



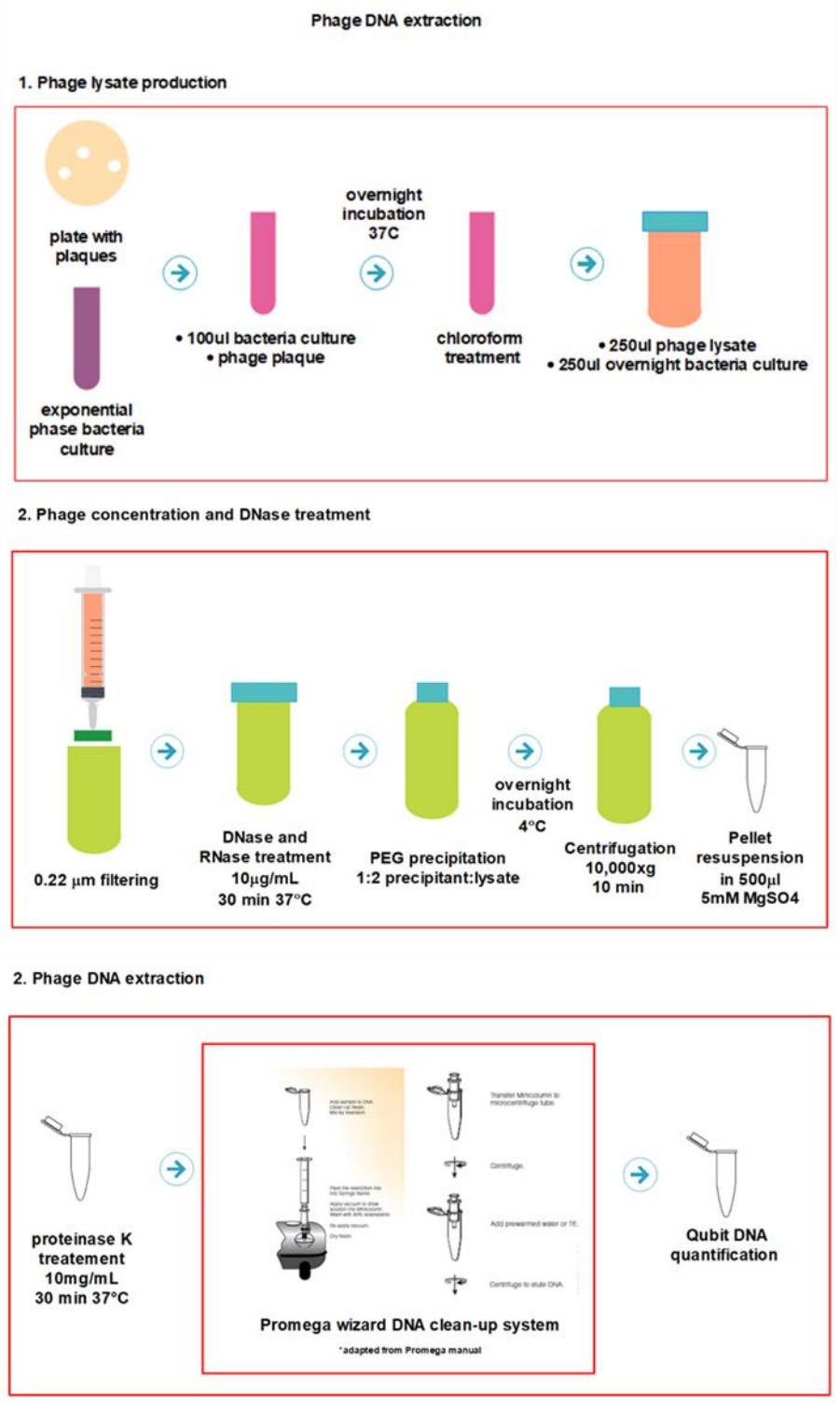
Supplemental Figure 5.3 Prophage induced in Achromobacter CF418 when infected with additional lytic phages. Contigs assembled from each phage lysate using SPADES were visualized in BANDAGE. A) Assembly with 100,000 reads. One bacteriophage genome present. B) Assembly with 100,000 reads. Two bacteriophage genomes present, one with high coverage of 250X (light purple) and another one with low coverage 5X (dark purple). One Achromobacter element present (green) C) Assembly with 100,000 reads. Two bacteriophage genomes present, one with high coverage of 250X (light purple) and another one with low coverage 5X (dark purple). One Achromobacter element present (green) D) Assembly with 100,000 reads. Two bacteriophage genomes present with similar coverage (dark purple). Achromobacter element present (green).



Supplemental Figure 5.4 Prophage induced in Achromobacter CF418 when infected with additional lytic phages. Genome coverage of bacteria reference genome. Region in red box is the prophage.



Supplemental Figure 5.5 Achromobacter prophage CF418-P1 and its closest phage relatives, genomes comparisons.



Supplemental Figure 5.6 Phages propagation and DNA extraction for genome sequencing.

Supplemental tables

Supplemental Table 5.1 Achromobacter phages Illumina sequencing information.

Phage lysate	**Illumina good quality reads**	***Hits to A. ruhlandii* NZ_CP017433.1**	***Hits to A. xylosoxidans* NZ_LN831029.1**	**% *Achromobacter* reads**	**Hits to human genome**	**% Human genome reads**
TL2-B-filtered	1,894,772	4	1	0.0003	3	0.0000
TL3-C-filtered	1,749,597	3	2	0.0003	0	0.0000
LB1-D-filtered	2,537,881	12	6	0.0007	1	0.0000
LB2-E-filtered	705	0	4	0.5674	0	0.0000
LB3-F-filtered	2,274,283	14	12	0.0011	0	0.0000
LB4-G-filtered	406	0	0	0.0000	0	0.0000
TL1-M-cscl	3,406,503	73	12	0.0025	10	0.0001
TL4-U-cscl	5,142,519	315	105	0.0082	35	0.0001
LB5-V-cscl	4,040,869	207	112,219	2.7822	0	0.0000
LB6-T-cscl	1,874,759	4	3,970	0.2120	0	0.0000
LB7-W-cscl	3,895,260	1,826	102,389	2.6754	145	0.0002
LB8-X-cscl	2,473,235	3,115	183,814	7.5581	23	0.0001

Supplemental Table 5.2 Achromobacter phages Nanopore sequencing information.

Sample type	Sample ID	DNA amount (nanograms)	BARCODE	Reads	% total reads
Phage	TL1	397.50	RB01	772	0.84
Phage	TL2	400.00	RB02	1867	2.04
Phage	TL3	400.00	RB03	1642	1.79
Phage	TL4	203.25	RB04	2700	2.95
Phage	LB4	60.75	RB05	5023	5.49
Phage	LB3	18.00	RB06	1127	1.23
Phage	LB7	10.50	RB09	585	0.64
Phage	LB8	3.26	RB10	231	0.25
Bacteria	Achromo CF418	213.75	RB12	21470	23.47
			Unclassified	35399	38.70
			Total	91482	

Supplemental Table 5.3 Phage lysates and phage genomes names

Isolation ID	Sequencing ID	Phage contig ID	Phage name	Host	Phage isolation source
IS	TL1	*Achromobacter phage nyaak* TL1	*Achromobacter phage nyaak*	*A. ruhlandii* CF116	Influent water sample
pond	TL2	*Achromobacter phage kewaak* TL2	*Achromobacter phage kewaak*	*A. ruhlandii* CF116	SDSU fishpond water
LM	TL4	*Achromobacter phage wiik* TL4	*Achromobacter phage wiik*	*A. ruhlandii* CF116	Lake Murray
IS2	LB2	Achromobacter phage *tuull* LB2	Achromobacter phage *tuull*	*A. ruhlandii* CF116	Influent water site 2
IS1	LB1	Achromobacter phage maay LB1	Achromobacter phage maay	*A. ruhlandii* CF116	Influent water site 1
IS3	LB3	Achromobacter phage xasilly LB3	Achromobacter phage xasilly	*A. ruhlandii* CF116	Influent water site 3
SA2	LB5	Achromobacter phage LB5-A	Achromobacter phage ahaak	*A.ruhlandii* CF418	Infuent water
SA3	LB5	Achromobacter phage LB5-B	Achromobacter phage shaaii	*A.ruhlandii* CF418	Infuent water
S315S	LB7	Achromobacter phage LB7-A	Achromobacter phage emuu	*A.ruhlandii* CF418	Influent water
S313L	LB8	Achromobacter phage LB8-B	Achromobacter phage ewii	*A.ruhlandii* CF418	Influent water
S2D	LB6-repeated	Achromobacter phage nyashin LB6 (repeated in 3 samples)	Achromobacter phage nyashin	*A.ruhlandii* CF418	Infuent water
S313L	LB8-prophage	Achromobacter prophage CF418-P1 (prophage from CF418)	Achromobacter prophage CF418-P1	*A.ruhlandii* CF418	Influent water
IS4	LB4	Achromobacter phage LB4 (partial genome ?)	Achromobacter phage kwar	*A. ruhlandii* CF116	Influent water site 4

Supplemental Table 5.4 Lifestyle prediction for Achromobacter phages.

Phage name	Lifestyle	probability of temperate lifestyle	sd for temperate lifestyle	probability of lytic lifestyle	sd for lytic lifestyle	Integrase in genome
Achromobacter phage nyaak		0.514	0.041	0.486	0.041	0
Achromobacter phage kewaak		0.512	0.040	0.488	0.040	0
Achromobacter phage wiik		0.509	0.045	0.491	0.045	0
Achromobacter phage *tuull*		0.504	0.043	0.496	0.043	0
Achromobacter phage maay		0.517	0.042	0.483	0.042	0
Achromobacter phage xasilly		0.509	0.041	0.491	0.041	0
Achromobacter phage ahaak		0.486	0.043	0.514	0.043	0
Achromobacter phage shaaii	Lytic	0.468	0.039	**0.532**	**0.039**	0
Achromobacter phage emuu		0.482	0.039	0.518	0.039	0
Achromobacter phage ewii		0.505	0.043	0.495	0.043	0
Achromobacter phage nyashin	Lytic	0.461	0.039	**0.539**	**0.039**	0
Achromobacter prophage CF418-P1	Temperate	**0.526**	**0.040**	0.474	0.040	**1**
Achromobacter phage kwar		0.508	0.037	0.492	0.037	0

Supplemental Table 5.5 Isolated Achromobacter phages for broader host range.

Phage ID	Phage name	Phage isolation host	Phage isolation source	Phage isolation source collection date
M1	*Achromobacter phage M1*	*Achromobacter* sp. VVP0357	Vallecitos Wastewater District. (Meadowlark Water Reclamation Facility), Carlsbad, CA.	2/20/2019
M2	*Achromobacter phage M2*	*Achromobacter* sp. VVP0357	Vallecitos Wastewater District. (Meadowlark Water Reclamation Facility), Carlsbad, CA.	2/20/2019
ENA1	*Achromobacter phage ENA1*	*Achromobacter* sp. VVP0357	Encina Wastewater Authority, Carlsbad CA. 92011	2/20/2019
MW2	*Achromobacter phage MW2*	*Achromobacter* sp. VVP0426	Vallecitos Wastewater District. (Meadowlark Water Reclamation Facility), Carlsbad, CA.	2/20/2019
SE2	*Achromobacter phage SE2*	*Achromobacter* sp. VVP0426	San Elijo Joint Powers Authority, Cardiff by the Sea, CA.	4/16/2019

Chapter 6 : Synthesis

In this dissertation the role of phages in an environmental and host-associated context was explored. Phages were studied at different scales, from a global environmental context (Chapter 1) to specific interactions in host-associated systems (Chapter 3 and Chapter 4). In addition, accurate methods to describe the effects of phages in microbiomes were developed (Chapter 2).

Phages abundance, diversity and lifestyle.

The most common mechanism by which phages influence microbial ecosystems is by controlling bacteria populations and therefore maintaining diversity. Phages exist as free viral particles (lytic phages) or inside bacteria genomes (lysogenic phages) and they can alternate between these two lifestyles. In host associated systems a third player is involved, the eukaryotic immune system. This tripartite association of host-bacteria-phage results in a more complex system.

Most phage particles on Earth are in soils and sediments (97%), these biomes have a high bacteria abundance and a virus to microbe ratio smaller than 10, which is indicative of a phage lysogenic lifestyle (Figure 6.1-A). The viral genetic diversity in soils and sediments is still underexplored. Since the publication of "Viruses as winners in the game of life" (Cobián Güemes et al. 2016), 35 new soils and sediment viromes have been published (Yoshida 2018; Bryson et al. 2015; Adriaenssens et al. 2017; Yu et al. 2019; Emerson et al. 2018; Trubl et al. 2018; Han et al. 2017; Scola et al. 2018; Corinaldesi, Tangherlini, and Dell'Anno 2017).

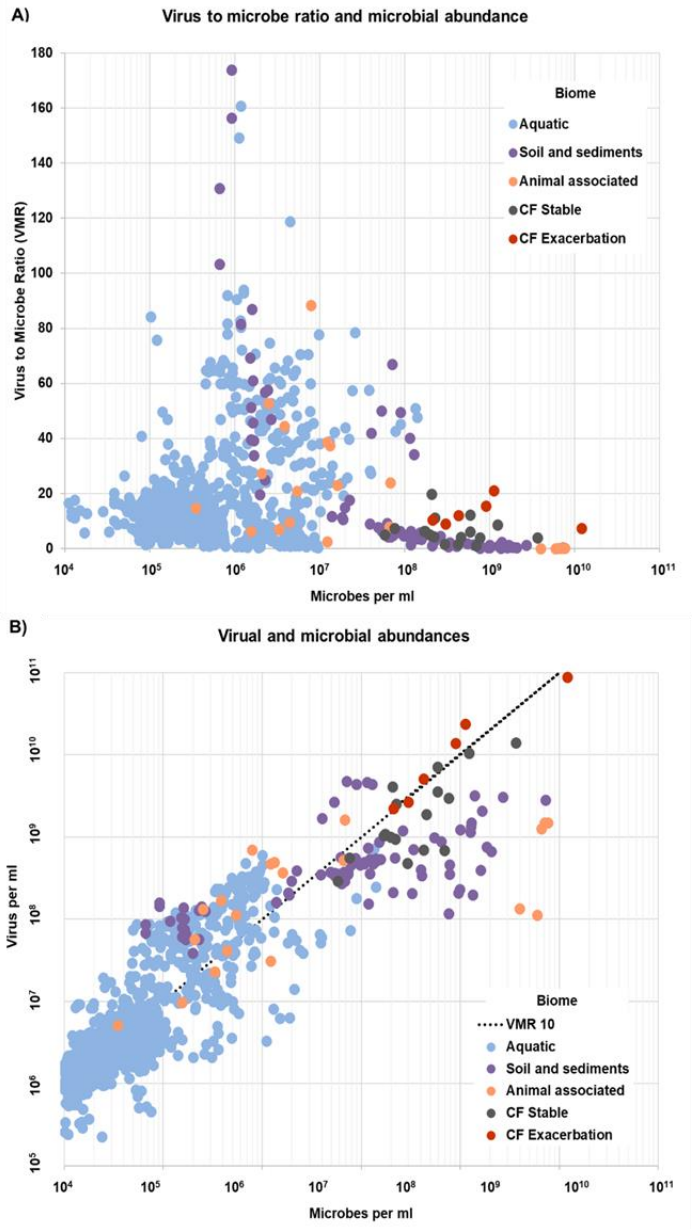


Figure 6.1Virus to microbe ratio, microbial and viral abundance in Earth biomes and in Cystic Fibrosis (CF) stable and exacerbation respiratory tract samples. Aquatic, soil and sediments, and animal associated data points were compiled from the literature and published in Cobián 2016 and Knowles and Silveira 2016. Cystic Fibrosis data points were obtained from Chapter 4 of this dissertation. A) Microbes per milliliter and the virus to microbe ratio. B) Microbial and viral abundances. A virus to microbe ratio of 10 is presented for reference.

Host-associated environments present high bacteria abundances. The lungs of Cystic Fibrosis (CF) patients have between 10^8 and 10^{10} microbes per mililiter. In such high bacteria abundances, viral abundances do not scale linearly. This phenomenon is an indication of phages integrating in bacteria genomes, a mechanism described as Piggyback the winner (Knowles and Silveira et al. 2016). In periods of stability in the lungs of CF patients, a phage lysogenic lifestyle is predominant; in exacerbations, a phage lytic lifestyle is observed (Figure 6.1-B). This finding suggests that phage inductions increase in exacerbation periods, an observation explored in Chapter 4. We predict that phage-pathogen intereactions are contributing to pathogenesis in the CF lung. Thus, characterization of bacterial and viral communities is needed to design effective therapeutic strategies during exacerbation events.

Are phages hard to find?

Phage identification in metagenomes and viromes is challenging since most phages are not characterized yet. In the exploration of the global virome, around 200,000 viral genotypes were identified, most of which represent new phages. To identify phages and mobile elements in CF metagenome assembled genomes (MAGs) an exact match strategy was developed (Chapter 2) and new phages were identified in every CF patient. This approach highlights the importance of accurate and detailed metagenomic studies. In this dissertation, methods to detect mobile elements in MAGs were developed and applied to successfully detect new phages and multispecies mobile elements.

The personalized nature of Cystic Fibrosis exacerbations

The CF lung is a harsh environment for microbes, nevertheless they colonize the mucosal surfaces persistently over the patient's lifetime. During this time the microbial

community adapts to the CF lung through constant exposure to antibiotics and host immune responses. Thus, microbial species that inhabit the lungs of each CF patient have unique adaptations. In the case of *Stenotrophomonas maltophilia*, the same species dominated exacerbation periods of two CF patients; however, each *S. maltophilia* carried unique insertions that represented 3% of the genome. One of these insertions was a filamentous phage encoding zonula occludens toxin. This finding highlights the importance of accurate and detailed metagenomic analysis for the study of the CF lung and other polymicrobial infections. In Chapters 3 and 4 of this dissertation, I provided the methods to detect mobile elements and illustrated their importance in understanding their effect in host-pathogen-phage interactions.

Model for Cystic Fibrosis acute exacerbations

CF acute exacerbations in which the lung function declines rapidly and there is no improvement with treatment are identified by the clinicians as a danger signal for CF patients. To improve survival rates of these patients, a CF rapid response strategy was adopted. Lessons learned from these acute exacerbations include: 1) the microbial community exhibits low diversity, 2) the dominant microbes employ mechanisms to cause direct damage to the host epithelium and access oxygen and nutrients, 3) the dominant microbes appear to replicate rapidly, 4) these infections are not recurrent, and 5) the phages lytic lifestyle dominates the system. Examples of CF microbes associated with acute exacerbations that follow the described principles are: *Achromobacter* spp. which encode hemolysins that attack the host tissue in order to access nutrients and oxygen; *Stenotrophomonas* spp. carrying a filamentous phage encoding zonula occludens toxin; *Escherichia coli* carrying shigatoxin; as well as

Haemophilus spp., *Staphylococccus* spp. and *Streptococcus* spp. carrying several virulence factors (Figure 6.2).

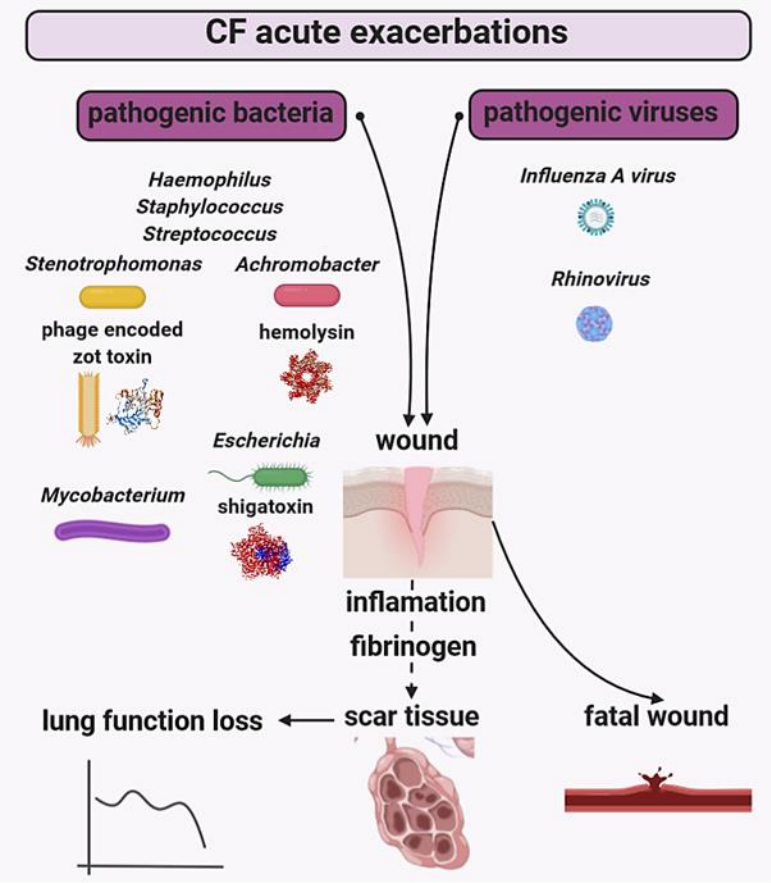


Figure 6.2 Model for Cystic Fibrosis acute exacerbations.

The Chaotic Neutral Phage

Phages are the winners in the game of life since they are the most abundant and diverse life form on Earth (Hendrix et al. 1999; Breitbart et al. 2002; Rohwer and Edwards 2002; Edwards and Rohwer 2005; Cobián Güemes et al. 2016). How they maintain a total of 6.03×10^{31} particles in the biosphere is still an open question, however we have some hints about their dominance strategies: 1) phages switch between a lytic and lysogenic lifestyle

which allow them to survive as free viral particles or in a latent state inside bacteria genomes (Ptashne 2004; Knowles et al. 2016; Zeng et al. 2010; Erez et al. 2017; Broussard et al. 2013); 2) phages encode their genomic information in different kinds of genetic material, such as dsDNA, ssDNA, dsRNA and ssRNA (Koonin, Dolja, and Krupovic 2015; Ofir and Sorek 2018), a feature that expands the genomic space they can explore and; 3) phages attack bacteria which leads to an arms race between them and allows for the development of multiple evasion and infection strategies (Benler et al. 2018, Lee et al. 2018, Rauch et al. 2017). The diverse strategies phages employ seem chaotic (Cook, Tweet, and Williams 2003) in the sense that they do not follow a set of established rules, nor do we understand the extent of these rules yet. Phages seem to be neutral in the sense that their effect to their host is sometimes beneficial and sometimes detrimental, but they do not have an agenda (i.e. evolution is not intentional). The question of the nature of phages is still open, maybe they are lawful evil, and do abide by a strict set of rules to dominate the world.

References

Adriaenssens, Evelien M., Rolf Kramer, Marc W. van Goethem, Thulani P. Makhalanyane, Ian Hogg, and Don A. Cowan. 2017. “Environmental Drivers of Viral Community Composition in Antarctic Soils Identified by Viromics.” Microbiome 5 (1): 1–14. https://doi.org/10.1186/s40168-017-0301-7.

Benler, Sean, Ana Georgina Cobián-güemes, Katelyn Mcnair, Shr-hau Hung, Kyle Levi, Rob Edwards, and Forest Rohwer. 2018. “A Diversity-Generating Retroelement Encoded by a Globally Ubiquitous Bacteroides Phage.” Microbiome 6 (191): 1–10.

Breitbart, Mya, Peter Salamon, Bjarne Andresen, Joseph M Mahaffy, Anca M Segall, David Mead, Farooq Azam, and Forest Rohwer. 2002. “Genomic Analysis of Uncultured Marine Viral Communities.” Proceedings of the National Academy of Sciences of the United States of America 99 (22): 14250–55. https://doi.org/10.1073/pnas.202488399.

Broussard, Gregory W., Lauren M. Oldfield, Valerie M. Villanueva, Bryce L. Lunt, Emilee E. Shine, and Graham F. Hatfull. 2013. “Integration-Dependent Bacteriophage Immunity Provides Insights into the Evolution of Genetic Switches.” Molecular Cell 49 (2): 237–48. https://doi.org/10.1016/j.molcel.2012.11.012.

Bryson, Samuel J.oseph, Andrew R. Thurber, Adrienne M.S. Correa, Victoria J. Orphan, and Rebecca Vega Thurber. 2015. “A Novel Sister Clade to the Enterobacteria Microviruses (Family Microviridae) Identified in Methane Seep Sediments.” Environmental Microbiology 17 (10): 3708–21. https://doi.org/10.1111/1462-2920.12758.

Cobián Güemes, Ana Georgina, Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton, and Forest Rohwer. 2016. “Viruses as Winners in the Game of Life.” Annual Review of Virology 3 (1): 197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Cook, Monte, Jonathan Tweet, and Skip Williams. 2003. Dungeons & Dragons Dungeon Master’s Guide : Core Rulebook II. Wizards of the Coast.

Corinaldesi, Cinzia, Michael Tangherlini, and Antonio Dell’Anno. 2017. “From Virus Isolation to Metagenome Generation for Investigating Viral Diversity in Deep-Sea Sediments.” Scientific Reports 7 (1): 1–12. https://doi.org/10.1038/s41598-017-08783-4.

Edwards, Robert A., and Forest Rohwer. 2005. “Viral Metagenomics.” Nature Reviews Microbiology 3 (June): 3–13. https://doi.org/10.1002/9781118010549.ch2.

Emerson, Joanne B, Simon Roux, Jennifer R Brum, Benjamin Bolduc, Ben J Woodcroft, Ho Bin Jang, Caitlin M Singleton, et al. 2018. “Host-Linked Soil Viral Ecology along a Permafrost Thaw Gradient.” Nature Microbiology. https://doi.org/10.1038/s41564-018-0190-y.

Erez, Zohar, Ida Steinberger-levy, Maya Shamir, Shany Doron, Avigail Stokar-avihail, Yoav Peleg, and Sarah Melamed. 2017. “Communication between Viruses Guides Lysis – Lysogeny Decisions.” Nature Publishing Group 541 (7638): 488–93. https://doi.org/10.1038/nature21049.

Han, Li Li, Dan Ting Yu, Li Mei Zhang, Ju Pei Shen, and Ji Zheng He. 2017. “Genetic and Functional Diversity of Ubiquitous DNA Viruses in Selected Chinese Agricultural Soils.” Scientific Reports 7: 1–10. https://doi.org/10.1038/srep45142.

Hendrix, R W, M C Smith, R N Burns, M E Ford, and G F Hatfull. 1999. “Evolutionary Relationships among Diverse Bacteriophages and Prophages: All the World’s a Phage.” Proceedings of the National Academy of Sciences of the United States of America 96 (5): 2192–97. https://doi.org/10.1073/pnas.96.5.2192.

Knowles, B, C B Silveira, B A Bailey, K Barott, V A Cantu, A G Cobián-Güemes, F H Coutinho, et al. 2016. “Lytic to Temperate Switching of Viral Communities.” Nature 531 (7595): 466–70. https://doi.org/10.1038/nature17193.

Koonin, Eugene V., Valerian V. Dolja, and Mart Krupovic. 2015. “Origins and Evolution of Viruses of Eukaryotes: The Ultimate Modularity.” Virology 479–480: 2–25. https://doi.org/10.1016/j.virol.2015.02.039.

Koonin, Eugene V, Kira S Makarova, Yuri I Wolf, and Mart Krupovic. 2019. “Ecolutionary Entanglement of Mobile Elements and Host Defense Systems : Guns for Hire.” Nature Reviews Genetics. https://doi.org/10.1038/s41576-019-0172-9.

Lee, Yan Jiun, Nan Dai, Shannon E. Walsh, Stephanie Müller, Morgan E. Fraser, Kathryn M. Kauffman, Chudi Guan, Ivan R. Corrêa, and Peter R. Weigele. 2018. “Identification and Biosynthesis of Thymidine Hypermodifications in the Genomic DNA of Widespread Bacterial Viruses.” Proceedings of the National Academy of Sciences of the United States of America 115 (14): E3116–25. https://doi.org/10.1073/pnas.1714812115.

Ofir, Gal, and Rotem Sorek. 2018. “Review Contemporary Phage Biology : From Classic Models to New Insights.” Cell 172 (6): 1260–70. https://doi.org/10.1016/j.cell.2017.10.045.

Ptashne, Mark. 2004. A Genetic Switch Phage Lambda Revisited. 3rd ed. Cold Spring Harbor Laboratory Press.

Rauch, Benjamin J., Melanie R. Silvis, Judd F. Hultquist, Christopher S. Waters, Michael J. McGregor, Nevan J. Krogan, and Joseph Bondy-Denomy. 2017. “Inhibition of CRISPR-Cas9 with Bacteriophage Proteins.” Cell 168 (1–2): 150-158.e10. https://doi.org/10.1016/j.cell.2016.12.009.

Rohwer, Forest, and Rob Edwards. 2002. “The Phage Proteomic Tree : A Genome-Based Taxonomy for Phage” 184 (16): 4529–35. https://doi.org/10.1128/JB.184.16.4529.

Scola, Vincent, Jean Baptiste Ramond, Aline Frossard, Olivier Zablocki, Evelien M. Adriaenssens, Riegardt M. Johnson, Mary Seely, and Don A. Cowan. 2018. “Namib Desert Soil Microbial Community Diversity, Assembly, and Function Along a Natural Xeric Gradient.” Microbial Ecology 75 (1): 193–203. https://doi.org/10.1007/s00248-017-1009-8.

Trubl, Gareth, Ho Bin Jang, Simon Roux, Joanne B. Emerson, Natalie Solonenko, Dean R. Vik, Lindsey Solden, et al. 2018. “Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing.” MSystems 3 (5): 1–21. https://doi.org/10.1128/msystems.00076-18.

Yoshida, Mitsuhiro. 2018. “Quantitative Viral Community DNA Analysis Reveals the Dominance of Single-Stranded DNA Viruses in Offshore Upper Bathyal Sediment from Tohoku , Japan.” Frontiers in Microbiology 9 (February): 1–10. https://doi.org/10.3389/fmicb.2018.00075.

Yu, Dan Ting, Ji Zheng He, Li Mei Zhang, and Li Li Han. 2019. “Viral Metagenomics Analysis and Eight Novel Viral Genomes Identified from the Dushanzi Mud Volcanic Soil in Xinjiang, China.” Journal of Soils and Sediments 19 (1): 81–90. https://doi.org/10.1007/s11368-018-2045-9.

Zeng, Lanying, Samuel O. Skinner, Chenghang Zong, Jean Sippy, Michael Feiss, and Ido Golding. 2010. “Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection.” Cell 141 (4): 682–91. https://doi.org/10.1016/j.cell.2010.03.034.