UNIVERSITY OF CALIFORNIA, SAN DIEGO

SAN DIEGO STATE UNIVERSITY

Polymicrobial Infections in Cystic Fibrosis Lungs: The Need for Personalized Treatment

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

in

Biology

by

Yan Wei Lim

Committee in charge:

University of California, San Diego

Professor Douglas Conrad
Professor Joseph Pogliano

San Diego State University

Professor Forest Rohwer, Chair
Professor Kelly Doran
Professor Robert Edwards

2015

The dissertation of Yan Wei Lim 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
2015

DEDICATION

I dedicate this dissertation to my good friend, Mike Furlan. I hope that this achievement will complete part of the dream you had in studying Cystic Fibrosis.

EPIGRAPH

"What if …" is where the ideas come from.

Unknown Source

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The Rohwer lab is a great place for science. It is a collaborative lab with dynamic interactions. Things were moving fast even though it did not seem so at certain moments. In addition to the lab members, there is a mastermind. None of this work would have been possible without Dr. Forest Rohwer, my teacher and mentor. Through his trust and support, I was given the opportunity to be an independent scientist while learning the skills from him. He led me to believe nothing is impossible in science and taught me to think outside the box - looking at things from all angles. Even though studying human diseases is my primary research goal, I was given the opportunity to venture into the underwater world and participated in the 2013 Southern Line Island expedition. I am grateful for that. We brought a sequencer to the most remote islands on earth, working alongside with one of the best computer scientists, Dr. Robert Edwards, and other excellent marine scientists.

I would also like to thank the members of my committee for all their feedback and support. I want to thank Dr. Robert Edwards for all the help he had given me. Unlimited access to computing and personnel resources in your lab has granted this work. To Dr. Douglas Conrad, I thank you for the opportunity to work with you and your team, as well as the patients in the clinic. Your perspective was key to the work I have done. I would also like to thank Dr. Kelly Doran for her amazing lab resources and lab members. Thank you for the opportunity to work in your lab, to learn the skills that I would not have otherwise have. Thank you Dr. Joe Pogliano for the amazing microbiology classes and your career advice.

To the Rohwer lab and Edwards lab members, past and present, thank you for your support and friendship. I have had the honor to learn from a group of talented scientists and individuals. A special thank you to Dr. Dana Willner, Dr. Matthew Haynes, and Mike Furlan who jump-started the path to this dissertation work. Another special thank you to Dr. Anca Segall, Dr. Peter Salamon, and Dr. Barbara Bailey for all your help throughout the years. Thank you to Daniel Cuevas, Genivaldo Gueiros Z. Silva and Katelyn M. for all your bioinformatics support. I would like to thank Gina Spidel and Leslie Rodelander for their friendship and administrative support. Without you, I would not be here physically and none of this work would have been possible. To all the Rohwer lab ladies. Thank you for your friendship and crazy times, the memories that I will hold on forever.

For financial support, I acknowledge the National Institute of Health (NIH), Cystic Fibrosis Research Inc. (CFRI), Inamori Foundation, California State University Program for Education and Research in Biotechnology (CSUPERB), Canadian Institute for Advanced Research (CIFAR), International Society of Microbial Ecology (ISME), and Mabel's Myers Memorial.

Lastly, I want to thank my family and friends. To my parents, I thank you for giving me the best education possible. Uncle John, and Aunty Wan Thai, I thank you for your unconditional love and support. To Rob, I thank you for your love, encouragement, and all you have done for me. I look forward to our future endeavors in life and science.

Chapter 2, in full, is published in the Journal of Visualized Experiments. Yan Wei Lim, Matthew Haynes, Mike Furlan, Charles E. Robertson, J. Kirk Harris, and Forest

Rohwer; 2014. The dissertation author was the primary investigator and author of this paper.

Chapter 3 in full is published in the Journal of Cystic Fibrosis. Yan Wei Lim, Robert Schmieder, Matthew Haynes, Dana Willner, Mike Furlan, Merry Youle, Katelynn Abbott, Robert Edwards, Jose S. Evangelista III, Douglas Conrad, and Forest Rohwer; 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is published in the Journal of Clinical Microbiology. Yan Wei Lim, Jose S. Evangelista III, Robert Schmieder, Barbara Bailey, Matthew Haynes, Mike Furlan, Heather Maughan, Robert Edwards, Forest Rohwer, and Douglas Conrad; 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is published in the PLoS ONE journal. Yan Wei Lim, Robert Schmieder, Matthew Haynes, Mike Furlan, T. David Matthews, Katrine Whiteson, Stephen J. Poole, Christopher S. Hayes, David A. Low, Heather Maughan, Robert Edwards, Douglas Conrad, and Forest Rohwer; 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is in preparation for submission. Yan Wei Lim, Katrine Whiteson, Barbara Bailey, Peter Salamon, Ben Felts, Andreas Haas, Robert Quinn, Mark Hatay, Douglas Conrad, Robert Edwards, and Forest Rohwer; 2015. The dissertation author was the primary investigator and author of this paper.

VITA

2015 Doctor of Philosophy in Biology, University of California, San Diego and San Diego State University

2009 – 2011 Research Assistant, Global Viral Forecasting Initiative Inc., San Francisco, CA

2006 – 2009 Research Assistant, Genome Institute of Singapore, Singapore

2006 Bachelor of Science (Honors) in Biomedical Sciences, University of Sunderland, United Kingdom

PUBLICATIONS

Lim Y.W., Whiteson K., Bailey B., Salamon, P., Felts B., Haas A., Quinn R., Hatay M., Conrad D., Edwards A. R., and Rohwer F. Metabolic activities of CF anaerobic communities and their responses to perturbations caused by oxygen and pressure (*in preparation*)

Lim Y.W., Silva G.G.Z.*, Cuevas D.A.*, Rohwer F., and Edwards A. R. The genetic differences of CF-derived *Prevotella* and *Veillonella* compared to oral isolates (*in preparation*)

Lim Y.W., Haynes M., Furlan M., Robertson C.E., Harris J.K., Rohwer F. (2014) Purifying the impure: Sequencing metagenomes and metatranscriptomes from complex animal-associated samples. Journal of Visualized Experiment

Lim Y.W., Cuevas DA, Silva GGZ, Aguinaldo K, Dinsdale EA, Haas AF, Hatay M, Sanchez SE, Wegley-Kelly L, Dutilh BE, Harkins TT, Lee CC, Tom W, Sandin SA, Smith JE, Zgliczynski B, Vermeij MJA, Rohwer F, Edwards RA. (2014) Sequencing at sea: challenges and experiences in Ion Torrent PGM sequencing during the 2013 Southern Line Islands Research Expedition. PeerJ

Lim Y.W., Evangelista J.S III, Schmieder R., Bailey B., Haynes M.R., Furlan M., Maughan H., Edwards R., Rohwer F., Conrad D. (2014) Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis. Journal of Clinical Microbiology

Lim Y.W., Schmieder R., Haynes M.R., Furlan M., Matthews T. D., Whiteson K., Poole S. J., Hayes C. S., Low D. A., Maughan H., Edwards R., Conrad D., Rohwer F. (2013)

Mechanismtic model of Rothia mucilaginosa adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data. PLoS ONE

Lim Y.W., Schmieder R., Haynes M.R., Willner D., Furlan M., Youle M., Abbott K., Edwards R., Evangelista J., Conrad D., Rohwer F. (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. Journal of Cystic Fibrosis http://dx.doi.org/10.1016/j.jcf.2012.07.009

Beyter D., Tang P.Z., Becker S., Hoang T., Bilgin D., **Lim Y.W**., Peterson T., Mayfield S., Haerizadeh F., Shurin J., Bafna V., and McBride R. Diversity, Productivity and Stability of an Industrial Microbial Ecosystem (*in revision*)

Quinn R.A., **Lim Y.W**., Whiteson K., Furlan M., Conrad D., Rohwer F., and Dorrestein P. Metabolomics of pulmonary exacerbation reveals the personalized nature of cystic fibrosis disease (*in revision*)

Quinn R.A., Vermeij M. J. A., Hartmann A.C., Galtier d'Auriac I., Benler S., Haas A., Quistad S.D., **Lim Y.W.**, Little M., Smith J., Dorrestein P., and Rohwer F. Metabolomics of holobiont interactions on pristine coral reefs (*in revision*)

Haas, A.F., Knowles, B., **Lim Y.W.**, McDole Somera, T., Kelly, L.W., Hatay, M., et al. (2014) Unraveling the Unseen Players in the Ocean - A Field Guide to Water Chemistry and Marine Microbiology. Journal of Visualized Experiment

Grasis J.A., Lachnit T., Anton-Erxleben F.A., **Lim Y.W.**, Schmieder R., Fraune S., et al. (2014) Species-specific viromes in the ancestral holobiont Hydra. PLos ONE

Kelly L.W., Williams G.J., Barott K.L., Carlson C.A., Dinsdale E.A., Edwards R.A., Haas A.F., Haynes M., **Lim Y.W.**, McDole T., Nelson C.E., Sala E., Sandin S.A., Smith J.E., Vermeij M.JA., Youle M., Rohwer F. (2014) Local genomic adaptation of coral reef-associated microbiomes to gradients of natural variability and anthropogenic stressors. PNAS

Whiteson K.L., Bailey B., Bergkessel M., Conrad D., Delhaes L., Felts B., Harris J.K., Hunter R., **Lim Y.W.**, Maughan H., Quinn R., Salamon P., Sullivan J., Wagner B., Rainey P. (2014) The upper respiratory tract as a microbial source for pulmonary infections in Cystic Fibrosis: Parallels from Island Biogeography. American Journal of Respiratory and Critical Care Medicine.

Garg N., Kapono C., **Lim Y.W.**, Koyama N., Vermeij M.J.A., Conrad D., Rohwer F., Dorrestein P. (2014) Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. International Journal of Mass Spectrometry

Quinn R., **Lim Y.W.**, Maughan H., Rohwer F., Whiteson K.L. (2014) Biogeochemical Forces Shape the Composition and Physiology of Polymicrobial Communities in the Cystic Fibrosis Lung. MBIO

Whiteson K.L., Meinardi S., **Lim Y.W.**, Schmieder R., Maughan H., Quinn R., Blake D., Conrad D., Rohwer F. (2014) Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. ISME Journal

Mokili J.L., Dutilh B.E., **Lim Y.W.**, Schneider B.S., Taylor T, et al. (2013) Identification of a Novel Human Papillomavirus by Metagenomic Analysis of Samples from Patients with Febrile Respiratory Illness. PLoS ONE 8(3): e58404. doi:10.1371/journal.pone.0058404

Schmieder R., **Lim Y.W.**, Edwards R. (2011) Identification and removal of ribosomal RNA sequences from metatranscriptomes Bioinformatics doi: 10.1093/bioinformatics/btr669

Willner D., Haynes M.R., Furlan M., Hanson N., Kirby B., **Lim Y.W.**, Rainey P.B., Schmieder R., Youle M., Conrad D., Rohwer F. (2011) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. American Journal of Respiratory Cell and Molecular Biology. rcmb.2011-0253OC

Willner D., Haynes M.R., Furlan M., Schmieder R., **Lim Y.W.**, Rainey P.B., Rohwer F., Conrad D. (2011) Spatial distribution of microbial communities in the cystic fibrosis lung. ISMEJ. 2011.104

Ng T.F.F., Willner D., **Lim Y.W.**, Schmieder R., Chau B., Nilsson C., Anthony S., Ruan Y., Rohwer F., Breitbart M. (2011) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS ONE 6(6): e20579.

Schmieder R., **Lim Y.W.**, Rohwer F. and Edwards R. (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics, 11(1), 341.

Ng T.F.F., Willner D., Nilsson C., **Lim Y.W.**, Schmieder R., Chau B., Ruan Y., Rohwer F., Breitbart M. (2010) Vector-based metagenomics for animal virus surveillance. International Journal of Infectious Diseases, DOI:10.1016/j.ijid.2010.02.461.

Rosario K., Nilsson C., **Lim Y.W.**, Ruan Y., Breitbart M. (2009) Metagenomic analysis of viruses in reclaimed water. Environmental Microbiology, 1462-2920

ABSTRACT OF THE DISSERTATION

Polymicrobial Infections in Cystic Fibrosis Lungs: The Need for Personalized Treatment

by

Yan Wei Lim

Doctor of Philosophy in Biology

University of California, San Diego, 2015
San Diego State University, 2015

Professor Forest Rohwer, Chair

Polymicrobial infection defines a combination of microorganisms, including viruses, bacteria, fungi, and sometimes parasites, which are simultaneously associated with an infected site. While the idea of polymicrobial infections is common nowadays, most clinical laboratories still focus on the culturing and identification of a single pathogen. The biological implications attributed to these singular microbial infections are deceptive, especially when considering the multiple complex interactions associated with polymicrobial infections. This dissertation used Cystic Fibrosis (CF) as a study system for polymicrobial infections in the lung. CF is a genetic disease caused by mutations in the Cystic Fibrosis Transmembrane Regulator (CFTR) gene. CF affects multiple organs across the body, but pulmonary infection remains the main cause of morbidity and mortality. Studying CF lungs is challenging. Here a set of comprehensive methods was developed to simultaneously study the viral, microbial, and host genetics using metagenomics and metatranscriptomics approaches. Many of the bacteria found in the lungs of CF patients are of oral origin. Detailed genetic analyses showed that these bacteria adapt to the CF lungs by acquiring essential genes while losing non-essential ones. One example is *Rothia mucilaginosa* – an organism that is present in more than 80% of San Diego patients. *De novo* assembly of a near-complete *R. mucilaginosa* genome showed potential physiological adaptations through the acquisition of multiple genes. A survey of more than 20 CF patients in San Diego revealed that every patient harbors a unique microbial community, suggesting that CF patients require personalized medicine for the treatment of their lung infections. Despite these personalized differences in microbial taxonomical profiles across CF patients, the functional potential across these different microbial communities is highly similar. The conserved functional potential in CF microbes allows them to carry out aerobic and anaerobic respirations, as well as fermentation, highlighting the importance of anaerobes in the CF lungs. CF microbes regulate their metabolic activities in response to perturbations. *In vitro* community culturing of CF microbes showed that the anaerobes were sensitive to antibiotics commonly used in CF patients and their metabolic activities could be associated with the patient's health. The role and importance of CF anaerobes and their survival mechanisms are illustrated in this study. In summary, this dissertation provides novel insights into polymicrobial infections in CF lungs and demonstrates the potential and advantages of coupling omics and clinical approaches for the study of other complex polymicrobial infections.

CHAPTER 1

Introduction

Polymicrobial Infections and Cystic Fibrosis

Polymicrobial infections are infections of mixed-microbial communities resulting in a phenotypic disease. Microbial colonization may reflect complex interactions and often the cause-and-effect relationships are complicated and difficult to resolve. The overarching goal of this dissertation is to understand the polymicrobial interactions and community level mechanisms occurring within the Cystic Fibrosis (CF) lung.

Current trend in the study of polymicrobial infections: Since the introduction of culture-independent sequence-based approaches, polymicrobial infections have seen increased attention in the study of infectious diseases. Early studies of microbial diseases relied on microscopy for the identification of the causative agent, often missing the complex interactions occurring within polymicrobial infections (Miller 1890). Polymicrobial infections are common in oral and respiratory diseases (Chen et al. 1996), abscesses in brain, liver, and lungs (Sibley et al. 2012), atherosclerotic plaque (Kozarov et al. 2005), as well as skin and soft tissue infections such as those seen in diabetes foot, cellulitis, and necrotic fasciitis (Sapico et al. 1984; Dryden 2010). Polymicrobial infections involve combined effects of microorganisms including viruses, bacteria, fungi, and sometimes parasites (Brogden et al. 2005). In some cases, the primary infection itself suppresses the host immune system resulting in increased host susceptibility to opportunistic and often fatal secondary infections (Slifka et al. 2003; Lawn 2004). In others, the colonization of mixed-microbes as a community can be attributed to the synergistic interactions between microbes or the predisposition from one microbe to another. Some examples of synergistic interactions between microbes in polymicrobial infections include plaque formation by *Streptococcus oralis* and *Actinomyces naeslundii* (Palmer et al. 2001), and abscesses caused by *Streptococcus milleri* strains and *Fusobacterium nucleatum* (Nagashima et al. 1999). Bacterial superinfections due to viral infection are common in otitis media (Heikkinen and Chonmaitree 2003) and respiratory diseases such as those involving *Streptococcus pneumonia* (Kash et al. 2011), *Staphylococcus aureus* (Denison et al. 2013), and *Neisseria meningitidis* (Raza et al. 1999).

Cystic Fibrosis (CF), a disease affecting multiple organs: CF is an autosomal recessive genetic disease caused by mutation in the Cystic Fibrosis Transmembrane Regulator (CFTR) gene (Riordan et al. 1989). To date, more than 1,800 CFTR variants across the 1,480 amino acids protein have been identified and associated with CF disease (CFTR2 database). CFTR is expressed in epithelial cells across many organs including the lungs, pancreas, liver, kidneys, and intestine (Jameson 1998). The CFTR protein functions as an anion channel regulated by cAMP-dependent phosphorylation that transports chloride and bicarbonate ions across the apical surface of epithelial cells (Anderson et al. 1991; Smith and Welsh 1992). The defective protein affects anion transport and fluid homeostasis on epithelial cells (Zielenski 2000; Reddy and Quinton 2001). Meconium ileus, blockage of the small intestine by neonatal feces, was a major complication in newborns with CF prior to the introduction of newborn screening programs. Nowadays, surgical intervention relieves the obstruction as soon as symptoms arise (Rescorla and Grosfeld 1993). Pancreatic insufficiency may lead to maldigestion and malnutrition in CF patients. Currently, it is compensated with pancreatic enzymes replacement therapy (Somaraju and Solis-Moya 2015). Dysfunctional CFTR proteins at the apical membrane of bile duct cells also cause hepatobiliary complications in CF patients (Colombo 2007). Early detection and treatment have improved patient outcome in recent years. Among the affected organs, pulmonary diseases remains the major cause of morbidity and mortality in CF patients (Gibson et al. 2003).

The progression of deadly pulmonary polymicrobial infections in Cystic Fibrosis: Functional defect of the CFTR proteins integrally shapes the landscape within CF lungs and leads to colonization by environmentally acquired opportunistic pathogens. Pulmonary disease in CF is characterized by polymicrobial infections (Sibley et al. 2008), airway mucus plugging (Fuchs et al. 1994), inefficient immune responses (Cohen and Prince 2012), and airway tissue remodeling. Despite aggressive treatments, CF patients undergo pulmonary exacerbations (CFPE). CFPE is characterized by a decline in lung function, as well as increased cough, and sputum production that results in airway scarring, eventual loss of respiratory function, and death.

In the respiratory epithelium, lack of proper CFTR function leads to an imbalance in Cl^- uptake (Quinton 1983), Na^+ hyper-reabsorption (Boucher 2007), and reduced HCO_3^- secretion (Quinton 2010). This leads to reduced airway fluid secretion, dehydration of the airway surface liquid (ASL), and results in impaired mucus clearance. The accumulation of thick and static mucus in the airway lumen causes plugs and blockages, creating an environment that allows for the colonization of opportunistic pathogens (Smith et al. 1996). Defective ion exchange in CF patients also leads to lower ASL pH, which reduces the effectiveness of antimicrobial immune responses (Pezzulo et al. 2012). The combined effect of opportunistic colonization and ineffective immune response (Stoltz et al. 2010) causes recurrent and chronic infections, leading to a progressive decline in lung function.

Microbiology and community complexity of Cystic Fibrosis airways: Based on standard microbiology culture data, *Staphylococcus aureus* and *Haemophilus influenza* are commonly found in younger CF patients (LiPuma 2010). As patients age, the communities often become dominated by *Pseudomonas aeruginosa*, *Burkholderia cepacia* complex, *Stenotrophomonas maltophilia*, and *Achromobacter* spp.; many of which are highly resistant against multiple antibiotics (LiPuma 2010; Cox et al. 2010). Using 16S rRNA gene analysis by microarray and high-throughput sequencing, more than 1,000 different taxa across ~90 genera have been associated with CF airways (Cox et al. 2010; Guss et al. 2011). Many of the species are environmental and oral-associated opportunistic pathogens, consisting of both aerobes and anaerobes. CF lungs present spatially structured microenvironments that range from highly oxygenated airways to anoxic mucus plugs (Worlitzsch et al. 2002). In CF pulmonary infections, the airways host a complex assemblage of opportunistic pathogens, including viruses, bacteria, and fungi (LiPuma 2010), which can stratify into the structured microenvironment. An ecological model of *Climax* and *Attack* communities introduced in 2013 by Conrad *et al.* suggests that the dynamic changes in microbial communities in CF airways formed two distinct functional communities (Conrad et al. 2013). The *Attack* community consists of transient viral and microbial populations such as respiratory syncytial viruses, rhinoviruses, *S. aureus,* and *H. influenza*. Due to their virulent properties, infections by *Attack* communities result in strong innate immune responses that initiate airways remodeling and create scarring. The resultant microenvironments facilitate the formation of the *Climax* community, characterized by slow microbial growth and high resistance to antimicrobial compounds. The concept is in line with the observation that the diversity of the bacterial communities peaks at an early age and decreases with increasing age, corresponding with the decrease in lung function (Cox et al. 2010).

Clinical microbiology and polymicrobial infections: In natural environments, no microbes live in isolation. The same is also true for microbes associated with the human body. While the idea of polymicrobial infections is well accepted, most clinical laboratories remain focus on growing microbes in isolation to comply with the gold standard diagnostic procedure as well as the antimicrobial spectrum and susceptibility testing procedures (Brook et al. 2013; Baron et al. 2013). Study of single isolates is important when studying microbial pathogenesis, the molecular and biochemical properties of the pathogen, and efficacy of available drugs. However, many of these biological properties are likely to be deceptive in polymicrobial infections (Miller 1890). In cases of known polymicrobial infections, multiple culture-based tests have to be explicitly ordered by the attending physician (Baron et al. 2013). Even though culture-independent technologies such as polymerase chain reaction (PCR) became more accessible in recent years, its use is limited to commonly known pathogens (Wolk and Dunne 2011).

A peek into the following chapters: Using pulmonary infections in the CF lungs as a study system, this dissertation presents a model for how sequence-based culture-independent approaches can be used to identify and characterize polymicrobial communities in a complex host-associated system (Chapter 2). These approaches discover the unique polymicrobial communities presented by individual patients and identify common mechanisms that may cause pulmonary exacerbation in CF patients (Chapter 3 and Quinn et al. 2014; Whiteson et al. 2014). Incorporating clinical data, Chapter 4 provides clinical insights from the metagenomic analysis of CF sputum. This represents the first attempt to couple clinical data with metagenomic sequencing as a proof of principle for personalized health and disease monitoring.

Preliminary findings suggest that anaerobes, specifically those originating from oral cavities, play a significant role in disease progression within CF lungs. In Chapter 5, a near complete *Rothia mucilaginosa* genome, reconstructed through *de novo* assembly of the metagenomics reads is presented. Detailed comparison of the CF-derived genome to a periodontitis isolate identified several physiological adaptations of CF isolates to the CF lung environment. Chapter 5 explores microhabitat-specific evolution of microbes as an explanation for the variable outcomes in polymicrobial infections.

A combination of metagenomics, metatranscriptomics, metabolomics, and clinical information led to the hypothesis that anaerobic microbial activities, specifically fermentation, lead to exacerbation events in CF patients. Chapter 6 presents an *in vitro* community culture model to examine the hypothesis and to study the anaerobic communities in CF lungs. Metagenomics, metatranscriptomics, and metabolomics were used to elucidate the important metabolic activities in the anaerobic communities. Finally in Chapter 7, a model is presented to describe the current hypothesis of CF lung pathogenesis and the importance of personalized medicine in CF patients. In addition, the potential use of these approaches in studying other chronic polymicrobial infections is discussed.

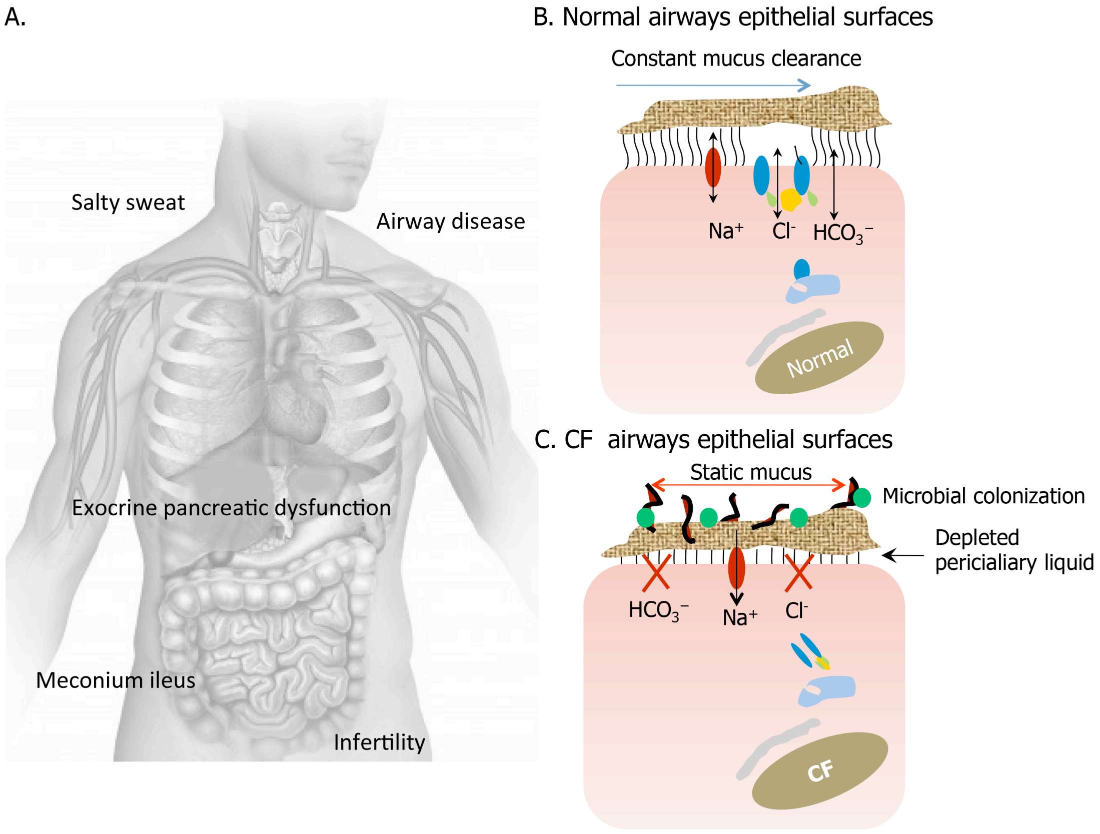


Figure 1.1: The global effect of CFTR dysfunction on multiple organs and airway epithelial cells. (A) Cystic Fibrosis Transmembrane Regulator (CFTR) gene is expressed in epithelial cells of many organs including the sweat gland, respiratory tract (salivary gland, nose, sinus, lung), pancreas, small intestine, colon, and kidney. (B) Airway epithelial surface with normal CFTR function allows constant mucus clearance that prevents colonization of opportunistic pathogens. (C) Defective CFTR function in CF airways accumulates thick and static mucus that seeds acute and chronic polymicrobial infection.

References

Anderson MP, Gregory RJ, Thompson S, Souza DW, Paul S, Mulligan RC, Smith AE, Welsh MJ (1991) Demonstration that CFTR is a chloride channel by alteration of its anion selectivity. Science 253:202–205.

Baron EJ, Miller JM, Weinstein MP, Richter SS, Gilligan PH, Thomson RB, Bourbeau P, Carroll KC, Kehl SC, Dunne WM, Robinson-Dunn B, Schwartzman JD, Chapin KC, Snyder JW, Forbes BA, Patel R, Rosenblatt JE, Pritt BS (2013) A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2013 recommendations by the Infectious Diseases Society of America (IDSA) and the American Society for Microbiology (ASM). Clin Infect Dis cit278. doi: 10.1093/cid/cit278

Boucher RC (2007) Evidence for airway surface dehydration as the initiating event in CF airway disease. J Intern Med 261:5–16. doi: 10.1111/j.1365-2796.2006.01744.x

Brogden KA, Guthmiller JM, Taylor CE (2005) Human polymicrobial infections. The Lancet 365:253–255. doi: 10.1016/S0140-6736(05)17745-9

Brook I, Wexler HM, Goldstein EJC (2013) Antianaerobic antimicrobials: Spectrum and susceptibility testing. Clin Microbiol Rev 26:526–546. doi: 10.1128/CMR.00086-12

CFTR2 database The Clinical and Functional TRanslation of CFTR (CFTR2). http://cftr2.org.

Chen PB, Davern LB, Katz J, Eldridge JH, Michalek SM (1996) Host responses induced by co-infection with Porphyromonas gingivalis and Actinobacillus actinomycetemcomitans in a murine model. Oral Microbiol Immunol 11:274–281.

Cohen TS, Prince A (2012) Cystic fibrosis: a mucosal immunodeficiency syndrome. Nat Med 18:509–519. doi: 10.1038/nm.2715

Colombo C (2007) Liver disease in cystic fibrosis. Curr Opin Pulm Med 13:529–536. doi: 10.1097/MCP.0b013e3282f10a16

Conrad D, Haynes M, Salamon P, Rainey PB, Youle M, Rohwer F (2013) Cystic fibrosis therapy: a community ecology perspective. Am J Respir Cell Mol Biol 48:150–156. doi: 10.1165/rcmb.2012-0059PS

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044. doi: 10.1371/journal.pone.0011044

Denison AM, DeLeon-Carnes M, Blau DM, Shattuck EC, McDougal LK, Rasheed JK, Limbago BM, Zaki SR, Paddock CD (2013) Molecular characterization of Staphylococcus aureus and Influenza virus coinfections in patients with fatal pneumonia. J Clin Microbiol 51:4223–4225. doi: 10.1128/JCM.02503-13

Dryden MS (2010) Complicated skin and soft tissue infection. J Antimicrob Chemother 65:iii35–iii44. doi: 10.1093/jac/dkq302

Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash MI, Ramsey BW, Rosenstein BJ, Smith AI, Wohl ME (1994) Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with Cystic Fibrosis. The New England Journal of Medicine 331:637–642.

Gibson RL, Burns JL, Ramsey BW (2003) Pathophysiology and management of pulmonary infections in Cystic Fibrosis. Am J Respir Crit Care Med 168:918–951. doi: <p>10.1164/rccm.200304-505SO</p>

Guss AM, Roeselers G, Newton ILG, Young CR, Klepac-Ceraj V, Lory S, Cavanaugh CM (2011) Phylogenetic and metabolic diversity of bacteria associated with Cystic Fibrosis. ISME J 5:20–29. doi: 10.1038/ismej.2010.88

Heikkinen T, Chonmaitree T (2003) Importance of Respiratory Viruses in Acute Otitis Media. Clin Microbiol Rev 16:230–241. doi: 10.1128/CMR.16.2.230-241.2003

Jameson JL (1998) Principles of molecular medicine (Page 331). Springer Science & Business Media

Kash JC, Walters K-A, Davis AS, Sandouk A, Schwartzman LM, Jagger BW, Chertow DS, Li Q, Kuestner RE, Ozinsky A, Taubenberger JK (2011) Lethal synergism of 2009 pandemic H1N1 influenza virus and Streptococcus pneumoniae coinfection is associated with loss of murine lung repair responses. mBio. doi: 10.1128/mBio.00172-11

Kozarov EV, Dorn BR, Shelburne CE, Dunn WA, Progulske-Fox A (2005) Human atherosclerotic plaque contains viable invasive Actinobacillus actinomycetemcomitans and Porphyromonas gingivalis. Arterioscler Thromb Vasc Biol 25:e17–18. doi: 10.1161/01.ATV.0000155018.67835.1a

Lawn SD (2004) AIDS in Africa: the impact of coinfections on the pathogenesis of HIV-1 infection. J Infect 48:1–12.

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clin Microbiol Rev 23:299 –323. doi: 10.1128/CMR.00068-09

Miller WD (1890) The micro-organisms of the human mouth. SS White Dent MFG Co Phila USA 25.

Nagashima H, Takao A, Maeda N (1999) Abscess forming ability of streptococcus milleri group: synergistic effect with Fusobacterium nucleatum. Microbiol Immunol 43:207–216.

Palmer RJ, Kazmerzak K, Hansen MC, Kolenbrander PE (2001) Mutualism versus Independence: Strategies of mixed-species oral biofilms in vitro using saliva as the sole nutrient source. Infect Immun 69:5794–5804. doi: 10.1128/IAI.69.9.5794-5804.2001

Pezzulo AA, Tang XX, Hoegger MJ, Abou Alaiwa MH, Ramachandran S, Moninger TO, Karp PH, Wohlford-Lenane CL, Haagsman HP, van Eijk M, Banfi B, Horswill AR, Stoltz DA, McCray PB, Welsh MJ, Zabner J (2012) Reduced airway surface pH impairs bacterial killing in the porcine Cystic Fibrosis lung. Nature 487:109–113. doi: 10.1038/nature11130

Quinn RA, Lim YW, Maughan H, Conrad D, Rohwer F, Whiteson KL (2014) Biogeochemical forces shape the composition and physiology of polymicrobial communities in the Cystic Fibrosis lung. mBio 5:e00956–13. doi: 10.1128/mBio.00956-13

Quinton PM (2010) Role of epithelial $HCO3^{-}$ transport in mucin secretion: lessons from cystic fibrosis. Am J Physiol Cell Physiol 299:C1222–1233. doi: 10.1152/ajpcell.00362.2010

Quinton PM (1983) Chloride impermeability in cystic fibrosis. Nature 301:421–422. doi: 10.1038/301421a0

Raza MW, El Ahmer OR, Ogilvie MM, Blackwell CC, Saadi AT, Elton RA, Weir DM (1999) Infection with respiratory syncytial virus enhances expression of native receptors for non-pilate Neisseria meningitidis on HEp-2 cells. FEMS Immunol Med Microbiol 23:115–124.

Reddy MM, Quinton PM (2001) Selective activation of cystic fibrosis transmembrane conductance regulator Cl- and HCO3- conductances. JOP J Pancreas 2:212–218.

Rescorla FJ, Grosfeld JL (1993) Contemporary management of meconium ileus. World J Surg 17:318–325.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–1073.

Sapico FL, Witte JL, Canawati HN, Montgomerie JZ, Bessman AN (1984) The infected foot of the diabetic patient: Quantitative microbiology and analysis of clinical features. Rev Infect Dis 6:S171–S176. doi: 10.1093/clinids/6.Supplement_1.S171

Sibley CD, Church DL, Surette MG, Dowd SE, Parkins MD (2012) Pyrosequencing reveals the complex polymicrobial nature of invasive pyogenic infections: microbial constituents of empyema, liver abscess, and intracerebral abscess. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol 31:2679–2691. doi: 10.1007/s10096-012-1614-x

Sibley CD, Parkins MD, Rabin HR, Duan K, Norgaard JC, Surette MG (2008) A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in Cystic Fibrosis patients. Proc Natl Acad Sci 105:15070 –15075. doi: 10.1073/pnas.0804326105

Slifka MK, Homann D, Tishon A, Pagarigan R, Oldstone MBA (2003) Measles virus infection results in suppression of both innate and adaptive immune responses to secondary bacterial infection. J Clin Invest 111:805–810. doi: 10.1172/JCI13603

Smith JJ, Welsh MJ (1992) cAMP stimulates bicarbonate secretion across normal, but not cystic fibrosis airway epithelia. J Clin Invest 89:1148–1153.

Smith, Travis S, E Greenberg, Welsh M (1996) Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. Cell 85:229–236.

Somaraju UR, Solis-Moya A (2015) Pancreatic enzyme replacement therapy for people with cystic fibrosis (Review). Paediatr Respir Rev 16:108–109. doi: 10.1016/j.prrv.2014.11.001

Stoltz DA, Meyerholz DK, Pezzulo AA, Ramachandran S, Rogan MP, Davis GJ, Hanfland RA, Wohlford-Lenane C, Dohrn CL, Bartlett JA, Nelson GA 4th, Chang EH, Taft PJ, Ludwig PS, Estin M, Hornick EE, Launspach JL, Samuel M, Rokhlina T, Karp PH, Ostedgaard LS, Uc A, Starner TD, Horswill AR, Brogden KA, Prather RS, Richter SS, Shilyansky J, McCray PB Jr, Zabner J, Welsh MJ (2010) Cystic fibrosis pigs develop lung disease and exhibit defective bacterial eradication at birth. Sci Transl Med 2:29ra31. doi: 10.1126/scitranslmed.3000928

Whiteson KL, Meinardi S, Lim YW, Schmieder R, Maughan H, Quinn R, Blake DR, Conrad D, Rohwer F (2014) Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. ISME J 8:1247–1258. doi: 10.1038/ismej.2013.229

Wolk DM, Dunne WM (2011) New technologies in clinical microbiology. J Clin Microbiol 49:S62–S67. doi: 10.1128/JCM.00834-11

Worlitzsch D, Tarran R, Ulrich M, Schwab U, Cekici A, Meyer KC, Birrer P, Bellon G, Berger J, Weiss T, Botzenhart K, Yankaskas JR, Randell S, Boucher RC, Döring G (2002) Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. J Clin Invest 109:317–325. doi: 10.1172/JCI13870

Zielenski J (2000) Genotype and phenotype in cystic fibrosis. Respir Int Rev Thorac Dis 67:117–133. doi: 29497

CHAPTER 2

Purifying the impure: Sequencing metagenomes and metatranscriptomes from complex animal-associated samples

Abstract

The accessibility of high-throughput sequencing has revolutionized many fields of biology. In order to better understand host-associated viral and microbial communities, a comprehensive workflow for DNA and RNA extraction was developed. The workflow concurrently generates viral and microbial metagenomes, as well as metatranscriptomes, from a single sample for next-generation sequencing. The coupling of these approaches provides an overview of both the taxonomical characteristics and the community encoded functions. The presented methods use Cystic Fibrosis (CF) sputum, a problematic sample type, because it is exceptionally viscous and contains high amount of mucins, free neutrophil DNA, and other unknown contaminants. The protocols described here target these problems and successfully recover viral and microbial DNA with minimal human DNA contamination. To complement the metagenomics studies, a metatranscriptomics protocol was optimized to recover both microbial and host mRNA that contains relatively few ribosomal RNA (rRNA) sequences. An overview of the data characteristics is presented to serve as a reference for assessing the success of the methods. Additional CF sputum samples were also collected to (i) evaluate the consistency of the microbiome profiles across seven consecutive days within a single patient, and (ii) compare the consistency of metagenomic approach to a 16S ribosomal RNA gene-based sequencing. The results showed that daily fluctuation of microbial profiles without antibiotic

perturbation was minimal and the taxonomy profiles of the common CF-associated bacteria were highly similar between the 16S rDNA libraries and metagenomes generated from the hypotonic lysis (HL)-derived DNA. However, the differences between 16S rDNA taxonomical profiles generated from total DNA and HL-derived DNA suggest that hypotonic lysis and the washing steps benefit in not only removing the human-derived DNA, but also microbial-derived extracellular DNA that may misrepresent the actual microbial profiles.

Introduction

Viral and microbial communities associated with the human body have been investigated extensively in the past decade through the application of sequencing technologies (Suau et al. 1999; Breitbart et al. 2002). The outcomes have led to the recognition of the importance microbes in human health and disease. The major initiative came from the human microbiome project that describes the bacteria (and some archaea) residing on human skin, and within oral cavities, airways, urogenital tract, and gastrointestinal tract (Proctor 2011). Further microbiome studies of healthy human airways through bronchoalveolar lavage (BAL) (Charlson et al. 2011; Pragman et al. 2012) and nasopharyngeal swabs (Charlson et al. 2011) have shown that the lung can serve as an environmental sampling device, results in transient microbial colonization in the airways. However, the impact of microbial colonization in impaired airway surfaces can lead to severe and chronic lung infections, such as those seen in Cystic Fibrosis (CF) patients.

CF is a lethal genetic disease caused by the mutation in Cystic Fibrosis Transmembrane Regulator (CFTR) gene (Kerem et al. 1989). These mutations give rise to defective CFTR proteins that in turn affect transepithelial ion transport across the apical surface of the epithelium. The disease affects multiple organ systems, but the majority of mortality and morbidity is attributable to CF lung disease (Kleven et al. 2008). The CF lung provides a unique ecosystem for microbial colonization. The defect in ion transport causes mucus to build up in the CF airways, creating microenvironments consisting of aerobic, microaerophilic, and anaerobic compartments anchored by a static nutrient-rich mucosal surface. This environment facilitates the colonization and proliferation of microbes, including viral, bacteria, and fungi. Acute and chronic pulmonary microbial infections lead to constant but ineffective immune responses, resulting in extensive airway remodeling, loss of pulmonary capacity, and ultimately pulmonary failure.

Bacterial communities associated with the CF lung have been well described using both culture-dependent and culture-independent approaches, which include using 16S ribosomal RNA (rRNA) gene sequencing (Fodor et al. 2012) and shotgun metagenomics (Lim et al. 2012; Lim et al. 2014). The 16S rRNA-based approach is able to characterize a wide range of microbial species and capture broad shifts in community diversity. However, it is limited in its resolution in defining the communities (summarized in Claesson et al. 2010) and the predictions of metabolic potentials are limited to those general functions known for the taxa identified. Therefore, 16S rRNA gene sequencing methods are insufficient for the necessary taxonomic and functional analytic accuracy of the diverse microbial communities present in CF lungs. The metagenomic approach described here complements the 16S rRNA-based approach, overcomes its limitations, and enables a relatively effective way to analyze both the microbial community taxonomy and genetic contents in CF lungs.

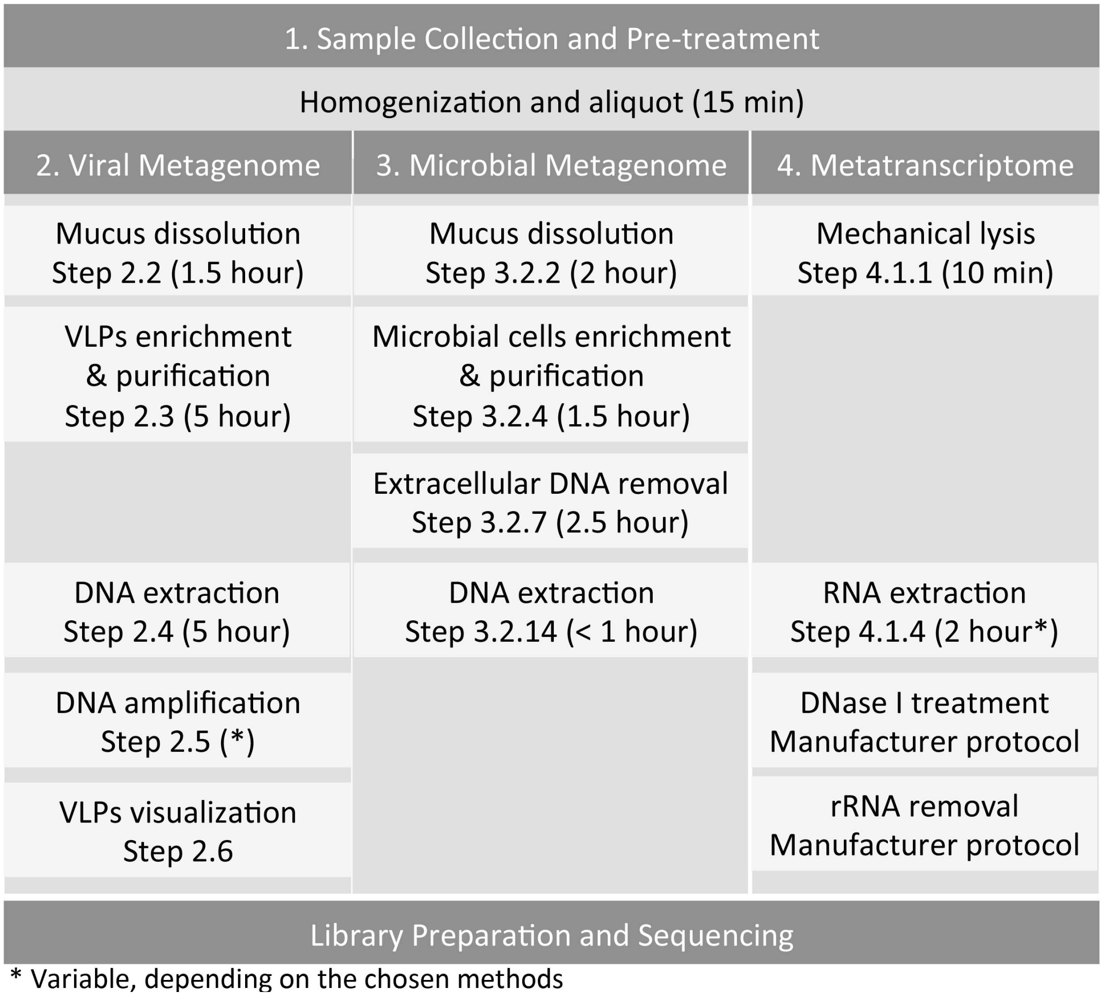
Microbial DNA isolated from animal-associated samples often contains a large amount of host DNA. CF sputum or lung tissue samples usually contain a large amount of human DNA released by neutrophils in the immune response, often greater than 99% of the total DNA (Lethem et al. 1990; Shak et al. 1990; Breitenstein et al. 1995a). Although some intact human cells may be present, most of this DNA is free in solution or adsorbed to the surface of microbes. In addition, the presence of exceptionally viscous mucus plugs, cellular debris, and other unknown contaminants further complicate isolation of microbial cells. Several methods were tested for depleting these samples of human DNA, including percoll gradients to separate human from microbial cells (Childs and Gibbons 1988), treatment with DNase I, ethidium bromide monoazide to selectively degrade human DNA(Lee and Levin 2006), and the MolYsis kit, all with limited success. To date the most effective microbial DNA purification procedure for CF sputum has been a modification of the process described by (Breitenstein et al. 1995b). This approach, herein known as hypotonic lysis (HL) method, uses a combination of β-mercaptoethanol to reduce mucin disulfide bonds, hypotonic lysis of eukaryotic cells, and DNase I treatment of soluble DNA (Lim et al. 2012). Despite the lack of alternatives the HL method raised some concerns due to (i) possible biases resulting from unwanted lysis of microbes and (ii) whether the observed fluctuations in community composition (Lim et al. 2012; Lim et al. 2014) are an artifact of variations associated with the sample processing. In addition to the generation of shogun metagenomes, we address these issues by comparing the 16S rRNA gene profiles of the total DNA and microbial DNA extracted from the HL method using the same set of sputum samples collected from a single patient across seven consecutive days.

Compared to microbial communities, the characterization of viral communities associated with animals is limited (Mokili et al. 2012; Bibby 2014). The viral communities in CF airways have only been characterized minimally (Willner et al. 2009; Willner and Furlan 2010; Willner et al. 2012). The first metagenomic study characterizing the DNA of viral communities in CF airways showed that most viruses associated with CF lungs are phages (Willner et al. 2009). The metabolic potential of phage in CF and non-CF individuals was significantly different. Specifically, the phage communities in CF individuals carried genes reflective of bacterial host adaptations to the physiology of CF airways, and bacterial virulence (Willner et al. 2009). Subsequent metagenomic studies of viruses in CF lung tissue demonstrated distinct spatial heterogeneity of viral communities between anatomical regions (Willner et al. 2012). In addition, CF lung tissue harbored the lowest viral diversity observed to date in any ecosystem (Willner et al. 2012). Most viruses identified were phages with the potential to infect CF pathogens. However, eukaryotic viruses such as herpesviruses, adenoviruses, and human papilloma viruses (HPV) were also detected. In one event, where cysts in the lung tissue were observed during dissection, more than 99% of a human papillomavirus genome was recovered, even though the patient was never diagnosed with a pulmonary papilloma or carcinoma. This indicates that the viral diversity present not only reflects the severity of tissue damage, but may also expose and explain an underlying uncharacterized disease. The protocols described here provide a simple, yet powerful way to isolate viral-like particles (VLPs) from samples that consist of large amounts of thick mucus, host and microbial cells, free DNA, as well as cell debris.

Complementing metagenomics, metatranscriptomics is used to monitor the dynamics in gene expression across the microbial community and the host (Bomar et al. 2011; Lim et al. 2012). In this case, both microbial and host mRNA need to be preferentially selected. Since bacterial mRNAs are not polyadenylated, an oligo-dT-based mRNA pull-down method cannot be exploited. Polyadenylation-dependent RNA amplification cannot be used in host-associated samples if the samples are known to contain large amounts of eukaryotic mRNA. Many animal-associated samples, including

CF sputum, contain a high density of cells in addition to high amounts of cellular debris and nucleases that include RNases. Therefore, another challenging task is to prevent extensive RNA degradation during metatranscriptome processing. In most cases, total RNA extracted from CF sputum is partially degraded, limiting the downstream applications and utility of the derived RNA. In recent years, several approaches for rRNA depletion have been developed and adapted in commercially available kits. The efficacy of these approaches is however limited, especially when working with partially degraded rRNA (He et al. 2010; Lim et al. 2012). The methods employed here allowed for the retrieval of partially degraded total RNA suitable for efficient downstream total rRNA removal. Direct comparison of the efficiency in rRNA removal from partially degraded total RNA comparing two different kits was illustrated by (Lim et al. 2012).

Overall, the goal of this manuscript is to provide a complete set of protocols (Figure 2.1) to generate viral and microbial shotgun metagenomes, and a metatranscriptome, from a single animal-associated sample, using induced sputum sample as an example. Molecular laboratory workflow should include separate pre- and post-amplification areas to minimize cross-contamination. The methods are easily adaptable to other sample types such as tissue (Willner et al. 2012), nasopharyngeal and oropharyngeal swabs (Mokili et al. 2013), bronchoalveolar lavage (BAL) and coral (unpublished data). Each sample should be processed immediately upon collection especially when microbial metagenomics and metatranscriptomics studies are desired. If the samples were frozen, it limits the isolation of intact microbial cells for microbial metagenomes as freezing potentially disrupt the cell integrity. However, freezing does not preclude metatranscriptomics and viral isolation, but the quality of RNA and amount of viral particles recovered may be affected through the freeze-thaw process. It is important to note that induced sputum has served as the primary source of samples in many studies associated with adult CF patients and other chronic pulmonary diseases (Henig et al. 2001; Rogers et al. 2006) as BAL can be too invasive. In our studies, sputum samples were collected with a careful and consistent sampling method, i.e. following mouthwash and rinsing of the oral cavity using sterile saline solution to keep oral microbes contamination within the sputum samples to a minimum.



1. Sample Collection and Pre-treatment

Homogenization and aliquot (15 min)

2. Viral Metagenome	3. Microbial Metagenome	4. Metatranscriptome
Mucus dissolution Step 2.2 (1.5 hour)	Mucus dissolution Step 3.2.2 (2 hour)	Mechanical lysis Step 4.1.1 (10 min)
VLPs enrichment & purification Step 2.3 (5 hour)	Microbial cells enrichment & purification Step 3.2.4 (1.5 hour)	
	Extracellular DNA removal Step 3.2.7 (2.5 hour)	
DNA extraction Step 2.4 (5 hour)	DNA extraction Step 3.2.14 (< 1 hour)	RNA extraction Step 4.1.4 (2 hour*)
DNA amplification Step 2.5 (*)		DNase I treatment Manufacturer protocol
VLPs visualization Step 2.6		rRNA removal Manufacturer protocol

Library Preparation and Sequencing

* Variable, depending on the chosen methods

Figure 2.1: Workflow for the preparation of host-associated samples, such as sputum sample, for virome, microbiome, and metatranscriptome sequencing.

Protocol

Induced sputum samples were collected in accordance with the University of California Institutional Review Board (HRPP 081500) and San Diego State University Institutional Review Board (SDSU IRB#2121), by the research coordinator of the University of California, San Diego (UCSD) adult CF clinic.

1. Sample collection and pre-treatment (Pre-treat samples within 30 min after collection)

1.1) Prior to sample collection, label four 15 ml tubes as: (i) Viral metagenome, (ii) Microbial metagenome, (iii) Metatranscriptome, and (iv) Extra sputum. Repeat for each sample. Add 2 ml of 0.1 mm silica beads into the tube labeled “Metatranscriptome”, followed by 6 ml of guanidine isothiocyanate-based RNA lysis buffer (GITC-lysis buffer).

1.2) During sample collection, use sterile saline solution (60 ml) as a mouth rinse to minimize contamination by oral microbes. Collect sputum samples over a thirty-minute time period after the inhalation of four milliliters of 7% hypertonic saline via a nebulizer. Process samples immediately, as described below.

1.3) Dilute the sample to a total volume of 8 ml.

1.3.1) Estimate sample volume by weighing empty sputum cup before and after sample collection.

1.3.2) If the volume of sample is less than 8 ml, add appropriate amount of 0.02 µm-filtered 1X PBS to generate a total sample volume of at least 8 ml.

1.3.3) Immediately homogenize the sample with a 3 ml syringe until no visible clumps remain within the sputum

1.3.4) Using the same syringe, draw up 2 ml of sputum and proceed immediately to step 1.4.

1.4) Preserving total RNA from sputum sample

1.4.1) Inject 2 ml sputum sample from step 1.3.4 into the “Metatranscriptome” tube containing silica beads and GITC-lysis buffer.

1.4.2) Close the lid and seal the tube securely with parafilm to avoid leakage.

1.4.3) Homogenize the sputum immediately at medium speed for 10 min. Depending on the vortexer available, place the tube horizontally and secure with tape if necessary.

1.4.4) Keep the tube at 4 °C or in an ice box and transport to the laboratory if necessary.

1.5) Using the same syringe, aliquot 2 ml of sputum each into the tubes labeled “Viral metagenome” and “Microbial metagenome” and transfer the remaining sputum from the sputum cup into the tube labeled “Extra sputum”.

1.6) Store all tubes at 4 °C or ice box and transport to the laboratory if necessary.

2. **Generating viral metagenome**

2.1) Preparation of buffers and solutions

2.1.1) Prepare 50 mM dithiothreitol (DTT) in advance and store at 4 °C. This is stable for 2 weeks.

2.1.2) Prepare Saline Magnesium (SM) buffer (250 ml): 1 M NaCl, 10 mM $MgSO_4$, 50 mM Tris-HCL; adjust pH to 7.4. Filter sterilize (0.02 μm pore size) and store at room temperature.

2.1.3) Prepare DNase I enzyme to 100 U μl^{-1} (in molecular grade water) from lyophilized bovine pancreas DNase I according to the activity defined by Dornase unit/mg dry weight.

2.1.4) Prepare 10X DNase I buffer (50 ml): 100 mM $MgCl_2$, 20 mM $CaCl_2$; adjust pH to 6.5. Filter sterilize (0.02 μm pore size) and store at room temperature.

2.1.5) Prepare 4% paraformaldehyde.

2.1.6) Prepare 200X TE Buffer: 2 M Tris-HCl (pH 8.5), 0.2 M EDTA. Filter sterilize (0.02 μm pore size) and store at room temperature.

2.1.7) Prepare 10 ml of 10% Sodium Dodecyl Sulfate (SDS) using molecular grade water.

2.1.8) Prepare 50 ml CTAB/NaCl (10% CTAB. 700 mM NaCl) using molecular grade water. Dissolve CTAB overnight. If precipitates persist, heat up the solution at 65 °C. Solution is highly viscous in room temperature.

NOTE: The filtration of buffers using 0.02 um filter allows the removal of viral-like particles within the solution, but not free nucleic acid contamination.

2.2) Sample pre-treatment

2.2.1) Prepare appropriate amount of fresh 6.5 mM dithiothreitol (DTT).

2.2.2) Dilute the homogenate by adding 0.02 µm-filtered SM buffer to generate a total volume of 6 ml.

2.2.3) To aid in mucus dissolution, add equal volume (6 ml) of 6.5 mM dithiothreitol (DTT) to the sample, vortex vigorously to mix and incubate for 1 hour at 37 °C.

2.2.4) Vortex the treated sample vigorously and spin at 10 °C, 3,056 x g for 15-20 min.

2.2.5) Collect the supernatant into a new 15 ml tube.

2.2.6) Repeat step 2.2.3 to 2.2.5 for the next sample.

2.2.7) Transfer and filter the supernatant with a 0.45 µm filter mounted on a syringe into a new 15 ml tube.

Note: If the filter clogs, retrieve the samples from the syringe and omit the filtration step.

2.2.8) Take a 100 µl subsample of the 0.45 µm – filtered sample, perform chloroform and DNase I treatment (see section 2.3.12 – 2.3.15) and add equal volume of 4% paraformaldehyde to fix the sample for epifluorescence microscopy (Figure 2.2A).

2.2.9) For a "catch-all" viral particles enrichment approach (see Discussion), go to step 2.3.12 to omit viral particles selection based on cesium chloride gradient ultracentrifugation. However, this may result in chloroform-resistant bacterial contamination and higher amount of host-DNA in the viral lysate.

2.3) Viral-like particles (VLPs) enrichment and purification

2.3.1) Prepare individual cesium chloride (CsCl) solutions by dissolving the appropriate amount of CsCl with non-filtered SM buffer to the desired density (1.7 g ml^{-1}, 1.5 g ml^{-1}, 1.35 g ml^{-1}, and 1.2 g ml^{-1}). Filter each solution through a 0.02 µm filter prior to use.

2.3.2) Set up CsCl gradient as shown in Figure 2.2B.

2.3.3) Load 1 ml of 1.7 g ml^{-1} into each tube, load 1 ml of 1.5 g ml^{-1} into each tube, load 1 ml of 1.35 g ml^{-1} into each tube, load 1.2 g ml^{-1} into each tube (optional), and finally load 6-8 ml sample into the respective tube. Mark individual layers to denote the location of each fraction.

2.3.4) Balance each opposing pair of tubes to within 1 mg.

2.3.5) Carefully load each tube into the spin bucket. Spin all buckets even if they are empty. Load the spin bucket onto the rotor.

2.3.6) Centrifuge at 82,844 x g at 4 °C for 2 hours.

2.3.7) Following centrifugation, carefully remove the tubes from the holder without disrupting the density gradients.

2.3.8) Using a 3 ml syringe with an 18 Ga needle, pierce the tube just below the 1.5 g ml^{-1} density layer (red arrow) and pull ~ 1.5 ml into the syringe.

2.3.9) Collect the upper fraction by slowly removing the needle and allowing the remaining fraction in the tube to drip into a new 15 ml tube through the puncture. Label this as “upper fraction waste”.

2.3.10) Collect the 1.5 g ml^{-1} fraction (containing VLPs) from the syringe by ejecting the contents into two new microfuge tubes.

2.3.11) Repeat step 2.3.8 – 2.3.10 for all samples.

2.3.12) Add 0.2 volume of chloroform into the viral concentrate, shake vigorously, incubate at RT for 10 min, spin at max speed for 5 min, and collect the aqueous phase.

2.3.13) Add 10X DNase buffer and DNase I (final concentration = 2.5 U μl^{-1}) into the chloroform-treated viral concentrate, and incubate at 37 °C for 1.5 – 2 hours.

2.3.14) Inactivate the DNase activity at 65 °C for 15 min.

2.3.15) Remove 15 μl of the chloroform- and DNase I-treated viral fraction into a new tube, and add 15 μl 4% paraformaldehyde to fix the sample for epifluorescence microscopy.

2.4) DNA extraction

2.4.1) Pool viral concentrates from each sample into a cleaned and autoclaved 50 ml Oak Ridge high-speed centrifuge tube.

2.4.2) Add the following: 0.1 volume 200X TE buffer, 10 μl 0.5 M EDTA per ml of sample, 1 volume of formamide, and 10 μl glycogen. Mix well and incubate at room temperature for 30 min.

2.4.3) Using the new volumes, add 2 volumes of room temperature 100% ethanol. Mix well and incubate at 4 °C for at least 30 min.

2.4.4) Pellet DNA by spinning the tube at 17,226 x g for 20 min, at 4 °C using a SS-34 rotor.

2.4.5) Discard the supernatant carefully by using a serological pipette. Wash the pellet twice with ice-cold 70% ethanol.

2.4.6) Remove as much liquid as possible and allow the pellet to air-dry at room temperature for 15 min.

2.4.7) Resuspend the DNA pellet in 567 µl of 1X TE buffer (pH 8.0).

NOTE: Allow at least 15 min for complete resuspension at room temperature. Store the resuspended DNA overnight at 4 °C until further processing.

2.4.8) Transfer the entire 567 µl of resuspended DNA solution into a new 1.5 ml microfuge tube. Add 30 µl of pre-warmed 10% SDS and 3 µl of proteinase K (20 µg ml^{-1}), mix thoroughly and incubate for 1 hour at 56 °C. Pre-warm CTAB/NaCl at 65 °C.

2.4.9) Add 100 µl of 5 M NaCl and mix thoroughly. Add 80 µl of pre-warmed CTAB/NaCl solution, vortex, and incubate for 10 min at 65 °C.

2.4.10) Add equal volume of chloroform, vortex to mix, and spin at 16,100 x g for 5 min.

2.4.11) Transfer the supernatant to a new 1.5 ml microfuge tube. Add an equal volume of phenol/chloroform, vortex to mix, and spin at 16,100 x g for 5 min.

2.4.12) Transfer the supernatant to a new 1.5 ml microfuge tube. Add an equal volume of chloroform, vortex to mix, and spin at 16,100 x g for 5 min.

2.4.13) Transfer the supernatant to a new 1.5 ml microfuge tube. Add equal volume of isopropanol to the supernatant fraction, mix, and incubate at -20 °C for at least 30 min.

2.4.14) Pellet the DNA, spin at 16,100 x g for 15 min at 4 °C. Pipette off the supernatant carefully and wash the pellet twice with ice-cold 70% ethanol.

2.4.15) Perform a short spin and remove the remaining ethanol from the tube. Air dry the pellet for 15 min.

2.4.16) Resuspend the DNA pellet with 50 µl of elution buffer (5 mM Tris, pH 8.5). Allow the pellet to rehydrate for at least 5 min in room temperature.

2.4.17) Quantify the DNA using a high-sensitivity fluorescence-based assay.

2.5) Amplification using Phi29 polymerase (Optional)

2.5.1) Prepare 2X annealing buffer: 80 mM Tris-HCl (pH 8.0), 20 mM $MgCl_2$.

2.5.2) Dilute Phi29 DNA polymerase to 5 U μl^{-1}.

2.5.3) Pre-mix sample buffer, comprised of 50 µl random hexamer primer (100 µM), 125 µl 2X Annealing buffer, and 25 µl water. Aliquot and store at -20 °C.

2.5.4) Pre-mix reaction buffer, comprised of 100 µl Phi29 10X buffer, 40 µl 10 mM dNTPs, and 560 µl water. Aliquot and store at -20 °C.

2.5.5) Add 1 µl template DNA into 4 µl sample buffer.

2.5.6) Incubate the mixture at 95 °C for 3 min and cool on ice.

2.5.7) Add 14 µl of reaction buffer into the mixture from 2.4.4, mix by pipetting up and down.

2.5.8) Add 1 µl of Phi29 DNA polymerase, mix by pipetting up and down, and incubate at 30 °C for 18 hours followed by 65 °C for 10 min.

2.5.9) Clean up the reactions using genomic DNA columns or phenol/chloroform and ethanol precipitations.

2.6) Epifluorescence Microscopy (Refer to Haas et al. 2014) for filtration system set-up)

Note: Following isolation and purification, epifluorescence microscopy with nucleic acid dyes can be used to verify the presence and purity of viral particles in samples (Figure 2.2A; Figure 2.2C). Free DNA in the sample can give rise to high background fluorescence. Therefore, the sample should be DNase I-treated prior to fixation and staining for micrographs.

2.6.1) Prepare mount solution (0.1% ascorbic acid, 50% glycerol). Add 100 µl of 10% ascorbic acid to 4.9 ml of 1X phosphate buffered saline (PBS), mix thoroughly. Add 5 ml of 100% glycerol to the mixture, mix thoroughly and label the tube as “mount”.

2.6.2) Filter mount using a 0.02 µm alumina matrix disposable syringe filter, aliquot into microfuge tubes, and store at -20 °C.

2.6.3) Aliquot 100 µl of sample into a new microfuge tube and add equal volume of 4% paraformaldehyde to fix the VLPs. Incubate the mixture at room temperature for at least 10 min.

2.6.4) Make up the volume to 1 ml by adding 800 µl of 0.02 µm-filtered water. Add 1 µl of SYBR-Gold stain into the tube and incubate at room temperature for 10 minutes.

2.6.5) Set up the filtration system by turning on the vacuum pump between -9 and -10 psi (-62.1 and -68.9 kPa).

2.6.6) Wash the pedestal with water and place a 0.02 µm alumina matrix filter with annular polypropylene support ring into the filter pedestal.

2.6.7) Place a filter tower on top of the filter pedestal with the filter and secure with a clamp.

2.6.8) Pipette the contents from the 1.5 ml microfuge tube into the filter tower, and allow a few minutes for sample to filter through.

2.6.9) Leave the vacuum on while removing the filter tower and clamp.

2.6.10) Label and pipette 10 µl of mount reagent into a microscopic slide.

2.6.11) Carefully remove the filter from the filter pedestal and blot the bottom of the filter with a Kimwipe, then place the filter directly on top of the mount on the microscopic slide.

2.6.12) Pipette another 10 µl of mount reagent onto the filter and place a coverslip over the filter.

3. Generating microbial metagenome

3.1) Preparation of buffers and solutions

3.1.1) Prepare 50 ml of 1X DNase buffer: 50 mM NaAc, 10 mM $MgCl_2$, 2 mM $CaCl_2$; adjust pH to 6.5. Filter sterilize (0.22 µm) and store at room temperature.

3.1.2) Prepare DNase I enzyme to 1000 U μl^{-1} (in molecular grade water) from lyophilized bovine pancreas DNase I according to the activity defined by Dornase unit/mg dry weight.

3.1.3) Prepare 100 ml of SE buffer: 75 mM NaCl, 25 mM EDTA; adjust pH to 7.5. Filter sterilize (0.22 µm) and store at room temperature.

3.2) Sample pre-treatment prior to DNA extraction

3.2.1) Dilute the homogenate by adding 5 volumes of 0.22 µm-filtered 1X PBS. For example, add 10 ml of 1X PBS into 2 ml of sample.

3.2.2) Add β-mercaptoethanol to 2% (v/v) final concentration. Rock the mixture (in the chemical hood) at room temperature for 2 hours.

3.2.3) Spin the sample at 10 °C and 3,056 x g for 15 min, and discard supernatant.

3.2.4) Resuspend the pellet in 10 ml of molecular grade water (or 0.22 µm filtered water), and incubate at room temperature for 15 min.

3.2.5) Repeat steps 3.2.3 and 3.2.4 once.

3.2.6) Spin at 10 °C and 3,056 x g for 15 min, and discard supernatant.

3.2.7) Resuspend the pellet in 5 ml 1X DNase buffer and add 15 µl DNase I (1,000 U μl^{-1}) per ml of sample.

3.2.8) Incubate at 37 °C with repeated mixing for 2 hours.

3.2.9) Spin at 10 °C and 3,056 x g for 15 min, and discard supernatant. Resuspend the pellet in 10 ml SE buffer.

3.2.10) Repeat step 3.2.9.

3.2.11) Spin at 10 °C and 3,056 x g for 15 min, and discard supernatant.

3.2.12) Resuspend the pellet in 2 ml SE buffer, and transfer to two microfuge tubes.

3.2.13) Pellet the cells in the microfuge rubes. Spin the tubes at 16,100 x g at room temperature for 15 min.

3.2.14) Remove the supernatant and extract DNA from the pelleted cells using a genomic DNA extraction kit, Gram-positive variation protocol.

4. Generating metatranscriptome

4.1) Sample pre-treatment

4.1.1) Perform mechanical lysis of cells by bead beating in GITC-lysis buffer immediately after sample collection and homogenization. See step 1.4.

4.1.2) Spin the mixture at 4 °C and 600 x g for 5 min to pellet the silica beads.

4.1.3) Transfer the supernatant into a new tube.

4.1.4) Add 200 µl of chloroform for every 750 µl of GITC-lysis buffer used, shake vigorously by hand for 15 seconds, incubate at room temperature for 10 minutes, and spin at 4 °C and 3,056 x g for 15 min. During this 15 min spin, prepare for step 4.2.

4.1.5) After the 15 min spin (a clear separation of aqueous phase-interphase-organic phase forms), extract the aqueous phase (without disrupting the interphase) into new RNase-free tube(s).

NOTE: The aqueous phase contains the RNA. Keep the tubes on ice until the next step.

4.2) Perform total RNA extraction and purification using commercially available column-based RNA purification kits or conventional isopropanol-based RNA precipitation.

4.2.1) Silica column-based RNA purification

4.2.1.1) Measure the total volume of the aqueous fraction obtained.

4.2.1.2) Add appropriate volume of RNA-binding buffer to sample and mix well.

4.2.1.3) Adjust mixture to appropriate binding condition according to manufacturer's protocol. Mix well and do a short spin.

4.2.1.4) Load the mixture into the RNA-column. For a large volume sample, use multiple loading and load each column up to 4 times. Otherwise, consider using multiple columns for each sample.

4.2.1.5) Wash the column appropriately according to the manufacturer's protocol.

4.2.1.6) Elute the RNA with at least 30 µl of RNase-free water. Double-elution will slightly increase the yield of RNA. However, this will dilute the RNA concentration.

4.2.1.7) Measure the RNA concentration and proceed directly to DNase I treatment. Use the Bioanalyzer to check the quality of the RNA (recommended).

4.2.2) RNA precipitation

4.2.2.1) Add an equal volume of isopropanol (*e.g.*, 500 µl of isopropanol into 500 µl of aqueous fraction) and 2 µl of 10 µg μl^{-1} RNase-free glycogen to the sample.

4.2.2.2) Incubate the mixture at room temperature for 10 min.

4.2.2.3) Spin at 12,000 x g and 4 °C for 15 min.

4.2.2.4) Carefully remove the supernatant, add 1 ml of RNase-free 75% ethanol. Spin the mixture at 7,500 x g and 4 °C for 5 min to make sure the pellet is intact.

4.2.2.5) Carefully remove the ethanol.

4.2.2.6) Repeat steps 4.2.2.4 and 4.2.2.5 once.

4.2.2.7) Air dry the pellet for 10 min.

4.2.2.8) Rehydrate the pellet in 50 µl of RNase-free water, incubate at 55 °C for 5 min and proceed directly to DNase treatment. Use the Bioanalyzer to check the quality of the RNA (Recommended).

4.2.2.9) Store RNA in aliquots at -20 °C, or -80 °C for long-term storage.

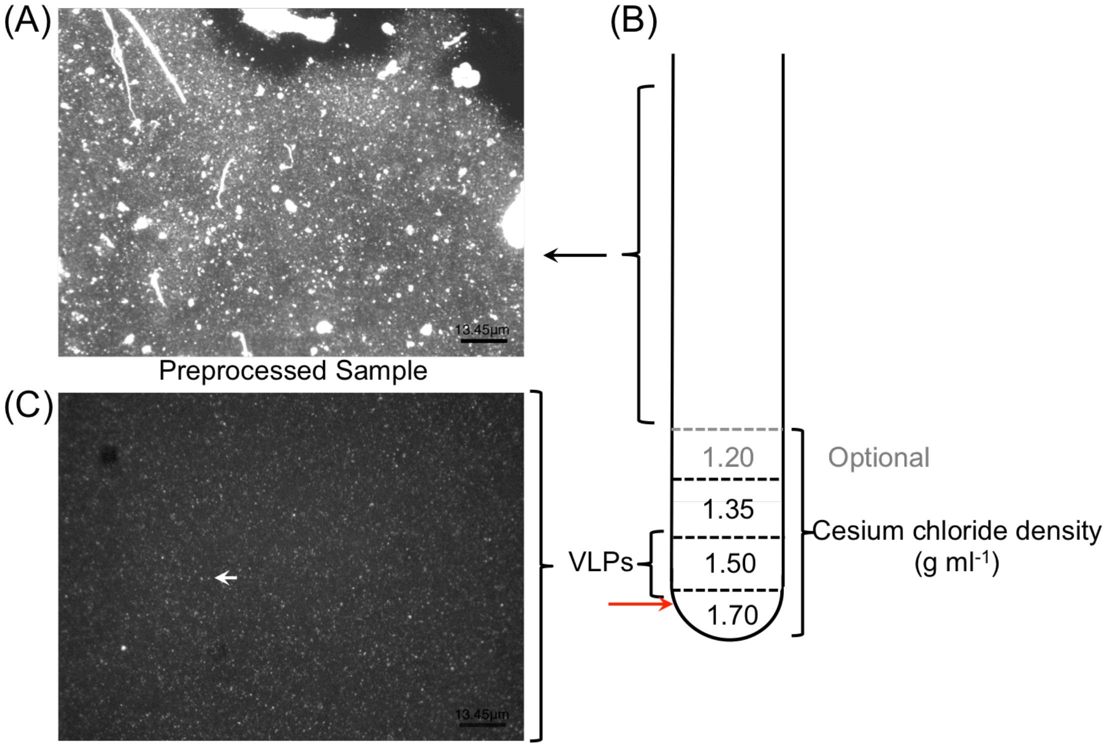


Figure 2.2: Cesium chloride density gradients ultracentrifugation facilitate the elimination of extracellular DNA and large particles (A), and allow for optimal isolation of viral-like particles from CF sputum. One milliliter of each gradient is layered on top of each other prior to loading the pre-treated sample (B). Following particles isolation and purification, epifluorescence microscopy with nucleic acid dyes such as SYBR Gold are used to verify the presence and purity of viral particles in samples. Clear viral-like particles (C; white arrow) were observed following the density gradient separation of CF sputum sample.

Results

Viral Metagenomes

CF sputum is exceptionally viscous and contains a high amount of mucin and free DNA (Figure 2.2A); the density gradient ultracentrifugation facilitates the elimination of host-derived DNA (Figure 2.2B). The results from a previous study (Lim et al. 2012) showing eight viromes generated from the presented workflow are summarized here (Table 2.1). Seven samples (CF1-D, CF1-E, CF1-F, CF4-B, CF4-C, CF5-A, and CF5-B; Table 2.1) were processed as described in Section 2. The generated viromes contained little (0.02% - 3.7%) human-derived sequences with only one exception (70%). CF4-A was omitted from the density gradient ultracentrifugation step (CF4-A) and the virome generated from this specific sample contained >97% human-derived sequences (Table 2.1). Figure 2 shows an example of the epifluorescence microscopy image of a typical CF sputum sample before (Figure 2.2A) and after (Figure 2.2C) density gradient ultracentrifugation of a typical CF sputum sample. Clear viral-like particles (VLPs) were observed in the micrographs without large particles following the density gradient separation. After VLPs DNA extraction, bacterial contamination is often tested using 16S rDNA amplification prior to the sequencing of VLPs DNA.

Table 2.1: Library characteristics of eight viromes generated from sputum samples using presented workflow. This table is extracted from Lim et al. (2012). Seven samples (CF1-D, CF1-E, CF1-F, CF4-B, CF4-C, CF5-A, and CF5-B) were processed as described in Section 2 and generated viromes that contained little (0.02% - 3.7%) human-derived sequences with one exception (70%). CF4-A was omitted from the density gradient ultracentrifugation step (CF4-A) and generated virome that contained > 97% human-derived sequences.

	CF1-D	**CF1-E**	**CF1-F**	**CF4-A**	**CF4-B**	**CF4-C**	**CF5-A**	**CF5-B**
Total number of reads	**224,859**	**87,891**	**106,189**	**93,301**	**140,020**	**1,558**	**272,552**	**217,438**
Preprocessed reads [a]	**109,389**	**73,624**	**67,070**	**82,011**	**68,617**	**1,137**	**215,808**	**158,432**
	49%	**84%**	**63%**	**88%**	**49%**	**73%**	**79%**	**73%**
Number of bases	47,239,573	33,351,525	28,922,479	27,667,695	29,386,841	243,986	95,205,805	69,581,811
Mean read length	432	453	431	337	428	215	441	439
Host sequences [b]	240	526	28	79,774	13	797	585	5,859
	0.21%	0.71%	0.04%	97.27%	0.02%	70.10%	0.27%	3.70%
Viral hits [c]	**7,214**	**23,550**	**4,070**	**737**	**4,642**	**22**	**6,466**	**5,981**
	6.59%	**31.99%**	**6.07%**	**0.90%**	**6.77%**	**1.93%**	**3.00%**	**3.78%**
Unassigned Reads [d]	**103,888**	**60,490**	**32,780**	**1,935**	**68,440**	**311**	**105,612**	**119,551**
	94.97%	**82.16%**	**48.87%**	**2.36%**	**99.74%**	**27.35%**	**48.94%**	**75.46%**

[a] Reads after data pre-processing by PRINSEQ.

[b] Human reads identified by DeconSeq plus reads with a best BLASTn hit (NCBI nucleotide database) to the phylum Chordata.

[c] tBLASTx hits against in-house viral genome database. The percentage was calculated using the total number of preprocessed reads.

[d] Reads with no BLASTn hit against the NCBI nucleotide database. The percentage was calculated using the total number of preprocessed reads. Some reads with no BLASTn hit against the NCBI nucleotide database were identified as viral at protein level in the tBLASTx analysis.

Microbial Metagenomes

Seven sputum samples presented here were collected from a single CF patient across seven consecutive days. The patient started on oral antibiotic (Ciprofloxacin and Doxycycline) on Day 3 after the sputum was collected. The volume of each sputum sample collected from this patient was 15 ml throughout the seven days; therefore, PBS was not added to the sample. The goal of this sampling event was to evaluate the protocols presented in this workflow by (i) evaluating the daily fluctuation of microbial community structure, and (ii) compare the microbial community structure and resolution between metagenomics and 16S rDNA sequencing. Therefore, total DNA and HL-DNA were extracted from each sample.

The HL-DNA concentration of each sputum sample following DNA extraction is presented in Table 2.2. The total yield of HL-DNA ranged from 210 ng to > 5 μg based. Illumina sequencing libraries were generated with a total starting material of 1 ng for each sample (Figure 2.3). The characteristics of the metagenomics data are presented in Table 2.2. All but one library yielded more than 1 million sequences and more than 85% high quality sequences were retained upon data preprocessing using the PRINSEQ (Schmieder and Edwards 2011a) software. All datasets were first preprocessed to remove duplicates and sequences of low quality (minimum quality score of 25), followed by further screening and removal of human-derived sequences using DeconSeq (Schmieder and Edwards 2011b). The amount of human-derived sequence contamination is highly dependent on the sample properties. Here, the total amount of human-derived sequences ranged from 14% – 46% (Table 2.2). The preprocessed sequences were then annotated using the Metaphlan (Segata et al. 2012) pipeline as well as MG-RAST (Meyer et al. 2008) server.

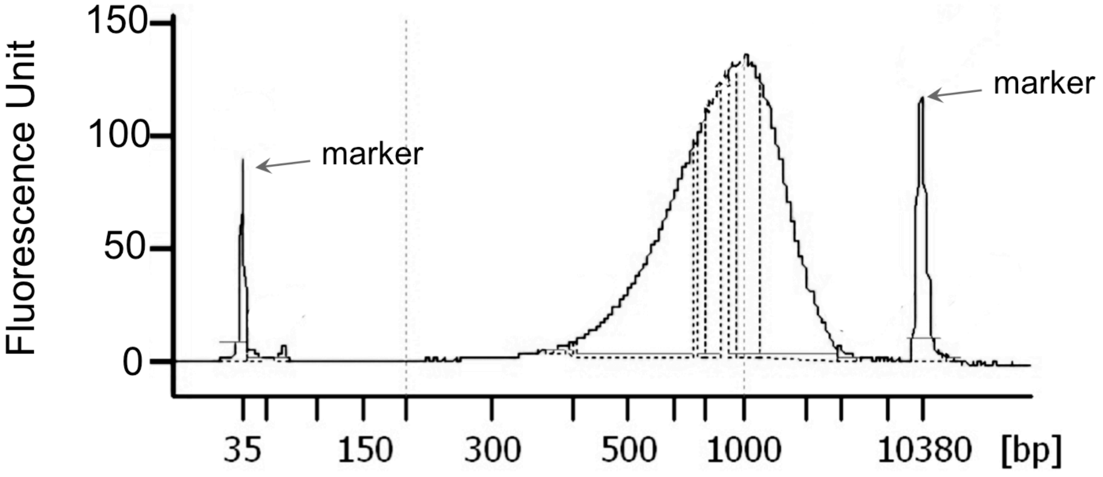


Figure 2.3: Example of the size distribution of Nextera XT libraries generated from 1 ng of HL-DNA that resulted in CF sputum microbiomes. Library normalization, pooling, and loading amount was done as described in the manufacturer protocol without any deviation.

In addition to metagenomes, 16S rDNA amplicon libraries were generated from both the total DNA and HL-DNA via primers targeting approximately 300 bp of the V1-V2 variable region in the 16S rRNA gene (Hara et al. 2012; Markle et al. 2013). PCR products from individual samples were normalized and pooled for sequencing using the Illumina 500-cycle paired-end sequencing performed on the MiSeq platform. Paired-end 16S rDNA amplicon sequences were sorted by sample via barcodes using a python script and the paired reads were assembled using phrap (Ewing et al. 1998; Ewing and Green

1998). Assembled sequence ends were trimmed until the average quality score was ≥ 20 using a 5 nt window. Potential chimeras were then removed using Uchime (Edgar et al. 2011) against a chimera-free subset of the SILVA (Quast et al. 2013) reference sequences. Taxanomy was assigned to the high quality reads with SINA (Pruesse et al. 2012) (version 1.2.11) using the 418,497 bacterial sequences from the SILVA(Quast et al. 2013) database. Sequences with identical taxonomic assignments were clustered to produce Operational Taxonomic Units (OTUs). This process generated 1,655,278 sequences for 16 samples (average size: 103,455 sequences/sample; min: 72,603; max: 127,113). The median Goods coverage score, a measure of completeness of sequencing, was ≥ 99.9%. The software package Explicet (Robertson et al. 2013) (v2.9.4, www.explicet.org) was used for analysis and figure generation. Alpha-diversity (intra-sample) and beta-diversity (inter-sample) were calculated in Explicet at the rarefaction point of 72,603 sequences with 100 bootstrap re-samplings.

The first question targeted by this study was whether hypotonic lysis preferentially selects for (i.e., preferentially retains or lyses) particular groups of microbes. After the first hypotonic lysis, re-suspended pelleted cells were subsampled from the first two samples (CF1-1A* and CF1-2A*) to compare with the same samples after the second hypotonic lysis (CF1-1A and CF1-2A). All samples were treated equally, i.e., treated with DNase I prior to DNA extraction, followed by DNA extraction and the sequencing pipeline. As shown in Figure 2.4, the microbial profiles of the subsamples are highly similar to the samples after two hypotonic lysis treatments. In addition, the second hypotonic lysis increases the fraction of non-human sequences by 6%-17% within the metagenomes (Table 2.2).

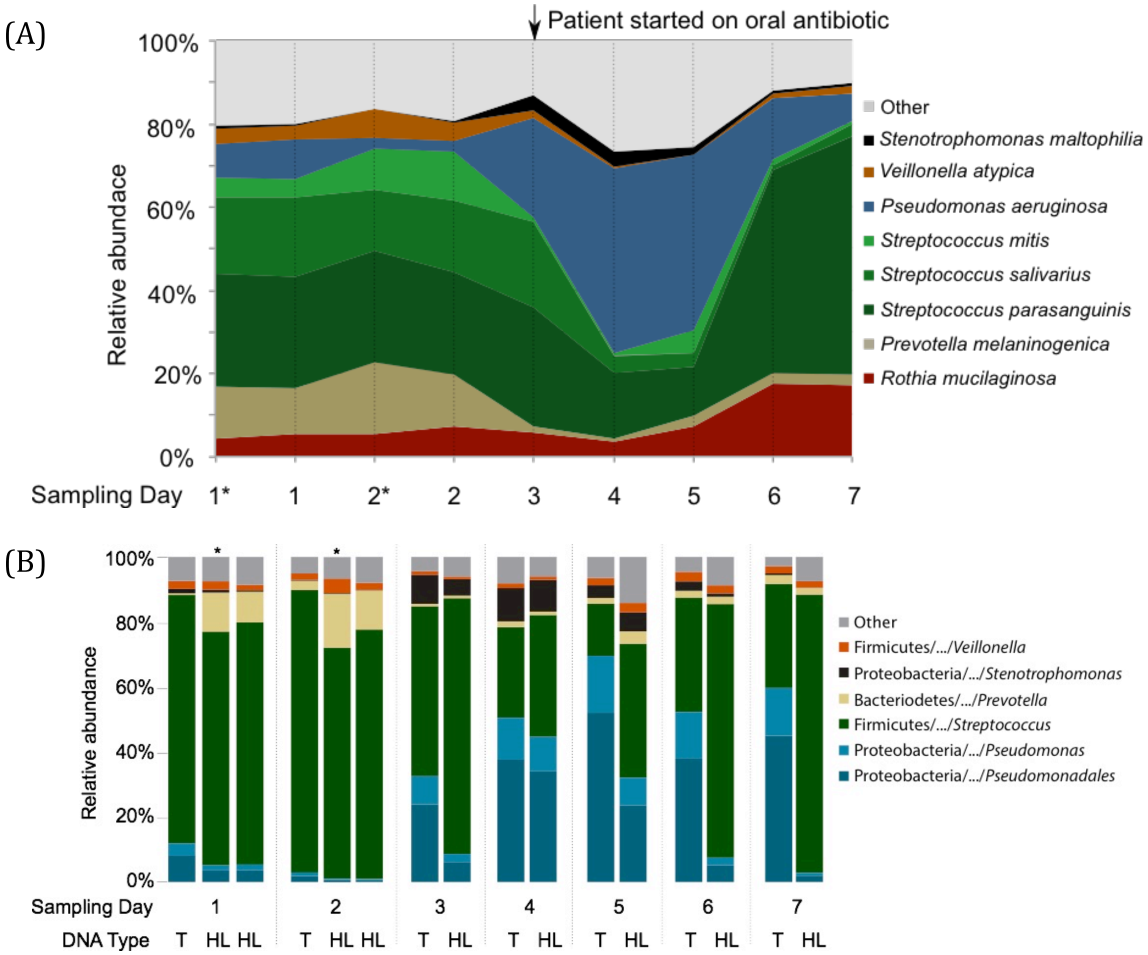


Figure 2.4: Taxonomic analysis of the microbial communities in nine samples collected longitudinally from one CF patient. (A) Microbial profiles based on the metagenomic libraries generated from hypotonic lysis method-based DNA. The species assignment was based on the Metaphlan pipeline following data preprocessing that remove duplicates and sequences with low quality and human sequence homology. In order to show that two-steps hypotonic lysis did not preferentially selects for particular groups of microbes, subsamples (*) after the first hypotonic lysis were included. (B) Microbial profiles based on the V1V2 region of 16S rRNA gene sequencing from total DNA (T) and hypotonic lysis method-based DNA (HL). These data have not been previously published.

Table 2.2: Characteristics of microbiomes generated from sputum samples using presented workflow. The DNA concentration of each sample in 100 µl elution buffer (5 mM Tris/HCl, pH 8.5) and the characteristics of sequence data are presented. A total of 1 ng was used to generate individual library using the Nextera XT library preparation kit.

Sample	Concentration	Total Yield	Total No. Reads	Total No. Reads	Non-Human Sequences
	(ng/µl)	(ng)	(Raw[a])	(Processed[b])	(%)
CF1-1A*	2.3	230	1,098,454	937,688	691,541 74%
CF1-1	13	1,300	2,212,756	1,958,910	1,574,520 80%
CF1-2A*	2.1	210	672,878	588,106	407,530 69%
CF1-2	5.2	520	1,944,012	1,697,010	1,455,174 86%
CF1-3	28.8	2,880	1,048,304	896,756	560,852 63%
CF1-4	24.1	2,410	1,154,922	984,702	621,098 63%
CF1-5	33.6	3,360	1,029,622	888,630	481,548 54%
CF1-6	43.2	4,320	1,434,016	1,256,504	725,858 58%
CF1-7	57.8	5,780	1,000,174	872,036	565,376 65%

* 1 ml of sample was subsampled from CF1-1 and CF1-2 following the first hypotonic lysis step (Step 3.2.4) before the second hypotonic lysis procedure. The cells were spun down as described in 3.2.6 and proceed through the remaining protocol without any modification.

[a] Unprocessed Illumina reads from a 2 X 300 bp MiSeq sequencing run.

[b] Reads were assessed, trimmed, and removed based on quality and length as described in the discussion.

To test for differences in microbial composition between metagenomic- and 16S rDNA-based profiling, and for changes before and after hypotonic lysis that might explain the differences previously seen between our studies and others, bacterial 16S rDNA sequencing libraries were generated from both the total DNA and HL-derived DNA (Figure 2.4B). At genus level, the taxonomy profiles of the common CF-associated bacteria such as *Pseudomonas*, *Stenotrophomonas*, *Prevotella*, *Veillonella*, and *Streptococcus* were highly similar between the 16S rDNA libraries and metagenomes generated from the HL-derived DNA. However, *Rothia* detection in the 16S rDNA libraries was not as abundant as with the metagenomic libraries. When comparing the 16S rDNA taxonomical profiles generated from total DNA and HL-derived DNA, Pseudomonas was differentially represented in the total DNA compared to the HL-derived DNA starting from Day 3.

Metatranscriptomes

Typically, the total RNA extracted from CF sputum is partially degraded and the size ranges from 25 - 4000 bps (Figure 2.5A and Figure 2.5C). Here, the representative results presented was previously published in (Lim et al. 2012). The fraction of rRNA within the non-depleted metatranscriptomes ranges from 27 – 83%, and the relative abundance of rRNA varied across samples (Table 2.3; data extracted from Lim et al.(Lim et al. 2012)). However, depletion with Ribo-Zero kit decreased the rRNAs relative abundance of rRNA to 1 – 5% with the exception of sample CF1-F. The variation in the effectiveness of rRNA removal could reflect the quality of extracted RNA, or differences in the microbial community present and hence the accessibility of rRNAs for probes hybridization (Lim et al. 2012). The electropherograms of a successful (Figure 2.5B) and unsuccessful (Figure 2.5D) rRNA removal procedure using the Ribo-Zero rRNA removal kit differ, at which rRNA peaks are visible in the unsuccessful removal.

The size range of cDNA libraries generated often reflects the size range of the starting RNA sample. The cDNA libraries presented here were generated with a whole transcriptome amplification kit (WTA2) upon rRNA depletion followed by Roche-454 sequencing library preparation (Lim et al. 2012). The cDNA generated contain fragments ranging from 50 - 4,000 bps (Figure 2.5E-F) and is highly consistent across samples (Lim et al. 2012). The availability of other platform-specific RNA-Seq library preparation kits currently provide more alternative options for one to combine cDNA synthesis and sequencing library preparation in optimum conditions. One recommended option to date is the ScriptSeq Complete Gold Kit combining rRNA removal reagents recommended above and RNA-Seq library preparation kit.

Table 2.3: Library characteristics of the metatranscriptomes with and without rRNA depletion. The data is extracted from Lim *et al.* (2012), which has additional comparison of other rRNA removal kits and the effect of cDNA nebulization prior to sequencing library preparation.

Sample	**CF1-D**		**CF1-F**		**CF4-B**		**CF4-C**	
Treatment	**None**	**Ribo-Zero**	**None**	**Ribo-Zero**	**None**	**Ribo-Zero**	**None**	**Ribo-Zero**
Preprocessed reads	**2,088**	**1,991**	**40,876**	**25,238**	**19,728**	**32,737**	**31,791**	**36,172**
Mean read length	275	245	262	270	233	259	240	267
Total rRNA reads	**1,737**	**91**	**29,499**	**17,267**	**5,285**	**291**	**16,371**	**1,761**
	83.20%	**4.60%**	**72.20%**	**68.40%**	**26.80%**	**0.90%**	**51.50%**	**4.90%**
Microbial rRNA	1,414	32	19,978	12,035	23	227	6,916	1,076
	67.70%	1.60%	48.90%	47.70%	0.10%	0.70%	21.80%	3.00%
Eukaryota rRNA	323	59	9,520	5,232	5,262	64	9,455	683
	15.50%	3.00%	23.30%	20.70%	26.70%	0.20%	29.70%	1.90%
% rRNA removed*	**0%**	**95%**	**0%**	**5%**	**0%**	**97%**	**0%**	**91%**
Non-rRNA reads	**351 (16.8%)**	**1,900 (95.4%)**	**11,377 (27.8%)**	**7,971 (31.6%)**	**14,443 (73.2%)**	**32,446 (99.1%)**	**15,420 (48.5%)**	**34,411 (95.1%)**
Total NR hits	**102 (4.9%)**	**691 (34.7%)**	**3,327 (8.1%)**	**2,857 (11.3%)**	**4,938 (25.0%)**	**10,751 (32.8%)**	**5,905 (18.6%)**	**15,766 (43.6%)**
Eukaryotic	74	407	2,790	2,524	4,614	10,227	4,553	8,274
Bacterial	26	283	520	312	287	471	1,326	7,442
Unassigned reads	**249 (11.9%)**	**1,209 (60.7%)**	**8050 (19.7%)**	**5,114 (20.3%)**	**9,505 (48.2%)**	**21,695 (66.3%)**	**9,515 (29.9%)**	**18,645 (51.5%)**

*The amount of rRNA removed expressed as a percentage of the amount present in the non-depleted aliquot.

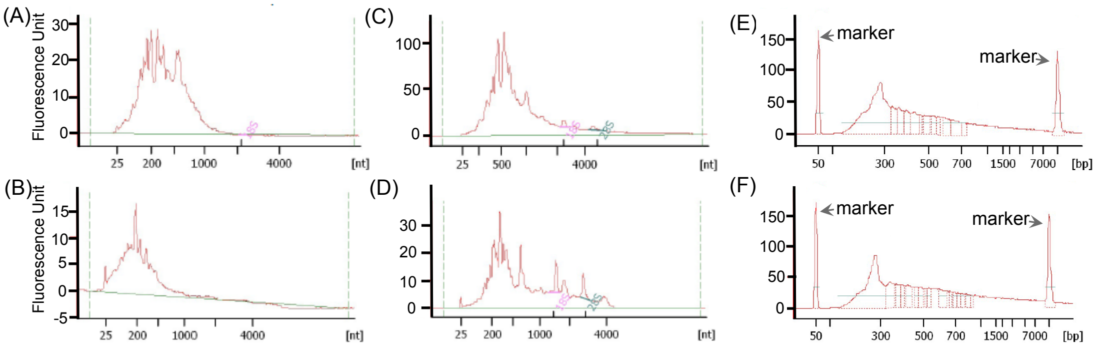


Figure 2.5: Examples of Agilent 2100 Bioanalyzer electropherograms of RNA (A-D) and cDNA (E-F) generated for the metatranscriptomic libraries, using RNA pico and high-sensitivity dsDNA chips respectively. (A) and (C) show the examples of electropherograms before rRNA removal procedures. The electropherograms of a successful (B) and unsuccessful (D) rRNA removal procedure using total rRNA Removal kit differ slightly, at which rRNA peaks are visible in the unsuccessful removal. The size range of cDNA (E-F) generated using the whole-transcriptome amplification kit (Sigma-Aldrich) is similar to the size range of the starting rRNA-depleted RNA, and highly consistent across the two different samples.

Discussion

Viral metagenomics

Viral particles are concentrated using polyethylene glycol (PEG) precipitation or small volume concentrators. In some cases, concentration may not be needed, but pre-filtration or low speed centrifugation steps are used to remove eukaryotic and microbial cells. Viral lysates will be further enriched and purified using density gradient ultracentrifugation (Thurber et al. 2009; Lim et al. 2012) or small size filters (e.g., 0.45 µm) to remove eukaryotic and large microbial cells (Mokili et al. 2013). Density gradient ultracentrifugation is typically performed with dense but inert solutions such as sucrose or cesium chloride to isolate and concentrate viral particles (Thurber et al. 2009). Physical separation is based on the size and buoyant density of viral particles. Therefore, proper choice of filter pore size and the rigorous preparation of gradients are essential to isolate specific viral communities, as the success of the physical recovery of VLPs determines the community isolated (Thurber et al. 2009) (i.e., viral particles that do not pass through the filter or fall within the extraction density will not be detected in the metagenome). After viral isolation and concentration, there may be contaminating non-viral genomic material present in the sample both in the form of free nucleic acids and microbial and eukaryotic cells. Therefore, it is critical to verify the purity of viral particles in samples (Figure 2.1A-2.1B). A chloroform treatment is commonly used to lyse remaining cells, followed by nuclease treatment to degrade free nucleic acids prior to nucleic acid extraction.

A caveat to the presented workflow was the use of density gradient separation to isolate viral particles as it may exclude enveloped viral particles that may be too buoyant to enter the CsCl gradient. An alternative “catch-all” method is to omit the density gradient separation and isolate the community DNA from the 0.45 µm – filtrates treated with chloroform and DNase I. This approach is also appropriate to accommodate small sample volumes such as those from swabs or blood plasma. However, this may result in chloroform-resistant bacterial contamination and higher amount of DNase I-resistant extracellular DNA.

Current sequencing protocols require 1 ng to 1 µg of nucleic acids for sequencing library preparation whereby higher DNA yields provide a wider choice of sequencing options. The DNA concentration of generated viromes often ranges from below the detection limit to more than 200 ng µl-1. The amount of viral nucleic acids recovered may be insufficient for direct sequencing library preparation. In such cases, nucleic acid amplification is essential. Linker amplification shotgun libraries (LASLs) (Breitbart et al. 2002; Henn et al. 2010; Duhaime et al. 2012) and whole genome amplification based on multiple displacement amplification (MDA) are the two methods most commonly used to generate sufficient DNA for sequencing. MDA methods such as those based on Phi29 DNA polymerase are known to suffer from amplification biases, and may preferentially amplify ssDNA and circular DNA, resulting in non-quantitative taxonomical and functional characterization (Yilmaz et al. 2010; Kim and Bae 2011). An optimized version of the LASLs approach has been shown to introduce only minimal biases, promotes higher sensitivity (for small amounts of starting material), and is easily adapted for different sequencing platforms (Duhaime et al. 2012). However, the approach has many steps, requires specialized equipment to minimize DNA loss, and is limited to dsDNA templates. In our laboratory, this approach has been successfully adapted to amplify detectable and undetectable amount of DNA extracted from bronchoalveolar lavage-, coral- and sea water-derived VLPs (unpublished and Hurwitz et al. 2013).

Developing data analysis pipelines has classically been one of the most challenging aspects of viral metagenomics analysis due to the highly diverse and largely unknown nature of the viral communities. While there are an estimated 108 viral genotypes in the biosphere, to date current viral databases contain ~ 4000 viral genomes, which is about 1/100,000th of this approximate total viral diversity. Therefore, similarity-based searches (such as BLAST (Altschul et al. 1990)) for taxonomic and functional assignment in viral metagenomes possess inherent challenges. Many sequences fail to have significant similarities to genomes in the database, and therefore, are classified as unknown. Even though homology-based searches are the most important applications for assigning taxonomy and function to sequence data, alternative approaches based on database-independent analysis have been developed (Angly et al. 2005; Angly et al. 2009; Dutilh et al. 2012). (Fancello et al. 2012) provide a complete review of computational tools and algorithms used in viral metagenomics.

Microbial metagenomics

Typically, the total amount of DNA extracted from hypotonic lysis-treated microbial communities (HL-DNA) range from 20 ng to 5 μg. The yield is highly dependent on the patient's health status and the amount of sputum sample collected, which explains the variations seen in the total yield of HL-DNA extracted in this study

(Table 2.2). The critical steps to generate good quality sequence data rely on the quality of sequencing libraries generated. Figure 2.2 shows a typical size range for the sequencing libraries generated from CF sputum-derived microbial DNA using an enzymatic-based DNA fragmentation procedure. The optimal library size is dependent on the choice of sequencing platform and application, and therefore, the fragmentation procedure can be optimized, if necessary, through alternative approaches such as sonication and nebulization. In addition to the presented representative results, the success of the presented method on CF sputum collected from multiple patients across multiple time points is also illustrated in (Lim et al. 2012) and (Lim et al. 2014).

Previous studies (Lim et al. 2012; Lim et al. 2014) suggest that every patient harbors a unique set of microbial community that shifts over time, thereby reflecting the persistence of the major players within the community while fluctuations are likely due to perturbations such as antibiotic treatments. Whether these fluctuations occur daily even without external perturbations or due to sampling procedure and sample processing, is still in question. Based on the HL-DNA metagenomic and 16S rDNA amplicon analysis, the 7-day longitudinal sampling shows that the daily fluctuation of microbial profiles without antibiotic perturbation (Day 1, 2, and 3) was minimal (Figure 2.3A-2.3B). Upon introduction of oral antibiotics immediately after the Day 3 sampling, changes in the community profile became apparent on Day 4. While the antibiotic ciprofloxacin targets a broad spectrum of known bacterial pathogens such as *P. aeruginosa, Staphylococcus aureus,* and *Streptococcus pneumonia*, the treatment increased the relative abundance of *P. aeruginosa* while decreasing the *Streptococcus* spp. and *P. melaninogenica*. By Day 6, the community slowly recovered to the initial starting community structure. The results suggest that fluctuations of microbial profiles within a single patient are more likely due to community perturbations in the airways.

Given the consistency between the microbial profiles of 16S rDNA libraries and metagenomes from HL-derived DNA, we ruled out the biases originating from the 16S rRNA primers used in this study. One possible explanation for the differences seen across 16S rDNA taxonomical profiles generated from total DNA and HL-derived DNA (Figure 2.3B) may be the presence of high amounts of *Pseudomonas* spp. extracellular DNA after the antibiotic treatment. This is supported by the findings that these differences were most apparent at Day 7, three days after the antibiotics treatment, which targets Pseudomonas spp. in addition to others. Ciprofloxacin is commonly used as the first-line treatment in patients with CF and chronic *P. aeruginosa* infection even though its spectrum of activity includes most CF-associated pathogens. We hypothesized that the antibiotic treatment eradicates susceptible communities including *Streptococcus* spp. and hence creating a niche filled by resistant *P. aeruginosa*. *Pseudomonas aeruginosa* may gain resistance through increasing its biofilm communities and extracellular DNA has been shown to be the main structural support of its biofilm architecture (Allesen-Holm et al. 2006). Even as the community structure recovered, extracellular DNA may have remained in the CF sputum. Therefore, these data suggest that hypotonic lysis and the washing steps presented in this workflow potentially benefit in not only removing the human-derived DNA, but also microbial-derived extracellular DNA that may misrepresent the actual microbial profiles.

Metatranscriptomics

A high quality metatranscriptome should contain relatively few ribosomal RNA (rRNA) sequences and represent an unbiased sampling of the community transcripts (mRNA). Due to the short half-life and limited amount of mRNA, it is critical that the protocol, as presented here, minimizes sample handling to maximize the number of transcripts recovered.

In recent years, several approaches for rRNA depletion have been developed and adapted in commercially available kits. These include MICROB*Enrich*, Ribo-Zero, and sample-specific subtractive hybridizations (Stewart et al. 2010) that are based on oligonucleotide hybridization, and the mRNA-ONLY kit that is based on exonuclease enzymatic activity targeting RNA containing a 5' monophosphate. In addition, several approaches for mRNA enrichments such as the MessageAmp II-Bacteria Kit that preferentially polyadenylates and amplifies linear RNA are also available. Some of these methods (e.g., mRNA-ONLY, MICROB*Express* and the MessageAmp) are used concurrently for optimal efficiency. However, the efficacy of all of these approaches are limited, especially when working with partially degraded rRNA, as often observed in total RNA extracted from CF samples. Polyadenylation-dependent RNA amplification cannot be used to generate metatranscriptomes consisting of both eukaryotic and prokaryotic mRNA. In addition, the poly(A) tail added to the sequences may reduces the amount of useful sequence data. Regions with homopolymer stretches will tend to have lower quality scores, causing a significant number of reads to be filtered out by sequencing and post-sequencing software, and the average useful read length after trimming off poly (A) tails will be reduced significantly (Frias-Lopez et al. 2008).

Dealing with complex CF microbial communities and partially degraded RNA (Figure 2.4A and Figure 2.4C), our previous study showed that the hybridization-capture method by the Ribo-Zero Gold kit was more effective in removing both human and microbial rRNA compared to the combine treatments using other kits (Lim et al. 2012) (Table 2.3). The resultant data allows concurrent analysis of both human host and microbial transcripts. Depending on the yield and quality of RNA, as well as ultimate choice of sequencing platform, many of these processes including cDNA synthesis can be streamlined with sequencing library generation. For example, Ribo-Zero treated RNA can be used to make metatranscriptome sequencing libraries using ScriptSeq RNA-Seq Library Preparation kit.

Metagenomic analysis of animal-associated communities provides a comprehensive representation of the overall functional entity that includes the host and its associated communities. The workflow presented here is adaptable to a variety of complex animal-associated samples, especially those that contain thick mucus, high amounts of cell debris, extracellular DNA, protein and glycoprotein complexes, as well as host cells in addition to the desired viral and microbial particles. Even though viral and microbial particles may be lost at every step, particles isolation and purification are essential to minimize the amount of host DNA. While the metagenomics data provides metabolic potentials of the communities examined, metatranscriptomics complement this by revealing the differential expression of encoded functions (Lim et al. 2012). A comprehensive assessment of the genomics and transcripts data has yielded new insights to the dynamics of community interactions and facilitates the development of improving therapies (Lim et al. 2012; Lim et al. 2013; Lim et al. 2014).

Acknowledgments

This work was supported by the National Institute of Health (1 R01 GM095384-01) awarded to Forest Rohwer. We thank Epicentre, an Illumina company for providing early access to Ribo-Zero Epidemiology kits. We thank Mark Hatay for the design and production of the ultracentrifugation tube holder. We thank Andreas Haas and Benjamin Knowles for critical readings and discussions of the manuscript, and Lauren Paul for assisting the filming process.

Chapter 2, in full, is published in the Journal of Visualized Experiments. Yan Wei Lim, Matthew Haynes, Mike Furlan, Charles E. Robertson, J. Kirk Harris, and Forest Rohwer; 2014. The dissertation author was the primary investigator and author of this paper. Article Link: http://www.jove.com/video/52117/purifying-impure-sequencing-metagenomes-metatranscriptomes-from

References

Allesen-Holm M, Barken KB, Yang L, Klausen M, Webb JS, Kjelleberg S, Molin S, Givskov M, Tolker-Nielsen T (2006) A characterization of DNA release in Pseudomonas aeruginosa cultures and biofilms. Mol Microbiol 59:1114–1128. doi: 10.1111/j.1365-2958.2005.05008.x

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6:41. doi: 10.1186/1471-2105-6-41

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F (2009) The GAAS metagenomic tool and Its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol. doi: 10.1371/journal.pcbi.1000593

Bibby K (2014) Improved bacteriophage genome data is necessary for integrating viral and bacterial ecology. Microb Ecol 67:242–244. doi: 10.1007/s00248-013-0325-x

Bomar L, Maltz M, Colston S, Graf J (2011) Directed culturing of microorganisms using metatranscriptomics. mBio 2:e00012–11. doi: 10.1128/mBio.00012-11

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci 99:14250–14255. doi: 10.1073/pnas.202488399

Breitenstein S, Tümmler B, Römling U (1995a) Pulsed field gel electrophoresis of bacterial DNA isolated directly from patients' sputa. Nucleic Acids Res 23:722–723.

Breitenstein S, Tümmler B, Römling U (1995b) Pulsed field gel electrophoresis of bacterial DNA isolated directly from patients' sputa. Nucleic Acids Res 23:722–723.

Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, Bushman FD, Collman RG (2011) Topographical continuity of bacterial populations in the healthy human respiratory tract. Am J Respir Crit Care Med 184:957–963. doi: 10.1164/rccm.201104-0655OC

Childs WC, Gibbons RJ (1988) Use of percoll density gradients for studying the attachment of bacteria to oral epithelial cells. J Dent Res 67:826 –830. doi: 10.1177/00220345880670050601

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38:e200. doi: 10.1093/nar/gkq873

Duhaime MB, Deng L, Poulos BT, Sullivan MB (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 14:2526–2537. doi: 10.1111/j.1462-2920.2012.02791.x

Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, Edwards RA, Mokili JL (2012) Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinforma Oxf Engl 28:3225–3231. doi: 10.1093/bioinformatics/bts613

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinforma Oxf Engl 27:2194–2200. doi: 10.1093/bioinformatics/btr381

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185.

Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434:162–174. doi: 10.1016/j.virol.2012.09.025

Fodor AA, Klem ER, Gilpin DF, Elborn JS, Boucher RC, Tunney MM, Wolfgang MC (2012) The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. PLoS ONE 7:e45001. doi: 10.1371/journal.pone.0045001

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci 105:3805 –3810. doi: 10.1073/pnas.0708897105

Haas AF, Knowles B, Lim YW, McDole Somera T, Kelly LW, Hatay M, Rohwer F (2014) Unraveling the unseen players in the ocean - a field guide to water chemistry and marine microbiology. J Vis Exp JoVE e52131. doi: 10.3791/52131

Hara N, Alkanani AK, Ir D, Robertson CE, Wagner BD, Frank DN, Zipris D (2012) Prevention of virus-induced type 1 diabetes with antibiotic therapy. J Immunol Baltim Md 1950 189:3805–3814. doi: 10.4049/jimmunol.1201257

He S, Kunin V, Haynes M, Martin HG, Ivanova N, Rohwer F, Hugenholtz P, McMahon KD (2010) Metatranscriptomic array analysis of "Candidatus Accumulibacter phosphatis"-enriched enhanced biological phosphorus removal sludge. Environ Microbiol 12:1205–1217. doi: 10.1111/j.1462-2920.2010.02163.x

Henig NR, Tonelli MR, Pier MV, Burns JL, Aitken ML (2001) Sputum induction as a research tool for sampling the airways of subjects with Cystic Fibrosis. Thorax 56:306–311. doi: 10.1136/thorax.56.4.306

Henn MR, Sullivan MB, Stange-Thomann N, Osburne MS, Berlin AM, Kelly L, Yandava C, Kodira C, Zeng Q, Weiand M, Sparrow T, Saif S, Giannoukos G, Young SK, Nusbaum C, Birren BW, Chisholm SW (2010) Analysis of high-throughput sequencing and annotation strategies for phage genomes. PLoS ONE 5:e9083. doi: 10.1371/journal.pone.0009083

Hurwitz BL, Deng L, Poulos BT, Sullivan MB (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environ Microbiol 15:1428–1440. doi: 10.1111/j.1462-2920.2012.02836.x

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the Cystic Fibrosis gene: Genetic analysis. Science 245:1073–1080. doi: 10.1126/science.2570460

Kim K-H, Bae J-W (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol 77:7663–7668. doi: 10.1128/AEM.00289-11

Kleven D, McCudden C, Willis M (2008) Cystic Fibrosis: Newborn screening in America.

Lee J-L, Levin RE (2006) Use of ethidium bromide monoazide for quantification of viable and dead mixed bacterial flora from fish fillets by polymerase chain reaction. J Microbiol Methods 67:456–462. doi: 16/j.mimet.2006.04.019

Lethem M, James S, Marriott C, Burke J (1990) The origin of DNA associated with mucus glycoproteins in Cystic Fibrosis sputum. Eur Respir J 3:19 –23.

Lim YW, Evangelista JS, Schmieder R, Bailey B, Haynes M, Furlan M, Maughan H, Edwards R, Rohwer F, Conrad D (2014) Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. J Clin Microbiol 52:425–437. doi: 10.1128/JCM.02204-13

Lim YW, Schmieder R, Haynes M, Furlan M, Matthews TD, Whiteson K, Poole SJ, Hayes CS, Low DA, Maughan H, Edwards R, Conrad D, Rohwer F (2013) Mechanistic model of Rothia mucilaginosa adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data. PLOS ONE 8:e64285. doi: 10.1371/journal.pone.0064285

Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. J Cyst Fibros Off J Eur Cyst Fibros Soc. doi: 10.1016/j.jcf.2012.07.009

Markle JGM, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS (2013) Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science 339:1084–1088. doi: 10.1126/science.1233521

Meyer F, Paarmann D, D’Souza M, Olson R, Glass E, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards R (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386. doi: 10.1186/1471-2105-9-386

Mokili JL, Dutilh BE, Lim YW, Schneider BS, Taylor T, Haynes MR, Metzgar D, Myers CA, Blair PJ, Nosrat B, Wolfe ND, Rohwer F (2013) Identification of a novel Human Papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. PLOS ONE 8:e58404. doi: 10.1371/journal.pone.0058404

Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77. doi: 10.1016/j.coviro.2011.12.004

Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE (2012) The lung microbiome in moderate and severe chronic obstructive pulmonary disease. PLoS ONE 7:e47305. doi: 10.1371/journal.pone.0047305

Proctor LM (2011) The human microbiome project in 2011 and beyond. Cell Host Microbe 10:287–291. doi: 10.1016/j.chom.2011.10.001

Pruesse E, Peplies J, Glöckner FO (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinforma Oxf Engl 28:1823–1829. doi: 10.1093/bioinformatics/bts252

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–596. doi: 10.1093/nar/gks1219

Robertson CE, Harris JK, Wagner BD, Granger D, Browne K, Tatem B, Feazel LM, Park K, Pace NR, Frank DN (2013) Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. Bioinforma Oxf Engl 29:3100–3101. doi: 10.1093/bioinformatics/btt526

Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Kehagia V, Connett GJ, Bruce KD (2006) Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with Cystic Fibrosis. J Clin Microbiol 44:2601–2604. doi: <p>10.1128/JCM.02282-05</p>

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6:e17288. doi: 10.1371/journal.pone.0017288

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814. doi: 10.1038/nmeth.2066

Shak S, Capon DJ, Hellmiss R, Marsters SA, Baker CL (1990) Recombinant human DNase I reduces the viscosity of Cystic Fibrosis sputum. Proc Natl Acad Sci 87:9188 –9192.

Stewart FJ, Ottesen EA, DeLong EF (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J 4:896–907.

Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, Doré J (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. Appl Environ Microbiol 65:4799–4807.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483. doi: 10.1038/nprot.2009.10

Willner D, Furlan M (2010) Deciphering the role of phage in the cystic fibrosis airway. Virulence 1:309–313. doi: 10.4161/viru.1.4.12071

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in Cystic Fibrosis and non-Cystic Fibrosis individuals. PloS One 4:e7370. doi: 10.1371/journal.pone.0007370

Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, Lim YW, Rainey PB, Schmieder R, Youle M, Conrad D, Rohwer F (2012) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. Am J Respir Cell Mol Biol 46:127–131. doi: 10.1165/rcmb.2011-0253OC

Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Meth 7:943–944. doi: 10.1038/nmeth1210-943

CHAPTER 3

Metagenomics and metatranscriptomics:
Windows on CF-associated viral and microbial communities

Abstract

Samples collected from CF patient airways often contain large amounts of host-derived nucleic acids that interfere with recovery and purification of microbial and viral nucleic acids. This study describes metagenomic and metatranscriptomic methods that address these issues. Microbial and viral metagenomes, and microbial metatranscriptomes, were successfully prepared from sputum samples from five adult CF patients. Contaminating host DNA was dramatically reduced in the metagenomes. Each CF patient presented a unique microbiome; in some *Pseudomonas aeruginosa* was replaced by other opportunistic bacteria. Even though the taxonomic composition of the microbiomes are very different, the metabolic potentials encoded by the community are very similar. The viral communities were dominated by phages that infect major CF pathogens. The metatranscriptomes reveal differential expression of encoded metabolic potential with changing health status. Microbial and viral metagenomics combined with microbial transcriptomics characterize the dynamic polymicrobial communities found in CF airways, revealing both the taxa present and their current metabolic activities. These approaches can facilitate the development of individualized treatment plans and novel therapeutic approaches.

Introduction

In the lungs of cystic fibrosis (CF) patients, the defective cystic fibrosis transmembrane regulator (CFTR) protein affects transepithelial ion transport, consequently hindering the normal airway clearance mechanisms(Boucher 2002; Riordan 2008).(Riordan 2008)(Riordan 2008) The resultant static mucoid environment is colonized by a dynamic and complex community of microbes, viruses, and fungi (reviewed in LiPuma 2010).

While standard microbial culture techniques had identified the key pathogens, more recent culture-independent approaches based on 16S rRNA gene sequencing revealed a much wider range of microbial species associated with CF lungs (Rogers et al. 2004; Harris et al. 2007; Bittar et al. 2008; Cox et al. 2010; Guss et al. 2011; Zhao et al. 2012). However, 16S rRNA-based methods are limited in taxonomic resolution and are subject to biases (summarized in Claesson et al. 2010); their predictions of metabolic activities are confined to those general functions known for the taxa, thus overlooking potentially important strain-specific variants. Metagenomics can overcome those limitations.

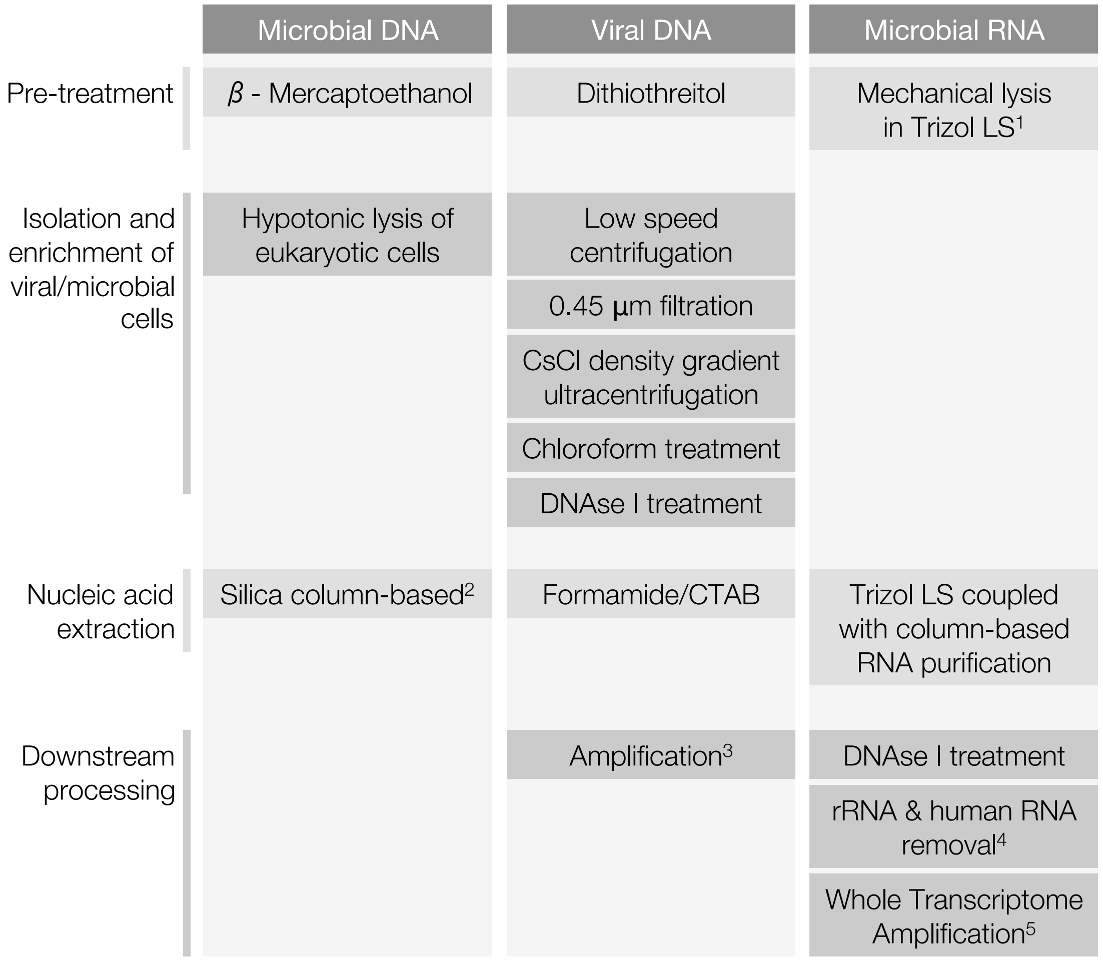
The metagenomic approach has been used to study viruses in human-associated environments such as blood (Breitbart and Rohwer 2005), feces (Breitbart et al. 2003; Zhang et al. 2006; Nakamura et al. 2009), and the lungs (Willner et al. 2009). It has also been successfully used to characterize the viral communities in sputum samples from CF and non-CF individuals (Willner et al. 2009). The presence of phages in CF airways is of particular relevance for clinical treatment, as environmental stress from the CF mucus and frequent antibiotic treatment is known to enhance phage mobility and promote the phage-mediated spread of antibiotic resistance genes in CF lungs (Rolain et al. 2009; Fothergill et al. 2011).

On the other hand, it is challenging to generate microbial metagenomes from CF samples. One reason for this is that microbial DNA isolated from CF sputum or lung tissue samples usually contains a large amount of human DNA, often greater than 99% of the total DNA recovered (Lethem et al. 1990; Shak et al. 1990; Breitenstein et al. 1995). Although some intact human cells may be present in the original sample, most of the contaminating DNA is extracellular and adsorbed to the surface of microbes, making isolation of pure microbial DNA particularly difficult.

Complementing metagenomics, metatranscriptomics characterizes the microbial genes expressed in an environment and can monitor shifts in their transcription or stability in response to perturbations, e.g., antibiotic treatments in CF patients. This approach has been used to investigate microbial community metabolism in marine (Poretsky et al. 2009; Hewson et al. 2010; McCarren et al. 2010) and soil (Leininger et al. 2006; Urich et al. 2008) environments, but its application to host-associated microbes has been limited to a few instances (Wittekindt et al. 2010; Gosalbes et al. 2011) due to technical challenges (Supplementary Table 3.1). One such challenge is that messenger RNAs (mRNAs) account for only ~5% of total cellular RNA. Various rRNA depletion methods have been developed to enrich samples for mRNA (Supplementary Table 3,2). Concurrent application of multiple methods (e.g., mRNA-ONLY^{TM}, MICROB*E*xpress^{TM} and MessageAmp^{TM}) can remove more of the rRNA in some instances, but efficacy remains limited, especially when working with partially degraded rRNA (He et al. 2010).

Metatranscriptomics of host-associated communities is particularly difficult. Amplification of the microbial RNA by methods that utilize synthetic polyadenylation is not applicable for samples that contain large amounts of eukaryotic mRNA. The appended poly-A tails reduce the amount of useful sequence data, especially when pyrosequencing technologies such as Roche/454 (Margulies et al. 2005) are used. Due to the short half-life and small quantity of mRNA, sample filtration and manipulation with buffer should be avoided prior to RNA extraction. This inevitably causes an increase in host RNA contamination when dealing with host-associated microbial samples. In the recent microbial metatranscriptomic study of mule deer lymph nodes by (Wittekindt et al. 2010), 99.3% of the taxonomically assigned reads were host-derived and <0.01 % were microbial.

Here we describe protocols to generate viral and microbial metagenomes, as well as microbial metatranscriptomes, from fresh CF sputum using 454 GS FLX Titanium pyrosequencing (Figure 3.1). These methods target and enrich for viral and microbial DNA, as well as microbial mRNA, while minimizing contamination with host nucleic acids. This is the first study to simultaneously survey the microbiome, virome, and community metatranscriptome in any ecosystem.



	Microbial DNA	Viral DNA	Microbial RNA
Pre-treatment	*β* - Mercaptoethanol	Dithiothreitol	Mechanical lysis in Trizol LS[1]
Isolation and enrichment of viral/microbial cells	Hypotonic lysis of eukaryotic cells	Low speed centrifugation 0.45 µm filtration CsCl density gradient ultracentrifugation Chloroform treatment DNAse I treatment	
Nucleic acid extraction	Silica column-based[2]	Formamide/CTAB	Trizol LS coupled with column-based RNA purification
Downstream processing		Amplification[3]	DNAse I treatment rRNA & human RNA removal[4] Whole Transcriptome Amplification[5]

Figure 3.1: Workflow for the preparation of CF sputum samples for microbiome, virome, and metatranscriptome sequencing.

Materials and Methods

Note: A detailed standard protocol describing each step can be downloaded from www.coralandphage.org or www.jove.com (See Chapter 1)

Sample collection: Eight sputum samples were collected from five CF volunteers (CF1 through CF5) at the Adult CF Clinic (San Diego, CA, United States) by expectoration into a sterile cup, with the exception of sample CF4-A that was a tracheal aspirate. All collection was in accordance with the University of California Institutional Review Board (HRPP 081500) and San Diego State University Institutional Review Board (SDSU IRB#2121). Clinical status at the time of collection was designated as *on treatment* (during systemic antibiotic treatment), *exacerbation* (prior to systemic antibiotic treatment), *post treatment* (upon completion of antibiotic treatment) or *stable* (when clinically stable and at their clinical and physiological baseline). Each sample was syringe-homogenized and divided into aliquots for metagenomic and metatranscriptomic analyses, culturing, and storage.

Virome protocol: (Supplementary Note 1) Dithiothreitol was added to the diluted sputum to aid mucus dissolution. Viral particles were purified by cesium chloride (CsCl) density gradient ultracentrifugation as described in Thurber *et al.* (2009)(Thurber et al. 2009). For one sample (CF4-A), the density gradient purification step was omitted for comparison. Viral DNA was extracted using CTAB/phenol:chloroform, and amplified with Phi29 DNA polymerase.

Microbiome protocol: (Supplementary Note 1) Sputum samples were treated with β-mercaptoethanol to disrupt mucus. Pelleted bacterial cells were repeatedly washed and centrifuged, then treated with DNase to remove human DNA.

Microbial metatranscriptome protocol: (Supplementary Note 1) Microbial cells in the sputum aliquots were mechanically lysed by vortexing with zirconia beads in TRIzol® LS (Life Technologies, NY). RNA was extracted using the Zymo Clean & Concentrator™ 25 kit (Zymo Research, Irvine, CA) with the small RNA removal protocol variation and treated with RNase-free DNase I (Ambion, Life Technologies: Grant Island, NY).

cDNA was generated using the WTA-2 kit (Sigma-Aldrich). The effect of nebulization on transcript length was assessed as described in Supplementary Note 2. Similarly, two rRNA depletion methods were tested: (i) the 'Ambion' method, i.e., MICROB*Enrich*™ and MICROB*Express*™, that removes bacterial rRNA as well as human rRNA and mRNA; and (ii) the Ribo-Zero™ method, i.e., Ribo-Zero™ rRNA Removal kit (Epidemiology version) (Epicentre, an Illumina company, Madison, WI) that removes bacterial and human rRNA.

Sequencing and data preprocessing/analysis: All samples were sequenced using the GS-FLX Titanium chemistry system. Primer tags in WTA amplified samples were removed using TagCleaner (Schmieder et al. 2010). All datasets were preprocessed to remove duplicates and reads of low quality using PRINSEQ (Schmieder and Edwards 2011a) (Supplementary Note 1). Metagenomic datasets were further screened and human-derived reads were removed using DeconSeq (Schmieder and Edwards 2011b).

The preprocessed metagenomes were annotated using BLASTn against the NCBI nucleotide database. Sequences assigned to the phylum Chordata and to vector or synthetic sequences were identified and removed. Virome sequences were then compared against an in-house boutique viral database containing 4,019 unique complete viral genomes using tBLASTx and normalized viral abundances were calculated. In the preprocessed metatranscriptomes, rRNA-like and non-rRNA reads were identified using BLASTn against the SILVA database (Pruesse et al. 2007). Non-rRNA reads were annotated using BLASTx against the NCBI non-redundant protein database. For details of database generation and content, normalization, as well as BLAST parameters, see Supplementary Note 1.

Taxonomic assignments: The best hit was assigned to the alignment with the highest coverage, identity, and score values. Query sequences with no BLAST hits above the defined threshold were designated as *unassigned*. The diversity of microbiomes was calculated based on the number of bacterial species identified in the datasets (Supplementary Note 1).

Metabolic pathways: Sequences from the metagenomes and metatranscriptomes (excluding all Chordata, vector, and synthetic sequences) were compared against the KEGG protein database using BLASTx. (The CF1-A metatranscriptome was omitted due to insufficient data.) For each pathway, the best hits and their abundances were identified and normalized using HUMAnN (Abubucker et al. 2012) . Normalized pathway abundance values were used to calculate similarities between samples based on random forests (Breiman 2001) and to partition the samples by Partitioning Around Medoids (PAM) clustering (Kaufman and Rousseeuw 2008).

Data accessibility: Sequence data was deposited in the NCBI Short Read Archive (SRA) with accession numbers SRP007749, SRP009392, and SRP009438.

Results and Discussion

This is the first study to describe a comprehensive workflow (Figure 3.1) for the generation of viromes, microbiomes, and microbial metatranscriptomes from any environment. The coupling of metagenomic and metatranscriptomic approaches provides an overview of both who is there and what they are doing, i.e., community taxonomy combined with the community’s encoded and expressed functional diversity. For this work, viral metagenomes (viromes), microbial metagenomes (microbiomes), and microbial metatranscriptomes were generated from twelve fresh sputum samples that had been collected from five adult CF patients (Table 3.1).

Table 3.1: Results summary for all viromes, microbiomes, and metatranscriptomes.

				Metagenomes		Metatranscriptomes
Patient ID	**Time Point**	**Date of Collection**	**Health Status**	**Microbial**	**Viral**	**Microbial**
				% (Microbial sequences)	% Viral sequences[a] VLPs Observed[b]	% (non-rRNA sequences)
CF1	A	09/02/2010	Stable	N/A[f]	N/A[f]	10.7% [d] (738)
	B	10/18/2010	Stable	N/A[f]	N/A[f]	30.3% [d] (41,789)
	C	11/12/2010	On Treatment	N/A[f]	N/A[f]	30.6% [d] (38,532)
	D	02/11/2011	Exacerbation	9% (14,691)	6.59% Yes	95.4%[e] (1,900)
	E	02/24/2011	On Treatment	58% (67,780)	31.99% Yes	N/A[f]
	F	03/14/2011	Post Treatment	79% (40,825)	6.07% No	31.6% [e] (7,971)
CF2	A	11/10/2010	On Treatment	N/A[f]	N/A[f]	87.6% [d] (68,976)
CF3	A	11/10/2010	On Treatment	N/A[f]	N/A[f]	89.6% [d] (93,375)
CF4	A	01/22/2011	Exacerbation	2% (3,834)	0.90%[c] No	86.8% [d] (59,394)
	B	02/01/2011	Post Treatment	0.2% (405)	6.77% No	99.1% [e] (32,446)
	C	03/20/2011	Stable	23% (41,636)	1.93% No	95.1% [e] (34,411)
CF5	A	10/07/2011	Exacerbation	2% (1,034)	3.00% Yes	N/A[f]
	B	10/21/2011	Post Treatment	1% (247)	3.78% Yes	N/A[f]

[a] Based on tBLASTx against viral genome database (threshold of 40% identity over at least 60% of the query sequence).

[b] Observation by epifluorescence microscopy of the viral fraction collected following cesium chloride density gradient centrifugation.

[c] Sample collected by filtration through 0.45 μm filter without cesium chloride density gradient ultracentrifugation.

[d] Following rRNA depletion by the Ambion kits.

[e] Following rRNA depletion by the Ribo-Zero$^{\text{TM}}$ (Epidemiology) kit.

[f] Sample not available.

Viruses in CF sputum: The metagenomic approach was successfully used previously to characterize viruses in CF lungs (Willner et al. 2009). In this study, viromes were generated from eight sputum samples obtained from three CF patients (Table 3.1, Supplementary Table 3.3). Two methods for purifying virus-like particles (VLPs) from sputum were compared. Seven samples were purified by filtration and cesium chloride density gradient ultracentrifugation, followed by chloroform and DNase I treatment. This procedure yielded viromes that contained little (0.02% – 3.7%) host-derived sequence (with one exception due possibly to its exceptionally high amount of mucins and free DNA, thus more viscous sputum; Supplementary Table 3.3). Omission of the density gradient ultracentrifugation step for the eighth sample (see Methods) resulted in a virome with 97% host-derived sequence. Cesium chloride density gradient centrifugation, previously shown to recover the majority of known phages (25) remains the method of choice for reducing host contamination when isolating viruses from complex samples such as CF sputum.

Analysis of the seven cesium chloride density-purified viromes using tBLASTx against the viral genome database identified more than 450 viral genotypes with each virome containing 319 – 456 genotypes (except CF4-C that contained only eight; Figure 3.2a). Unknowns accounted for 49% to >99% of the total reads in most of the viromes (Supplementary Table 3.3), which is typical for viral metagenomes (Willner et al. 2009). The exceptions were those samples highly contaminated by host sequences (CF4-A and CF4-C; Figure 3.2a). The high number of “unknown” sequences implies the presence of novel viruses that cannot be identified by database similarity, as had been found in previous studies (Breitbart and Rohwer 2005; Willner et al. 2009).

The majority of the viruses identified were phage, predominantly those that infect known CF pathogens. Their predicted bacterial hosts were tallied and the top 21 were used to construct predicted host range profiles (Figure 3.2a). The profiles were highly similar between patients, and even more so for multiple samples from the same patient. They were dominated by phages that infect major CF pathogens such as *Streptococcus*, *Burkholderia*, *Mycobacterium*, *Enterobacteria*, and *Pseudomonas* genera. *Streptococcus* phage (particularly Dp-1) were found in high abundance in the samples with the greatest abundance (>30%) of *Streptococcus* spp. (i.e., CF1-D and CF1-E; Figure 3.4a).

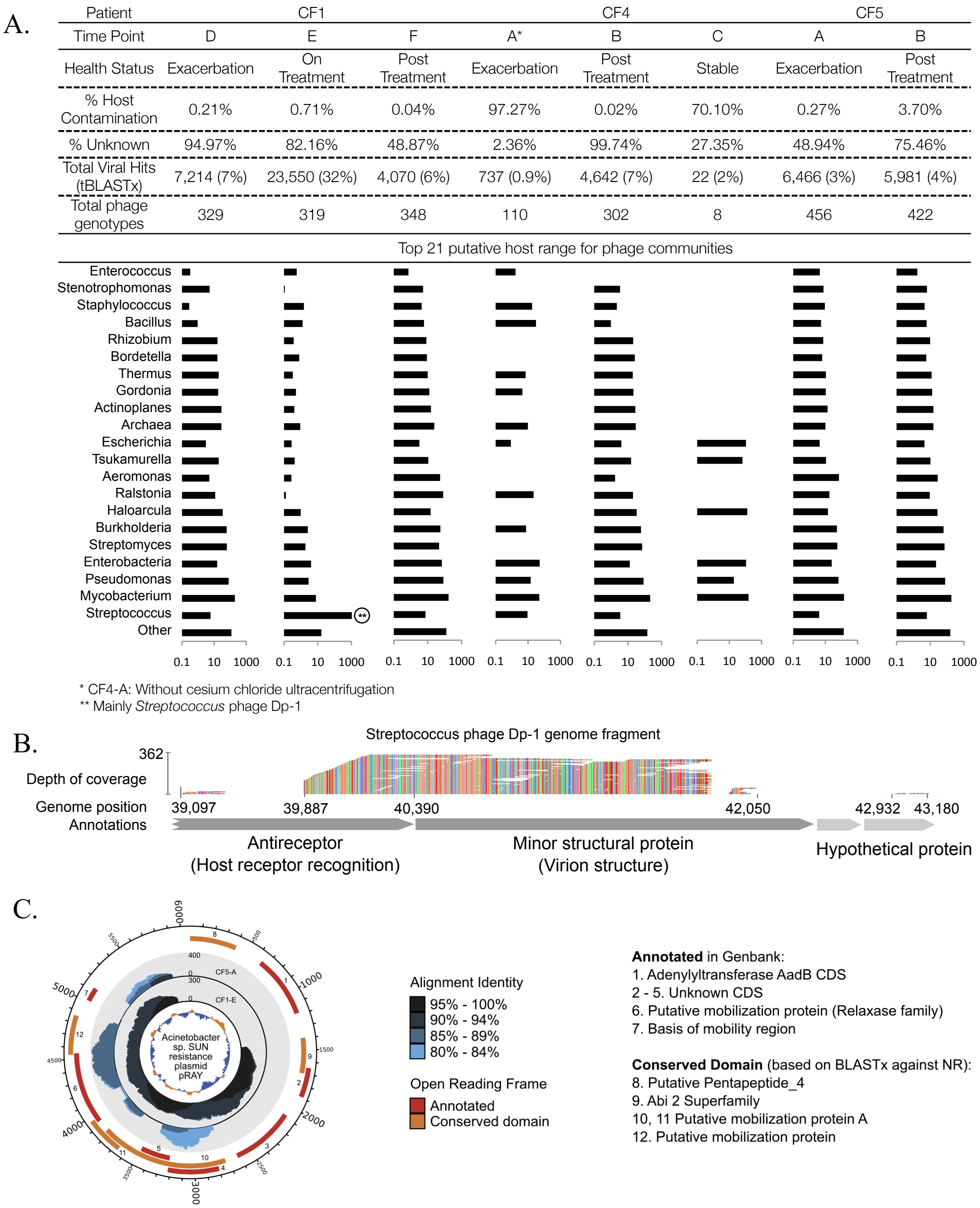
Streptococcus phage Dp-1 had been first isolated in 1975 from patients presenting with upper respiratory symptoms and was described as a virulent phage (McDonnell et al. 1975). Here tBLASTx analysis detected phage Dp-1 genes for DNA replication and packaging, host-receptor recognition, tail and capsid structural proteins, and host lysis (endolysin). The endolysin suggests possible top-down control of the *Streptococcus* spp. by lytic phage predation in these patients (Rodríguez-Cerrato et al. 2007).

When reads from the CF1-E virome were mapped against the Dp-1 reference genome (GI:327198314), high depth of coverage was observed for a Dp-1 genome fragment in a 3 kbp region that codes for an antireceptor and a minor structural protein (Figure 3.2b). Similarly, high coverage of regions of the *Acinetobacter* sp. SUN resistance plasmid pRAY was also observed in three samples (CF1-E, CF5-A, and CF5-B), including regions encoding a domain of the Abi-2 superfamily (proteins that confer resistance to phage infection) and mobilization proteins (Figure 3.2c). Finding these short sequences from these two genomes highly enriched in the viromes implies that they must be present in many other genomes, as well, likely the result of active horizontal gene transfer (HGT) in CF lungs. HGT is an important mechanism by which microbes evolve and adapt to the CF lung environment (Qiu et al. 2009), and phage can potentially facilitate this process.

The archaeal virus BJ1 (GI: 119756985) was identified in every virome—the first finding of an archaeal virus in the lungs. The hypersaline surface liquid in CF airways may be ecologically similar to the hypersaline lake in Inner Mongolia where the virus was isolated (Pagaling et al. 2007). Archaea were identified in low abundance in one microbiome (CF4-C; <0.1%) and all metatranscriptomes (<1%; data not shown), suggesting that they could play a role in the CF lung ecosystem. However, since more than 71% of the predicted ORFs for this archaeal virus show no similarity to any known genes, its genome sequence provides no clues as to what that role might be.

Eukaryotic DNA viruses in CF individuals have been shown to be dominated by a few viral genomes (Willner et al. 2009) that could potentially cause persistent infections, exacerbations, tumorigenesis, and poor clinical outcomes (Winnie and Cowan 1992; van Ewijk et al. 2008; Klein et al. 2009). The eukaryotic viruses identified in a previous study of the lungs of CF patients included torque teno virus (TTV), retroviruses, and human herpesviruses (Willner et al. 2009). In the current study, eukaryotic viruses, including human herpesviruses and retroviruses, were found in all samples (Supplementary Table 3.4), with torque teno viruses in high abundance in one sample (CF1-D).

Because RNA viruses are involved in the majority of respiratory infections, a filtration-based method was used to isolate RNA viruses from CF sputum (data not shown). However, this method was unable to recover identifiable RNA viruses, likely due their low abundances and technical challenges in their isolation.



Patient	CF1			CF4			CF5	
Time Point	D	E	F	A*	B	C	A	B
Health Status	Exacerbation	On Treatment	Post Treatment	Exacerbation	Post Treatment	Stable	Exacerbation	Post Treatment
% Host Contamination	0.21%	0.71%	0.04%	97.27%	0.02%	70.10%	0.27%	3.70%
% Unknown	94.97%	82.16%	48.87%	2.36%	99.74%	27.35%	48.94%	75.46%
Total Viral Hits (tBLASTx)	7,214 (7%)	23,550 (32%)	4,070 (6%)	737 (0.9%)	4,642 (7%)	22 (2%)	6,466 (3%)	5,981 (4%)
Total phage genotypes	329	319	348	110	302	8	456	422

* CF4-A: Without cesium chloride ultracentrifugation
** Mainly *Streptococcus* phage Dp-1

Figure 3.2: Taxonomic analysis of CF viromes. A. Putative host range profiles for phage communities. Each bar represents the sum of the normalized abundance values for all phage genotypes with the same putative bacterial host. Only the top 21 hosts are shown. B. Nucleotide-level alignment of CF1-E virome sequences against a region of the *Streptococcus* phage Dp-1 genome. Depth of coverage was based on 90% nucleotide identity. C. Coverage plot of *Acinetobacter* sp. SUN resistance plasmid pRAY recovered from CF1-E and CF5-A.

Microbes in CF sputum: When characterizing a microbial community, metagenomics surpasses a 16S rRNA-based approach in that it (1) frequently permits high-confidence species-level taxonomic assignment; (2) allows prediction of specialized functional capabilities of the adapted community, rather than inferring function from taxonomy; and (3) avoids the bias inherent in the selection of any universal target for PCR amplification. However, preparations of 'microbial' DNA derived from CF sputum or lung tissue are typically dominated by human DNA that was extracellular or adsorbed to the microbes. Several standard methods, including separation of human and microbial cells by percoll gradients (Childs and Gibbons 1988), treatment with DNase I, selective degradation of human DNA by ethidium bromide monoazide (Lee and Levin 2006), and use of the MolYsis kit (Molzym Life Science), have failed to reduce human DNA in CF samples (personal communications). In this study, the most effective procedure was found to be a modification of the method described by Breitenstein *et al.* (1995)(Breitenstein et al. 1995) that employs a combination of β-mercaptoethanol to reduce biofilm disulfide bonds, hypotonic lysis of eukaryotic cells, and DNase I treatment of soluble DNA (Supplementary Figure 3.1; Supplementary Table 3.5). Sufficient microbial DNA was extracted by this procedure to make amplification prior to sequencing unnecessary, thus avoiding potential amplification bias.

The amount of human contamination (13% – 97% of total preprocessed reads) was highly dependent on the sample properties. Samples collected from patients during exacerbations might be expected to contain higher levels of host DNA due to greater inflammation and neutrophil activity than those collected during and immediately following treatment. However, our metagenomic data showed no significant correlation between the fraction of host DNA and a patient's health status even though the amount of host DNA varies markedly among the metagenomes.

With high-throughput pyrosequencing, even a relatively small proportion of non-host sequence data can be sufficient to provide significant information. After the removal of eukaryotic reads, the microbiomes contained >75% bacterial reads (Table 3.1) and 2% – 12% unknowns, with the remainder being artificial and cloning vectors or synthetic constructs (Supplementary Table 3.5). The number of bacterial species identified, including aerobes and anaerobes, ranged from 24 to 256 (Supplementary Table 3.6).

Each patient presented a unique microbial profile (Figure 3.3a). The predominant groups persisted across the time points assayed but the relative abundance varied with exacerbations and antibiotic treatments. This suggests complex community dynamics in which the predominant groups adapt and persist, while others come and go in response to antibiotic treatment or other perturbations.

CF4 presented a classic CF lung microbiome where *P. aeruginosa* was one of the main players at all time points. In contrast, CF1 was colonized mainly by *Rothia mucilaginosa* and *Streptococcus* spp. during exacerbation. Effective treatments decreased *R. mucilaginosa*, thereby increasing the proportion of *P. aeruginosa.* The *Rothia dentocariosa* that colonized CF5 during exacerbations was eliminated by treatments, and the patient was subsequently colonized by *Pseudomonas fluorescence* instead of the common CF pathogen, *P. aeruginosa.* The microbiome profiles also showed that *P. aeruginosa* can be replaced as the main player by other opportunistic bacteria from the environment, as evidenced here by the colonization of (i) CF5, a landscape architect, by soil-dwelling *P. fluorescens* instead of the more common CF pathogen, *P. aeruginosa*, and (ii) CF1 by the oral flora *R. mucilaginosa* (Figure 3.3a). By going beyond the traditional tracking of particular recognized CF pathogens, metagenomics offers the possibility of personalized clinical treatment plans.

In some situations, metagenomics can yield in-depth genomic analysis (Narasingarao et al. 2012; Iverson et al. 2012). In this study, the CF1-E microbiome provided 7.8X average coverage over 93.56% of the genome of the most abundant species, *R. mucilaginosa.* Mapping of short reads against the reference genome DY-18 (GI: 283133067) (Figure 3.3b) revealed only 41 gaps that were >1,000 bp. Of those gaps, almost 20% were located in non-coding regions, 20% in regions annotated as hypothetical proteins, and the rest in genes of known function (Supplementary Table 3.7). In-depth analysis and interpretation of the genomic changes will be presented elsewhere (Chapter 5).

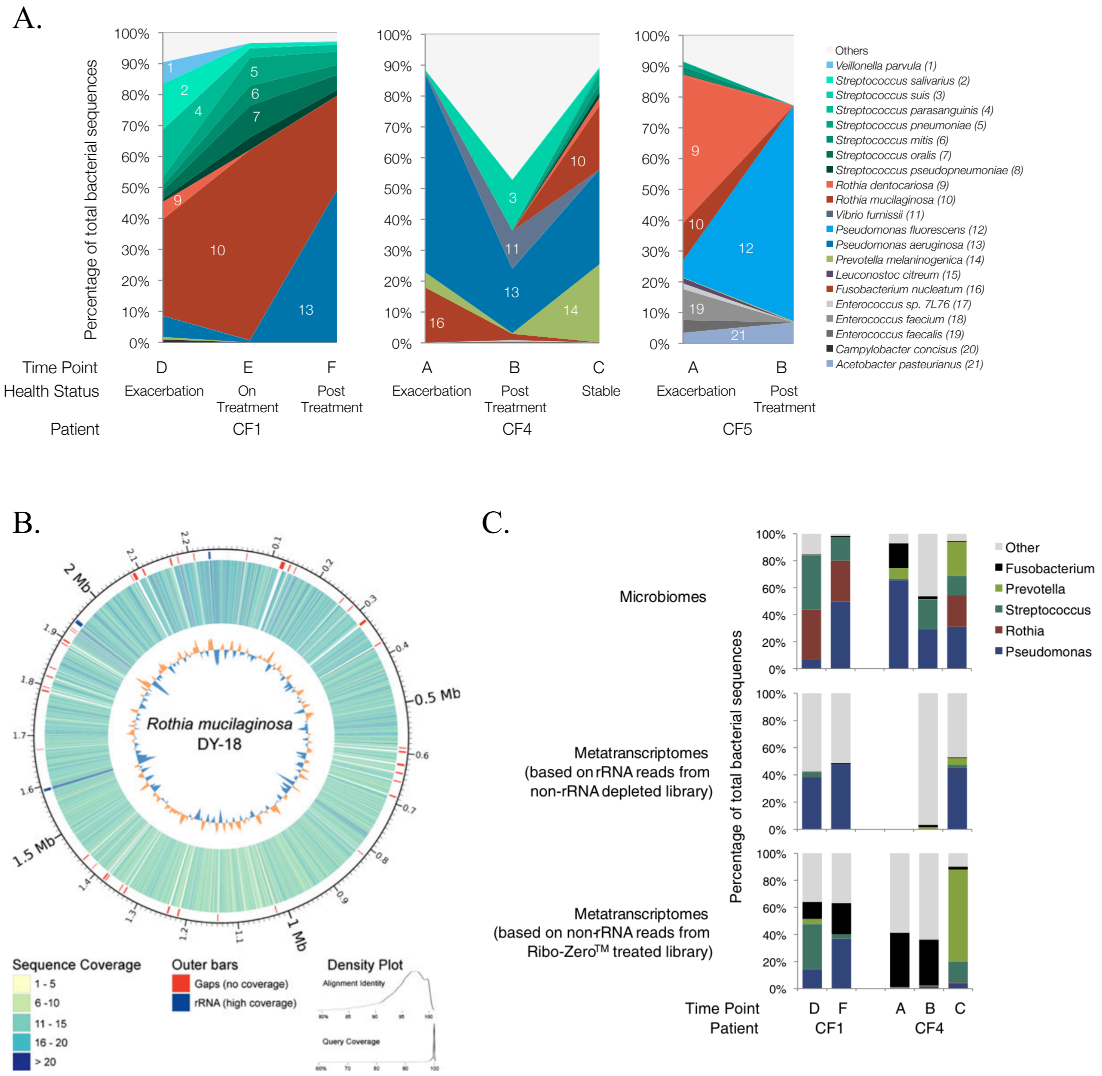


Figure 3.3: Taxonomic analysis of the microbial communities in three CF patients across multiple time points. A. Species-level comparisons between microbiomes. Identification was based on unique best hits using BLASTn against the NCBI nucleotide database. All species shown from the same genus are assigned similar colors. B. Sequence coverage of the *Rothia mucilaginosa* DY-18 genome by reads from the CF1-E microbiome. C. Genus-level comparisons between microbiomes and metatranscriptomes. (The CF4-A taxonomy is not shown here because an rRNA removal kit was used during metatranscriptome preparation.)

Evaluation of microbial metatranscriptome preparation: A high quality metatranscriptome contains relatively few rRNA reads and an unbiased sampling of RNAs of various lengths. Nebulization, the first step during preparation of a sequencing library, is a potential source of transcript size-induced bias since the size of our cDNA ranged from 50 to 4,000 bp (Supplementary Figure 3.2). While nebulization of high molecular weight DNA creates random fragments, application to lower molecular weight cDNA may result in non-uniform coverage or the loss of short transcripts (Torres et al. 2008).

Here, the effect of nebulization on transcript length was tested on four samples (Supplementary Note 1). There was no difference in the relative translated protein length profiles with and without nebulization (Supplementary Figure 3.3). The median translated polypeptide length (345 – 412 amino acids; Supplementary Table 3.8) is in the high range of previously described microbial protein lengths (Brocchieri and Karlin 2005), possibly due to the use of the small RNA removal protocol in the RNA Clean & Concentrator^{TM} kit (Zymo).

Nebulization also did not affect the relative proportion of rRNA-like and non-rRNA reads, although the proportion of rRNA varied from patient to patient (9.4% to 70.8%; Figure 3.4a). A reduced rRNA fraction was often associated with an increased proportion of eukaryote and unidentified reads.

Effective mRNA enrichment by rRNA removal using subtractive hybridization (e.g., the MICROB*Enrich*^{TM} from Ambion) had been previously demonstrated on synthetic microbial communities (He et al. 2010). However, efficacy depended on the integrity of the RNA and community composition. Here, four samples (CF1-D, CF1-F, CF4-B, and CF4-C) were used to compare two hybridization-based commercial rRNA removal kits. Each sample was divided into three aliquots for alternative treatments: (A) MICROB*Enrich*^{TM} + MICROB*E*xpress^{TM} (**A**mbion); (E) Early access version of Ribo-Zero^{TM} rRNA removal kit - Epidemiology (**E**picentre); and (N) **N**o treatment. In total, these twelve metatranscriptomes yielded 425,523 reads (105 Mbp), 48 – 77% of which were retained after data preprocessing (average read length = 252 bp).

With either treatment, the proportion of rRNA was reduced significantly for CF1-D but minimally for CF1-F (Figure 3.4b, Supplementary Table 3.9). Of the two treatment methods, the Ribo-Zero^{TM} is more effective in eukaryotic rRNA removal as evidenced by CF4-B and CF4-C. In these samples that contained a larger proportion of eukaryotic reads, the Ribo-Zero^{TM} treatment removed 96% and 90% of the rRNA, while the Ambion treatments increased the relative rRNA content.

Notably these rRNA removal methods were markedly less effective for sample CF1-F. Even with Ribo-Zero^{TM}, the most effective for all other samples, the treatment yielded only a 5% reduction in total rRNA. This inter-sample variation in rRNA removal could reflect differences in the microbial community present, the quality of extracted RNA, the accessibility of rRNAs for probe hybridization, and/or the degree of homology between the designed rRNA probes and the unknown community members.

Use of either rRNA depletion treatment precludes subsequent rRNA-based analysis of the sample because both methods distort the apparent relative abundances of microbial taxa (Supplementary Figure 3.4). Bias was apparent for all samples except CF1-F; neither rRNA depletion method had significant effect on that sample.

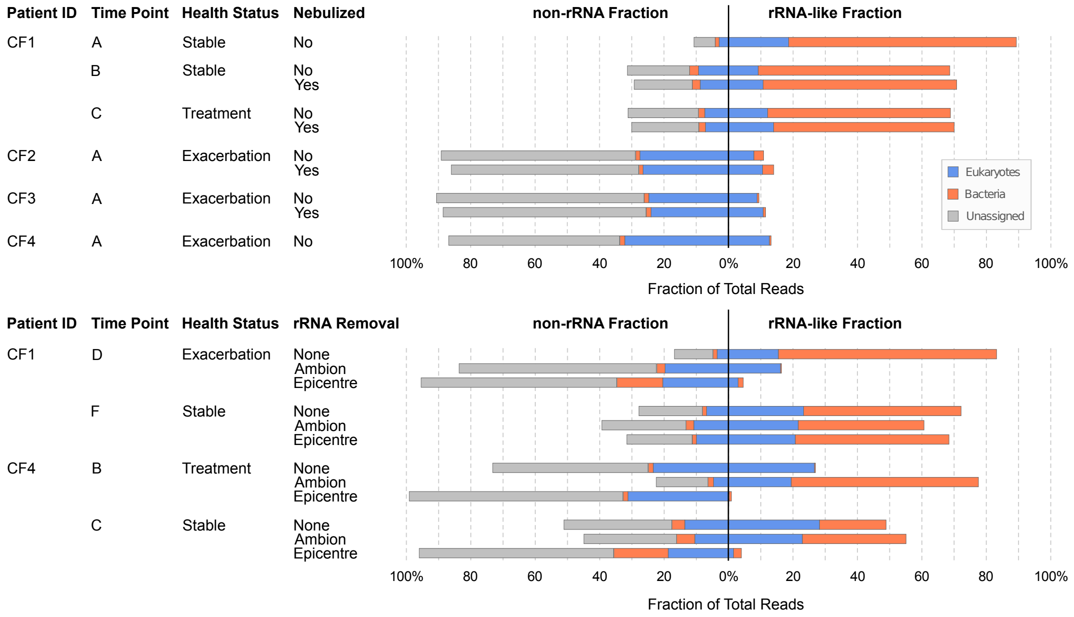


Figure 3.4: Evaluation of the effects of nebulization and rRNA depletion on the relative amounts of rRNA and non-rRNA in metatranscriptomes. (a) Effects of nebulization. All samples were treated by the Ambion rRNA depletion kits. (b) Comparison of rRNA depletion methods. "Ambion" method uses a combination of MICROB*Enrich*^{TM} + MICROB*Express*^{TM}; "Epicentre" method uses Ribo-Zero^{TM} rRNA Removal kit (Epidemiology version).

Microbial taxonomy three ways: Three methods were used to determine the relative abundances of the microbial genera present within a patient sample: (1) annotation of microbiomes; (2) 16S rRNA-based annotation of metatranscriptomes; and (3) annotation of metatranscriptomes based on encoded protein sequences. The marked differences observed among the three (Figure 3.3c) indicate that some community members are more transcriptionally active, thus contribute more to community metabolism than their relative numbers would predict. This is further evidenced by functional characterization (see below).

Community Metabolic Profiles: Whereas metagenomics surveys the functional capabilities encoded by members of the microbial commmunity, adding metatranscriptomic data provides insights into the current metabolic activities, insights that can assist in tailoring an effective treatment. To compare these approaches, the viromes, microbiomes, and metatranscriptomes were functionally annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. A total of 216 metabolic pathways and 212 modules (collection of functional units) were identified. Multidimensional scaling (MDS) and clustering of all datasets based on the normalized abundance value (see Methods) yielded three distinct groups, thus showing that a different view of community metabolism can be obtained from each method (Figure 3.5). The results demonstrate the internal consistency of each method. The few exceptions were (i) the clustering of the CF4-C metatranscriptome with the microbiomes; (ii) the clustering of the CF4-B and CF5-B microbiomes with the metatranscriptomes attributable to the low number of reads; (iii) the CF4-A virome, purified without the gradient utracentrifugation, appearing as an outlier in the virome cluster.

Overall, the metabolic profiles derived from the microbiomes were the most similar between patients as well as between time points for each patient (Supplementary Figure 3.5), indicating a shared pool of metabolic genes required for survival in the CF environment. The greatest variation, likely reflecting specialized adaptations within the viral and microbial communities, is seen in the principal component analysis (PCA) plot (Supplementary Figure 3.6).

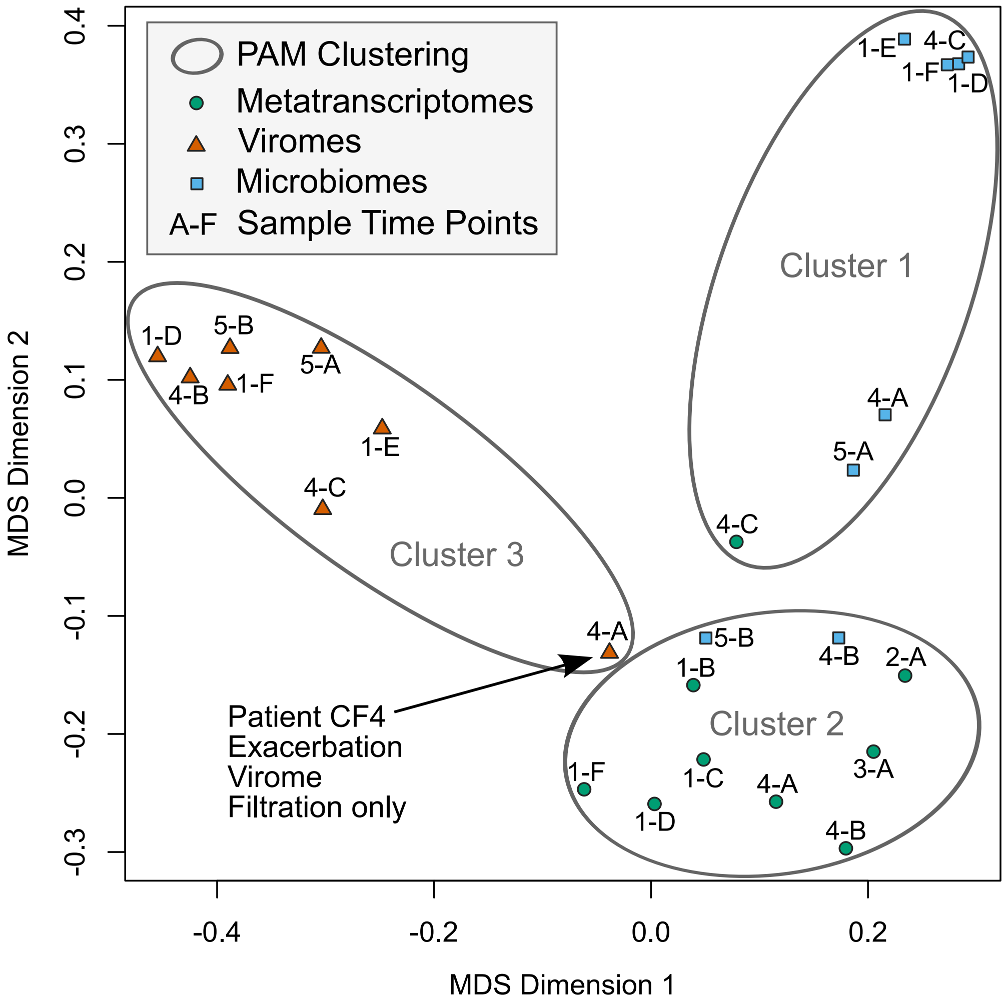


Figure 3.5: Comparison of KEGG metabolic pathways identified in viromes, microbiomes, and metatranscriptomes as shown by multidimensional scaling (MDS). Grouping by Partitioning Around Medoids (PAM) clustering placed all samples in the appropriate cluster with the exception of the CF4-C metatranscriptome. CF1-A was omitted from both analyses due to insufficient data.

Clinical implications: The picture of dynamic and diverse polymicrobial communities presented here deviates from classic CF clinical profiles derived from culturing, thereby challenging one-size-fits-all treatment regimes. For example, current treatments targeting the classic CF pathogen, *P. aeruginosa*, would not be effective against *P. fluorescence*, *R. mucilaginosa*, or *R. dentocariosa*—all of which were abundant in these microbiomes. The ability to identify the resident viruses and microbes that could potentially trigger exacerbation events makes effective individual treatment plans a possibility, including intervention based on predicted disease progression. Ongoing surveillance can monitor inter-patient transmission and inform infection control measures. In addition, shifting the focus from pathogen taxonomy to the community metabolisms associated with periods of stability and exacerbation opens the door to novel therapeutic approaches that change the airway environment to favor less pathogenic communities.

Summary

The combination of metagenomic and metatranscriptomic approaches demonstrated here can provide insight into the complex and dynamic interations between the host and both the microbial and viral communities present in CF lungs.

- The methods described successfully recover viral DNA, microbial DNA, and microbial mRNA from CF sputum, while minimizing contamination with host nucleic acids.
- Of the viruses identified in the virome reads, most are phage that infect major CF pathogens. These likely include vectors for clinically-significant microbe-microbe horizontal gene transfer. However, the majority of virome reads are "unknown," thus potentially novel viruses.
- To identify the microbes present, the microbiomes were annotated using BLASTn against the NCBI nucleotide database. Each CF patient possessed a unique microbial profile that shifted over time and sometimes reflected the acquisition of persistent opportunistic bacteria from the environment. High genome coverage for the most abundant species allowed in-depth genomic analysis.
- The third concurrent approach, microbial metatranscriptomics, monitors the active community metabolism, as opposed to the metabolic potential encoded in the genomes. Of the three measures, the metatranscriptomes showed the greatest variation between patients and over time, thus is best able to capture the dynamic nature of these complex communities.

Acknowledgments

This work was supported by the National Institute of Health and Cystic Foundation Research Inc. through grants (1 R01 GM095384-01 and CFRI #09-002) awarded to Forest Rohwer. We thank Epicentre, an Illumina company for providing early access to Ribo-Zero^{TM} Epidemiology kits. We thank Peter Salamon, Ben Felts, Katie Barott, Jeremy Barr, and Katrine Whiteson for critical readings and discussions of the manuscript.

Chapter 3 in full is published in the Journal of Cystic Fibrosis. Yan Wei Lim, Robert Schmieder, Matthew Haynes, Dana Willner, Mike Furlan, Merry Youle, Katelynn Abbott, Robert Edwards, Jose Evangelista, Douglas Conrad, and Forest Rohwer; 2013. The dissertation author was the primary investigator and author of this paper. Article Link: http://www.sciencedirect.com/science/article/pii/S1569199312001403

References

Abubucker S, Segata N, Goll J, Schubert A, Rodriguez-Mueller B, Zucker J, team tHMPMR, Schloss P, Gevers D, Mitreva M, Huttenhower C (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLos Computational Biology. doi: 10:10.1371/journal.pcbi.1002358

Bittar F, Richet H, Dubus J-C, Reynaud-Gaubert M, Stremler N, Sarles J, Raoult D, Rolain J-M (2008) Molecular detection of multiple emerging pathogens in sputa from Cystic Fibrosis patients. PLoS ONE 3:e2908. doi: 10.1371/journal.pone.0002908

Boucher RC (2002) An overview of the pathogenesis of Cystic Fibrosis lung disease. Advanced Drug Delivery Reviews 54:1359–1371. doi: 10.1016/S0169-409X(02)00144-8

Breiman L (2001) Random forests. machine learning. Springer Netherlands 45:5–32. doi: 10.1023/A:1010933404324

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

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

Breitenstein S, Tümmler B, Römling U (1995) Pulsed field gel electrophoresis of bacterial DNA isolated directly from patients' sputa. NAR 23:722–723.

Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. NAR 33:3390–3400. doi: 10.1093/nar/gki615

Childs WC, Gibbons RJ (1988) Use of percoll density gradients for studying the attachment of bacteria to oral epithelial cells. Journal of Dental Research 67:826 – 830. doi: 10.1177/00220345880670050601

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. NAR 38:e200. doi: 10.1093/nar/gkq873

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044. doi: 10.1371/journal.pone.0011044

Fothergill JL, Mowat E, Walshaw MJ, Ledson MJ, James CE, Winstanley C (2011) Effect of antibiotic treatment on bacteriophage production by a Cystic Fibrosis epidemic strain of Pseudomonas aeruginosa. Antimicrobial Agents and Chemotherapy 55:426 –428. doi: 10.1128/AAC.01257-10

Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, Latorre A, Moya A (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS ONE 6:e17447. doi: 10.1371/journal.pone.0017447

Guss AM, Roeselers G, Newton ILG, Young CR, Klepac-Ceraj V, Lory S, Cavanaugh CM (2011) Phylogenetic and metabolic diversity of bacteria associated with Cystic Fibrosis. ISME J 5:20–29. doi: 10.1038/ismej.2010.88

Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurso FJ, Pace NR (2007) Molecular identification of bacteria in bronchoalveolar lavage fluid from children with Cystic Fibrosis. PNAS 104:20529–20533. doi: 10.1073/pnas.0709804104

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R, Hugenholtz P (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Meth 7:807–812. doi: 10.1038/nmeth.1507

Hewson I, Poretsky RS, Tripp HJ, Montoya JP, Zehr JP (2010) Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. Environmental Microbiology 12:1940–1956. doi: 10.1111/j.1462-2920.2010.02198.x

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590. doi: 10.1126/science.1212665

Kaufman L, Rousseeuw P (2008) Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, Inc, Hoboken, NJ, USA

Klein F, Amin Kotb WFM, Petersen I (2009) Incidence of human papilloma virus in lung cancer. Lung Cancer 65:13–18. doi: 10.1016/j.lungcan.2008.10.003

Lee J-L, Levin RE (2006) Use of ethidium bromide monoazide for quantification of viable and dead mixed bacterial flora from fish fillets by polymerase chain reaction. Journal of Microbiological Methods 67:456–462. doi: 16/j.mimet.2006.04.019

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442:806–809. doi: 10.1038/nature04983

Lethem M, James S, Marriott C, Burke J (1990) The origin of DNA associated with mucus glycoproteins in Cystic Fibrosis sputum. European Respiratory Journal 3:19 –23.

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clinical Microbiology Reviews 23:299 –323. doi: 10.1128/CMR.00068-09

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. doi: 10.1038/nature03959

McCarren J, Becker JW, Repeta DJ, Shi Y, Young CR, Malmstrom RR, Chisholm SW, DeLong EF (2010) Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. PNAS 107:16420 –16427. doi: 10.1073/pnas.1010732107

McDonnell M, Ronda-Lain C, Tomasz A (1975) “Diplophage”: A bacteriophage of Diplococcus pneumoniae. Virology 63:577–582. doi: 10.1016/0042-6822(75)90329-3

Nakamura S, Yang C-S, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, Maeda N, Kawai J, Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4:e4219. doi: 10.1371/journal.pone.0004219

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6:81–93. doi: 10.1038/ismej.2011.78

Pagaling E, Haigh RD, Grant WD, Cowan DA, Jones BE, Ma Y, Ventosa A, Heaphy S (2007) Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC Genomics 8:410. doi: 10.1186/1471-2164-8-410

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environmental Microbiology 11:1358–1375. doi: 10.1111/j.1462-2920.2008.01863.x

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. NAR 35:7188 –7196. doi: 10.1093/nar/gkm864

Qiu X, Kulasekara BR, Lory S (2009) Role of horizontal gene transfer in the evolution of Pseudomonas aeruginosa virulence. Genome Dyn 6:126–139. doi: 10.1159/000235767

Riordan JR (2008) CFTR function and prospects for therapy. Annu Rev Biochem 77:701–726. doi: 10.1146/annurev.biochem.75.103004.142532

Rodríguez-Cerrato V, García P, del Prado G, García E, Gracia M, Huelves L, Ponte C, López R, Soriano F (2007) In vitro interactions of LytA, the major pneumococcal autolysin, with two bacteriophage lytic enzymes (Cpl-1 and Pal), cefotaxime and moxifloxacin against antibiotic-susceptible and -resistant Streptococcus pneumoniae strains. Journal of Antimicrobial Chemotherapy 60:1159–1162. doi: 10.1093/jac/dkm342

Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Bruce KD (2004) Characterization of bacterial community diversity in Cystic Fibrosis lung infections by use of 16S ribosomal DNA terminal restriction fragment length polymorphism profiling. J Clin Microbiol 42:5176–5183. doi: <p>10.1128/JCM.42.11.5176-5183.2004</p>

Rolain J-M, Francois P, Hernandez D, Bittar F, Richet H, Fournous G, Mattenberger Y, Bosdure E, Stremler N, Dubus J-C, Sarles J, Reynaud-Gaubert M, Boniface S, Schrenzel J, Raoult D (2009) Genomic analysis of an emerging multiresistant Staphylococcus aureus strain rapidly spreading in Cystic Fibrosis patients revealed the presence of an antibiotic inducible bacteriophage. Biology Direct 4:1. doi: 10.1186/1745-6150-4-1

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6:e17288. doi: 10.1371/journal.pone.0017288

Schmieder R, Lim Y, Rohwer F, Edwards R (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics 11:341. doi: 10.1186/1471-2105-11-341

Shak S, Capon DJ, Hellmiss R, Marsters SA, Baker CL (1990) Recombinant human DNase I reduces the viscosity of Cystic Fibrosis sputum. PNAS 87:9188 –9192.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protocols 4:470–483. doi: 10.1038/nprot.2009.10

Torres TT, Metta M, Ottenwälder B, Schlötterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Research 18:172 –177. doi: 10.1101/gr.6984908

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS ONE 3:e2527. doi: 10.1371/journal.pone.0002527

Van Ewijk BE, van der Zalm MM, Wolfs TFW, Fleer A, Kimpen JLL, Wilbrink B, van der Ent CK (2008) Prevalence and impact of respiratory viral infections in young children with cystic fibrosis: prospective cohort study. Pediatrics 122:1171–1176. doi: 10.1542/peds.2007-3139

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in Cystic Fibrosis and non-Cystic Fibrosis individuals. PLoS ONE 4:e7370. doi: 10.1371/journal.pone.0007370

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

Wittekindt NE, Padhi A, Schuster SC, Qi J, Zhao F, Tomsho LP, Kasson LR, Packard M, Cross P, Poss M (2010) Nodeomics: pathogen detection in vertebrate lymph nodes using meta-transcriptomics. PLoS ONE 5:e13432. doi: 10.1371/journal.pone.0013432

Zhang T, Breitbart M, Lee WH, Run J-Q, Wei CL, Soh SWL, Hibberd ML, Liu ET, Rohwer F, Ruan Y (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4:e3. doi: 10.1371/journal.pbio.0040003

Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, Young VB, LiPuma JJ (2012) Decade-long bacterial community dynamics in Cystic Fibrosis airways. PNAS 109:5809–5814. doi: 10.1073/pnas.1120577109

Appendix for Chapter 3

Supplementary Notes

Supplementary Note 1: Methods and bioinformatics for data pre-processing and analysis

Sample description, collection and processing: Each sample was given a unique patient ID (CF<number>) followed by the time point when the sample was collected (represented by a letter, <A-Z>). Sample properties, e.g., "viscosity" and volume, varied with the differing health status of the patients. The initial sputum volume ranged from 2 to 6 ml, varying between patients and across time points. Sputum samples were collected in the clinic during the patient's visit. Patient clinical status at that time was designated by the clinician based on the commonly used Fuch's criteria, lung function tests, and the patient's reported outcome. The volume of the sputum was measured and an appropriate volume of 1X Phosphate Buffered Solution (PBS) was added to make the volume up to at least 6 ml. The sample was then syringe-homogenized and split into 5 aliquots for metagenomic (2) and metatranscriptomic (1) studies, culturing (1), and storage (1). Immediately after homogenization, each homogenate for microbial metatranscriptomic study was transferred to a 15 ml falcon tube containing 1 volume of 0.1 mm zirconia beads and 3 volumes of Trizol LS and was immediately vortexed for 10 minutes at medium speed to mechanically lyse microbial cells while maintaining the RNA intact. All samples were then transported on ice to the lab. Sputum samples were processed within two hours of collection for CF1-CF3, and within 30 minutes for the other samples.

Samples and corresponding metagenomes and metatranscriptomes: The generation of microbial metagenomes and community metatranscriptomes requires fresh

CF sputum. Initially, samples CF1-A, CF1-B, CF1-C, CF2, and CF3 were collected to test only the metatranscriptomics approach, whereas CF5 samples were collected before the metatranscriptomic approach was developed. Therefore, there were no appropriate remaining sample aliquots available for constructing the corresponding metagenomes or metatranscriptomes for these samples (Table 3.1).

Virome protocol: The sputum homogenate aliquot was diluted with suspension medium (SM) buffer (1M NaCl, 10 mM $MgSO_4$, 50 mM Tris-HCL pH 7.4) to at least 6 ml. An equal volume of 6.5 mM dithiothreitol was added and incubated for 1 h at 37 °C to aid mucus dissolution. The sample was centrifuged at 3,800×g for 15 min at 10 °C to pellet cells and debris. The supernatant was collected and syringe filtered through a 0.45 micron filter (Milipore, Billerica, MA). Viral particles were purified using cesium chloride (CsCl) gradient ultracentrifugation as described in (Thurber et al. 2009). The sample was added onto a cesium chloride density step gradient (1.7 g/ml, 1.5 g/ml, and 1.35 g/ml), and spun at 22,000 x g at 4 °C for 2 hours. The 1.5 g/ml density layer was then collected and examined by epifluorescence microscopy to verify the presence of virus-like particles (VLPs) and the absence of microbial and eukaryotic cells. For one of the eight samples (CF4-A), the cesium chloride density gradient purification step was omitted for comparison.

10X DNase buffer (500 mM NaAc, 100 mM $MgCl_2$, 20 mM $CaCl_2$, pH6.5) and 2.5 unit DNase I (Sigma-Aldrich: St. Louis, MO) was added per µl of sample and the mixture was incubated for 2 h at 37 °C. The reaction was terminated by incubation at 65

°C for 15 minutes. Viral DNA was extracted using CTAB/phenol:chloroform, and amplified using multiple displacement amplification with Phi29 DNA polymerase.

Microbiome protocol: Sputum sample aliquots were diluted 1:5 with PBS, and β-mercaptoethanol was added to a final concentration of 2% (v/v). This mixture was incubated on a rocking platform at room temperature for 2 h, then centrifuged at 3800×g for 15 min at 10 °C. The pellets were resuspended in 10 ml deionized water, held at room temperature for 15 min, and centrifuged as above. This step was repeated, and then the pellets were resuspended in 10 ml DNase buffer (50 mM Na acetate, 10 mM $MgCl_2$, 2 mM $CaCl_2$, pH 6.5). DNase I was added to a final concentration of 15,000 U/ml, and the suspensions were incubated at 37 °C for 2 hrs. After centrifugation, the pellets were resuspended in 10 ml SE buffer (75 mM NaCl, 25 mM EDTA, pH 7.5) and centrifuged as above. This step was repeated, and the final pellet was resuspended in ~500 µl SE. DNA was extracted from this cell suspension with the Nucleospin® Tissue kit (Macherey-Nagel, Düren, Germany), using the Gram-positive protocol variation.

Microbial metatranscriptome protocol: Sputum sample aliquots were immediately transferred to 15 ml falcon tubes containing 1 volume of 0.1 mm zirconia beads (BioSpec, Bartlesville, OK) and 3 volumes of TRIzol® LS reagent (Life Technologies, Grant Island, NY) and vortexed for 10 min at medium speed to mechanically lyse microbial cells but leave the RNA intact. The lysed homogenate was spun for 2 min at 500×g at 4 °C to pellet the zirconia beads. RNA extraction from the supernatant was performed using the Zymo Clean & Concentrator™ 25 kit (Zymo Research, Irvine, CA), using the small RNA removal protocol variation. Following extraction, RNA was treated with RNase-free DNase I (Ambion, Life Technologies: Grant Island, NY).

RNA and cDNA quality and quantity: The quality and quantity of the total RNA were assessed by measurement on the NanoDrop-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and the Agilent Bioanalyzer using the Agilent RNA 6000 Pico kit. The quality and length of the cDNA were assessed using the Agilent Bioanalyzer with the Agilent DNA 7500 kit.

Detailed protocol: Detailed standard protocols was published in the Journal of Visualized Experiment (JoVE) and Chapter 1. The protocols include initial sample pre-processing and pre-treatment prior to microbial cell and virus enrichment. Due to the large files describing every method, and possible deviations from the standard operating procedure, the website allows researcher to choose the sample type and procedure to be done, and automatically generates an appropriate corresponding workflow.

Bioinformatics: All samples were sequenced using the GS-FLX Titanium chemistry system. Multiplexed SFF sequence data files were separated according to their unique identifiers, and FASTA formatted sequences and corresponding quality scores were extracted using the GS-SFF tools software package (Roche: Brandord, CT).

(i) Data Preprocessing

All datasets were preprocessed using PRINSEQ(Schmieder and Edwards 2011a) to remove low quality reads, reads shorter than 60 bp, duplicate reads, and low complexity reads. The command used was:

```
perl prinseq-lite.pl -verbose -log -fastq file.fastq -derep 1245 -lc_method entropy -lc_threshold 50 -trim_qual_right 15 -trim_qual_left 15 -trim_qual_type mean -trim_qual_rule lt -trim_qual_rule lt -trim_qual_window 2 -trim_tail_left 5 -trim_tail_right 5 -min_len 60 -min_qual_mean 15 -ns_max_p 1 -rm_header
```

Viral and microbial metagenomes were further processed using DeconSeq(Schmieder and Edwards 2011b) to remove all human-like sequences with at least 90% query length coverage and 90% identity. This was done using the web version available at http://edwards.sdsu.edu/deconseq

Reference Databases: The rRNA reference database (hereafter **rRNAdb**) was constructed from bacterial, archaeal and eukaryotic sequences in the SILVA SSURef and LSURef databases(Pruesse et al. 2007). The truncated FASTA files (*_tax_silva_trunc.fasta.tgz) were downloaded from http://www.arb-silva.de/ (data release 108). The sequences were filtered for exact duplicate entries for a given organism ID to reduce redundancy. The remaining 584,956 LSU and SSU sequences were combined into the rRNAdb database and formatted for BLAST searches using the formatdb command. The taxonomy information for sequences in the rRNAdb was retrieved from the FASTA file headers. The SILVA taxonomy differed from the NCBI taxonomy due to incorrect annotations in their submissions (such as Oryza sativa classified sequences that were of bacterial origin).

The NCBI non-redundant protein (hereafter **NR**) database (version Feb 14, 2012) was downloaded from the NCBI FTP server: ftp://ftp.ncbi.nih.gov/blast/db/

The NCBI non-redundant nucleotide (hereafter **NT**) database (version Feb 14, 2012) was downloaded from the NCBI FTP server: ftp://ftp.ncbi.nih.gov/blast/db/

The viral database (created Feb 23, 2012) includes: 4,019 unique viral genomes downloaded from the NCBI FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/ and ftp://ftp.ncbi.nih.gov/refseq/release/viral/

(ii) Database searches

Database searches were performed using the BLAST program. Unless specified, the default command-line options were used. Fine-tuning of the options based on the characteristics of the input data may yield better performance and/or results. Analysis of the BLAST output was performed using in-house Perl scripts. BLAST version 2.2.24 was downloaded from: ftp://ftp.ncbi.nih.gov/blast/executables/release/LATEST/

(iii) *Data Analysis*

Annotation of the metagenomes and metatranscriptomes: rRNA-like and non-rRNA reads were identified in the preprocessed metatranscriptomes using BLASTn against the SILVA database (threshold of 50% query coverage and 75% alignment identity). Non-rRNA reads were annotated using BLASTx against the NCBI non-redundant protein database (threshold of 40% identity over at least 60% of the query sequence).

The preprocessed metagenomes were annotated using BLASTn against the NCBI nucleotide database (threshold of 40% identity over at least 60% of the query sequence). Sequences assigned to the phylum Chordata and to vector or synthetic sequences were identified and removed. Virome sequences were then compared against an in-house boutique viral database containing 4,019 unique viral genome sequences (Supplementary Note 1) using tBLASTx (threshold of 40% identity over at least 60% of the query sequence).

The abundance of each virus was normalized to the size of the metagenomes and weighted by the total number of base pairs in the database divided by the length of each viral sequence as described in Willner et al. (2009)(McDonnell et al. 1975). The normalized abundance values for phages that have >4 hits in at least one sample are shown in Supplementary Table 9. The normalized values for all phages belonging to the same putative host were summed and the top 21 were plotted as a bar graph as shown in Fig. 2a.

Best hit designation: The *best hit* designation was assigned to the alignment with the highest coverage, identity and score values within the specified thresholds. For BLASTx against NR, If there were multiple amino acid alignments (within the top 50 BLAST hits) against the same database sequence without overlap in both the query and database sequence, and within the length of the query sequence, the combined coverage, identity and score values were calculated for each query sequence to account for possible frame-shifts.

Taxonomic and functional assignments: The query sequence taxonomy and/or function were assigned based on the best matching database sequence(s). If there were multiple best hits with the same coverage, identity and score values that belonged to different taxa, or the matching database sequence belonged to different taxa, then the taxonomies/functions were randomly assigned using 100,000 bootstraps. This approach is similar to assigning an equal fraction to all possible taxa, but additionally provides the standard deviation for each assigned mean value. Query sequences with no BLAST hits and those unassigned due to the defined threshold were classified as “unassigned” or “unknown”. The diversity of microbiomes was calculated based on the number of bacterial species identified in the datasets.

Clustering of the functional assignments: The random forest was generated using the R package randomForest version 4.6-2 using default parameters with the exceptions of using a seed of 1 and growing to 100,000 trees. PAM clustering is a more robust version of k-mean clustering. It was performed using the cluster package version 1.13.3 using the default parameters with exceptions where the number of clusters was set to three and using (1-the proximity value calculated by the random forest).

Circular plots: Circular plots were generated using Circos version 0.56(Krzywinski et al. 2009). The coverage values are based on reference mapping using a modified version of BWA-SW 0.5.9 (Li and Durbin 2010). The reference mapping of Rothia sp. reads against the published reference genome DY-18 was performed using a minimum threshold of at least 80% nucleotide identity on 60% coverage value. Therefore, the Rothia sp. reads identified here have at least 80% nucleotide identity on the mapped region compared to the reference genome.

Calculating diversity for microbial metagenomes: Diversity is calculated based on the number of bacterial species identified in the datasets. The predicted number of species is the sum of all different species assigned to the bacterial domain based on the selected threshold of BLASTn against the NCBI nucleotide (NT) database. Species richness was calculated as:

$$\text{Richness} = E + \left(\frac{E-1}{E}\right) * \text{k}$$

where:

E = total number of species assigned to the bacterial domain

k = number of unique species

Simpson's Index was calculated as:

$$\text{D} = \frac{-\sum_{i=1}^{s}(\,n_i(n_i\ -\ 1))}{N\ (\,N-1)}$$

where:

n_i = number of individuals of species *i* that are counted

N = total number of all individuals counted

Shannon's Index was calculated as:

$$\text{H}' = -\sum_{i=1}^{S} \text{p}_\text{i} \ln \text{p}_\text{i}$$

where:

p_i = fraction of the individuals belonging to the *i*-th species

Analysis of metabolic pathways using KEGG database and HUMAnN pipeline: Chordata, synthetic, and vector sequences were identified and removed from the datasets based on the BLAST results. Additional potential Chordata sequences in the metatranscriptomes were removed based on the BLASTx against NR results without any threshold parameters to account for errors (for example, frame shifts) that caused too short alignments. All the remaining sequences from metagenomes, viromes and metatranscriptomes were compared by BLASTx against the Kyoto Encyclopedia of Genes and Genomes (KEGG) protein database (as of April 3, 2011). The best hits and the abundance of each pathway were identified and normalized using HUMAnN version 0.98. Gene-level abundances are normalized relative to sequencing depth and gene length, very much like RPKM for single-organism transcriptomics, and pathway-level abundances are normalized relative to the number of genes in the pathway, The normalized pathway abundance values were then used to calculate similarities between samples based on 100,000 random forests and to partition the samples using Partitioning Around Medoids (PAM) clustering. Multidimensional scaling (MDS) was used for plotting. The 20 pathways with highest variance between the three sample groups (viromes, microbiomes, and metatranscriptomes) were used in a principle component analysis (PCA) (Supplementary Fig. 6). The heatmap (Supplementary Fig. 5) was plotted based on the Euclidian distance between the metabolic profiles of each sample.

Supplementary Note 2: The effect of cDNA nebulization on transcript lengths

The nebulization step during the Roche/454 library preparation is a potential source of transcript size-induced bias as the size of our cDNA ranged from 50 – 4,000 bp. We investigated the effect of nebulization followed by the selection of 400 – 1000 bp fragments during Roche-454 platform-specific library preparation on the length of the translated polypeptide sequences. Eight metatranscriptomes were generated in pairs of two in parallel for this purpose. The median protein length was calculated as suggested by (Brocchieri and Karlin 2005) using the microbial genes identified through BLASTx against the NCBI protein (nr) database (Supplementary Table 7). The overall median translated protein length from our results was in the upper range compared to previously reported findings. This bias might be due to initial elimination of small RNA and all RNA below 200 nucleotides using the Zymo RNA Clean & Concentrator^{TM} kit for all RNA clean-up steps. However, there was no difference in the relative translated protein lengths between the nebulized and non-nebulized libraries from the same sample (Supplementary Fig. 3).

References

Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. NAR 33:3390–3400. doi: 10.1093/nar/gki615

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: An information aesthetic for comparative genomics. Genome Res 19:1639–1645. doi: 10.1101/gr.092759.109

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595. doi: 10.1093/bioinformatics/btp698

McDonnell M, Ronda-Lain C, Tomasz A (1975) "Diplophage": A bacteriophage of Diplococcus pneumoniae. Virology 63:577–582. doi: 10.1016/0042-6822(75)90329-3

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. NAR 35:7188 –7196. doi: 10.1093/nar/gkm864

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6:e17288. doi: 10.1371/journal.pone.0017288

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protocols 4:470–483. doi: 10.1038/nprot.2009.10

Supplementary Table 3.1: Summary of sequencing data from published microbial metatranscriptomes using various rRNA removal methods.

Reference	Environment/ Host	Sequencing method	Size of dataset	rRNA Removal /mRNA enrichment method	Total rRNA	Bacterial fraction of total rRNA	Included*
Ottesen E.A. ***et al.*** **(2011)(Ottesen et al. 2011)**	Marine	Roche-454	529.7 Mbp	Hybridization; Polyadenylation	33-40%		MG
Feike J. ***et al.*** **(2011)(Feike et al. 2011)**	Marine	Roche-454	~76,104,800	Hybridization; Polyadenylation	49-76%	N/A	RT
Gosalbes M.J. ***et al.*** **(2011)(Gosalbes et al. 2011)**	Human Gut	Roche-454	85,000,000	Polyadenylation	87-96%	N/A	None
Steward F.J. ***et al.*** **(2011)(Stewart et al. 2011)**	Marine	Roche-454	275,200,000	Sample-specific hybridization, Polyadenylation	37-61%	N/A	MG
Turnbaugh P.J. ***et al.*** **(2010)(Turnbaugh et al. 2010)**	Human Gut	Illumina GAII	>490 Mbp	Hybridization	N/A	N/A	MG, AM
Wu J. ***et al.*** **(2010)(Wu et al. 2011)**	Marine	Roche-454	22,300,000	None	N/A	85.82%	CL
Wittekindt N.E. ***et al*** **(2010)(Wittekindt et al., n.d.)**	Mule deer	Roche-454	110,678,290	Hybridization	N/A	<0.01%	MG, AM
McCarren J. ***et al.*** **(2010)(McCarren et al. 2010)**	Marine	Roche-454	N/A	None	94-97%	N/A	MG, FC

Supplementary Table 3.1: Summary of sequencing data from published microbial metatranscriptomes using various rRNA removal methods. (continued)

Reference	Environment/ Host	Sequencing method	Size of dataset	rRNA Removal /mRNA enrichment method	Total rRNA	Bacterial fraction of total rRNA	Included*
Vila-Costa *et al*. (2010)(Vila-Costa et al. 2010)	Marine	Roche-454	126,713,774	Exonuclease; hybridization; polyadenylation	49%	81.7%	16S
Steward F.J. *et al.* (2010) (Stewart, Ottesen, and DeLong 2010)	Marine	Roche-454	625,432,150	Sample-specific hybridization	52-61%	29%-45%	None
Hewson I. *et al.* (2010)(Hewson et al. 2010)	Marine	Roche-454	>200,000,000	Exonuclease; hybridization; polyadenylation	<66%	N/A	None
Turnbaugh P.J. *et al.* (2009)(Turnbaugh et al. 2009)	Humanized mouse gut	Illumina GAII	N/A	Hybridization	N/A	N/A	MG, AM
Poretsky R.S. *et al.* (2009)(Poretsky et al. 2009)	Marine	Roche-454	>24,042,200	Exonuclease; hybridization; polyadenylation	37%	N/A	16S-clone libraries, RT
Hewson I. *et al.* (2009)(Hewson et al. 2009)	Marine	Roche-454	N/A	Exonuclease; hybridization; polyadenylation	N/A	N/A	None
Gillbert J.A. *et al.* (2008)(Gilbert et al. 2008)	Marine	Roche-454	323,161,989	Hybridization	<0.33%		MG, RT, PCR

Supplementary Table 3.1: Summary of sequencing data from published microbial metatranscriptomes using various rRNA removal methods. (continued)

Reference	Environment/ Host	Sequencing method	Size of dataset	rRNA Removal /mRNA enrichment method	Total rRNA	Bacterial fraction of total rRNA	Included*
Urich T. *et al.* (2008)(Urich et al. 2008)	Soil	Roche-454	25,324,278	N/A	75%	73%	MG
Frias-Lopez J. *et al.*(2008)(Frias-Lopez et al. 2008)	Marine	Roche-454	14,700,000	Polyadenylation	53%		MG, RT
Leininger S. *et al.*(2006)(Leininger et al. 2006)	Soil	Roche-454	30,270,000	None	>34%	N/A	RT
Poretsky R.S. *et al.* (2005)(Poretsky et al. 2005)	Marine	Sanger	407clones	Hybridization; primer-targeted PCR	<20%	N/A	None

*16S: 16S-amplicon sequencing; AM: Amplicon; CL: Clone libraries; FC: Flow cytometry; MA: Microarray; MG: Metagenomes; N/A: Information not available; RT: Reverse transcription-qPCR; PCR: PCR validation

Supplementary Table 3.2: Commercially-adapted rRNA removal and mRNA enrichment kits and peer-reviewed publications reporting their usage.

Name	Methods	Company	Method Based on	Recommended for	References
MICROB *Express*^{TM}	Bacterial rRNA removal	Ambion	Subtractive Hybridization	Intact rRNA	This study;Poretsky (2005); Gilbert (2008); Poretsky (2009); Hewson I. (2009); Turnbaugh P.J. (2009); Hewson I. (2010);Vila-Costa M. (2010); Turnbaugh P.J. (2010)
MICROB *Enrich*^{TM}	Mammalian rRNA removal	Ambion	Subtractive Hybridization	Intact rRNA	This study; Stewart F.J. (2011)
Ribo-Zero^{TM}	Gram-negative/Gram positive rRNA removal	Epicentre	Microspheres	Intact and fragmented rRNA(Sooknanan, Pease, and Doyle 2010)28	None
Ribo-Zero^{TM}	Human/Mouse/ Rat rRNA removal	Epicentre	Microspheres	Intact and fragmented rRNA(Sooknanan, Pease, and Doyle 2010)28	None
Ribo-Zero^{TM}	Human & Bact rRNA removal	Epicentre	Microspheres	Intact and fragmented rRNA(Sooknanan, Pease, and Doyle 2010)28	*This study
mRNA-ONLY^{TM}	Prokaryotic mRNA enrichment	Epicentre	Terminator^{TM} 5'-Phosphate-Dependent (Sooknanan, Pease, and Doyle 2010)28 Exonuclease	Intact rRNA	Poretsky (2009); Hewson I. (2009); Hewson I. (2010); Vila-Costa M. (2010)

Supplementary Table 3.2: Commercially-adapted rRNA removal and mRNA enrichment kits and peer-reviewed publications reporting their usage. (continued)

Name	Methods	Company	Method Based on	Recommended for	References
WT-Ovation^{TM} Pico RNA Amplification System	Prokaryotic mRNA enrichment	NuGEN	Random-amplification	N/A	Wu J. (2010)
MessageAmp^{TM}	Prokaryotic mRNA enrichment	Ambion	Polyadenylation and amplification	Prokaryotic mRNAs	Poretsky (2009); Hewson I. (2010); Vila-Costa M. (2010); Stewart F.J. (2010); Stewart F.J. (2011) Gosalbes M.J. (2011)

* The Ribo-Zero^{TM} kits were combined into the Ribo-Zero^{TM} Meta Kit by the Research and Development group of Epicenter Biotechnology for our purposes in generating human-associated microbial metatranscriptomes.

Supplementary Table 3.3: Library characteristics of the eight viromes generated from sputum samples from three CF patients.

	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Total number of reads	**224,859**	**87,891**	**106,189**	**93,301**	**140,020**	**1,558**	**272,552**	**217,438**
Preprocessed reads [a]	**109,389 (49%)**	**73,624 (84%)**	**67,070 (63%)**	**82,011 (88%)**	**68,617 (49%)**	**1,137 (73%)**	**215,808 (79%)**	**158,432 (73%)**
Number of bases	47,239,573	33,351,525	28,922,479	27,667,695	29,386,841	243,986	95,205,805	69,581,811
Mean read length	432	453	431	337	428	215	441	439
Host sequences [b]	240 (0.21%)	526 (0.71%)	28 (0.04%)	79,774 (97.27%)	13 (0.02%)	797 (70.10%)	585 (0.27%)	5,859 (3.70%)
Viral hits [c]	**7,214 (6.59%)**	**23,550 (31.99%)**	**4,070 (6.07%)**	**737 (0.90%)**	**4,642 (6.77%)**	**22 (1.93%)**	**6,466 (3.00%)**	**5,981 (3.78%)**
Unassigned Reads [d]	**103,888 (94.97%)**	**60,490 (82.16%)**	**32,780 (48.87%)**	**1,935 (2.36%)**	**68,440 (99.74%)**	**311 (27.35%)**	**105,612 (48.94%)**	**119,551 (75.46%)**

[a] Reads after data pre-processing by PRINSEQ (see Supplementary Note 1).

[b] Human reads identified by DeconSeq plus reads with a best BLASTn hit (NCBI nucleotide database) to the phylum Chordata.

[c] tBLASTx hits against in-house viral genome database. The percentage was calculated using the total number of preprocessed reads.

[d] Reads with no BLASTn hit against the NCBI nucleotide database. The percentage was calculated using the total number of preprocessed reads. Some reads with no BLASTn hit against the NCBI nucleotide database were identified as viral at protein level in the tBLASTx analysis.

Supplementary Table 3.4: Eukaryotic viruses in CF viromes, based on tBLASTx against the in-house viral genome database (see Supplementary Note 1). Other includes viruses that are <2% in the population.

Virome	**Eukaryotic viruses**
CF1-D	Herpesviruses (70.6%), Bovine popular stomatitis virus (5.2%), Torque Teno Virus (4.9%), Molluscum contagiosum virus (2.7%), Adenoviruses (3.1%), Other (13.6%)
CF1-E	Herpesviruses (75.1%), Adenoviruses (7.6%), Other (17.3%)
CF1-F	Herpesviruses (73.2%), Adenoviruses (3.8%), Bovine popular stomatitis virus (3.7%), Molluscum contagiosum virus (3.7%), Other (15.7%)
CF4-A	Herpesviruses (42.4%), Taterapox virus (20.3%), Rabbit fibroma virus (7.8%), Other (29.4%)
CF4-B	Herpesviruses (75.4%), Bovine popular stomatitis virus (5.2%), Adenoviruses (3.4%), Molluscum contagiosum virus (3.0%), Other (12.9%)
CF4-C	Herpesviruses (61.5%), Bovine popular stomatitis virus (15.4%), Molluscum contagiosum virus (7.7%), Other (15.4%)
CF5-A	Herpesviruses (68.9%), Adenoviruses (4.0%), Bovine popular stomatitis virus (3.6%), Molluscum contagiosum virus (2.9%), Other (20.5%)
CF5-B	Herpesviruses (73.6%), Bovine popular stomatitis virus (4.3%), Adenoviruses (4.1%), Crocodilepox virus (2.7%), Other (15.4%)

Supplementary Table 3.5: Library characteristics of the eight microbiomes generated from sputum samples from three CF patients.

	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Total number of reads	**165,433**	**128,574**	**54,673**	**238,702**	**234,241**	**194,487**	**54,353**	**35,584**
Preprocessed reads[a]	**154,882 (94%)**	**117,456 (91%)**	**51,769 (95%)**	**207,480 (87%)**	**211,252 (90%)**	**183,604 (94%)**	**50,919 (94%)**	**32,991 (93%)**
Number of bases	60,206,563	46,024,694	21,977,934	70,148,108	80,082,439	73,458,029	19,139,110	11,920,534
Mean read length	389	392	425	338	379	400	376	361
Host sequences [b]	130,484 (84.25%)	37,149 (31.63%)	6,935 (13.40%)	196,872 (94.89%)	204,882 (96.98%)	119,190 (64.92%)	47,693 (93.66%)	31,805 (96.41%)
Non-host sequences	**24,398 (15.75%)**	**80,307 (68.37%)**	**44,834 (86.60%)**	**10,608 (5.11%)**	**6,370 (3.02%)**	**64,414 (35.08%)**	**3,226 (6.34%)**	**1,186 (3.59%)**
Bacterial hits [c]	**14,691 (9%)**	**67,780 (58%)**	**40,825 (79%)**	**3,834 (2%)**	**405 (0.15%)**	**41,636 (23%)**	**1,034 (2%)**	**247 (1%)**
Unassigned reads [d]	9,550 (6%)	12,330 (10%)	3,952 (8%)	6,585 (2%)	5,771 (3%)	22,583 (12%)	2,163 (4%)	914 (3%)

[a] Reads after data pre-processing by PRINSEQ (see Supplementary Note 1).

[b] BLASTn hits against the NCBI nucleotide database (threshold of 40% identity over at least 60% of the query sequence).

[c] Percentage was calculated using the number of preprocessed reads.

[d] Reads with no BLASTn hit against the NCBI nucleotide database. Percentage was calculated using the total number of preprocessed reads.

Supplementary Table 3.6: Bacterial community diversity in eight microbiomes generated from sputum samples from three CF patients.

Patient ID	Time point	Bacterial Sequences	Species	Richness	Evenness	Shannon index	Simpson index
CF1	D	14,691	245.29	245.94	0.45	2.49	0.16
	E	67,781	131.79	132.43	0.31	1.49	0.40
	F	40,825	109.48	110.14	0.32	1.50	0.34
CF4	A	3,834	91.26	91.88	0.31	1.40	0.46
	B	405	56.96	57.51	0.67	2.72	0.11
	C	41,636	256.36	257.09	0.38	2.12	0.21
CF5	A	1,034	51.00	51.64	0.53	2.08	0.26
	B	247	24.44	25.01	0.38	1.23	0.57

Supplementary Table 3.7: Gaps identified from the mapping of CF1-E microbiome reads against *Rothia mucilaginosa* DY-18. Genes that were not presence or not functional are highlighted in the table.

Size (bp)	**Start**	**End**	**Gene(s) affected/missing; (Copies in the genome)**
1330	50122	51452	glycosyltransferase involved in cell wall biogenesis CDS
7751	119456	127207	hemoglobin-like flavoprotein CDS (2), acyl-coenzyme A synthetase CDS (3), carbonic anhydrase CDS (3), Rhs family protein CDS (4), protein involved in cell division CDS (2), riboflavin synthase alpha chain CDS (1), hypothetical protein CDS, transcriptional regulator CDS (62), uncharacterized protein conserved in archaea CDS (37), alpha-mannosidase CDS (2), protein involved in cell division CDS (2), ABC-type phosphate transport system (1), ATPase component CDS (49)
2483	136791	139274	Exopolysaccharide (2)
1497	149641	151138	dinucleotide-utilizing enzyme CDS (3), hypothetical protein CDS
1729	260126	261855	SAM-dependent methyltransferase related to tRNA (uracil-5-)-methyltransferase CDS (2), permease of the major facilitator superfamily CDS (17)
1724	283683	285407	predicted hydrolase CDS (16)
5035	319381	324416	acyl-CoA dehydrogenase CDS (5), adenylate cyclase family 3 CDS (4), putative peptidoglycan-binding domain-containing protein CDS (2), ABC-type antimicrobial peptide transport system(7), ATPase component CD (49), permease component CDS (48)
2007	373312	375319	flavodoxin CDS (1)
1849	588374	590223	predicted membrane protein CDS (18)
3038	595573	598611	UDP-N-acetylglucosamine enolpyruvyl transferase CDS (2)
3628	620890	624518	ABC-type amino acid transport system (4), permease component CDS (48), ABC-type amino acid transport/signal transduction system (4), periplasmic component CDS (33), 6-pyruvoyl-tetrahydropterin synthase CDS (1), ABC-type polar amino acid transport system (3), ATPase component CDS (49)
2658	638197	640855	non-coding region
2071	655324	657395	predicted transcriptional regulator CDS, di- and tricarboxylate transporter CDS (1)

Supplementary Table 3.7: Gaps identified from the mapping of CF1-E microbiome reads against *Rothia mucilaginosa* DY-18. Genes that were not presence or not functional are highlighted in the table. (continued)

Size (bp)	**Start**	**End**	**Gene(s) affected/missing; (Copies in the genome)**
2555	686303	688858	non-coding region
1379	816600	817979	hypothetical protein CDS
1682	1021448	1023130	non-coding region
1746	1138421	1140167	predicted transcriptional regulator CDS
2818	1216482	1219300	putative peptidoglycan-binding domain-containing protein CDS (2), 3-phosphoglycerate kinase CDS (1), glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase CDS (2)
2955	1237151	1240106	glutathione S-transferase CDS (1), mismatch repair ATPase CDS (1)
2415	1331409	1333824	type II secretory pathway (3), component PulF CDS (2)
1689	1341702	1343391	nuclear protein export factor CDS (1)
1757	1374974	1376731	non-coding region
3117	1379244	1382361	non-coding region
1074	1410289	1411363	predicted permease CDS (6)
2091	1433964	1436055	amidase related to nicotinamidase CDS (1), hypothetical protein CDS
1048	1671561	1672609	chromosome segregation ATPase CDS (3)
2888	1788138	1791026	predicted glycosyltransferase CDS (2), ABC-type multidrug transport system (21), ATPase and permease component CDS (8), protein-tyrosine-phosphatase CDS (2)
1303	1796434	1797737	non-coding region
2115	1817898	1820013	hypothetical protein CDS, thiamine pyrophosphate-requiring enzyme CDS (3) , NTP pyrophosphohydrolase including oxidative damage repair enzyme CDS (10)
1540	1836359	1837899	soluble lytic murein transglycosylase CDS (2)
1618	1889247	1890865	non-coding region
2872	1895728	1898600	uncharacterized low-complexity protein CDS

Supplementary Table 3.7: Gaps identified from the mapping of CF1-E microbiome reads against *Rothia mucilaginosa* DY-18. Genes that were not presence or not functional are highlighted in the table. (continued)

Size (bp)	**Start**	**End**	**Gene(s) affected/missing; (Copies in the genome)**
1167	1920109	1921276	non-coding region
1039	1924608	1925647	predicted transcriptional regulator CDS
1167	2079868	2081035	permease of the major facilitator superfamily CDS (17)
3839	2084616	2088455	protein chain release factor B CDS (2), ABC-type amino acid transport/signal transduction system (4), periplasmic component CDS (27), cystathionine beta-lyase/cystathionine gamma-synthase CDS (3), hypothetical protein CDS, cystathionine beta-lyase/cystathionine gamma-synthase CDS (3)
1970	2088513	2090483	phosphoribosylaminoimidazole synthetase CDS (2)
2751	2105030	2107781	hypothetical protein CDS, type IV secretory pathway (3), VirD4 component CDS
2896	2162251	2165147	cobalamin biosynthesis protein CobN CDS (2), preprotein translocase subunit SecD CDS (3), hypothetical protein CDS
1606	2177676	2179282	Rhs family protein CDS (4)
1656	2209669	2211325	predicted transcriptional regulator CDS
1606	2177676	2179282	Rhs family protein CDS (4)
1656	2209669	2211325	predicted transcriptional regulator CDS

Supplementary Table 3.8: Metatranscriptomes evaluating the effect of nebulization on transcript length.

	CF1-B	CF1-B nebulized	CF1-C	CF1-C nebulized	CF2-A	CF2-A nebulized	CF3-A	CF3-A nebulized
Preprocessed reads	**68,516**	**69,457**	**59,121**	**67,018**	**54,237**	**23,981**	**37,188**	**67,408**
Mean read length	305	294	312	305	273	262	287	290
Median protein length (amino acids)	387	377	377	378	401	345	381	412
Total rRNA reads	**47,033 (68.6%)**	**49,151 (70.8%)**	**40,700 (68.8%)**	**46,907 (70.0%)**	**5,886 (10.9%)**	**3,356 (14.0%)**	**3,498 (9.4%)**	**7,723 (11.5%)**
Microbial rRNA	40,741 (59.5%)	41,725 (60.1%)	33,542 (56.7%)	37,548 (56.0%)	1,634 (3.0%)	832 (3.5%)	216 (0.6%)	494 (0.7%)
Eukaryota rRNA	6,292 (9.2%)	7426 (10.7%)	7158 (12.1%)	9,358 (14.0%)	4252 (7.8%)	2,524 (10.5%)	3281 (8.8%)	7,229 (10.7%)
Non-rRNA reads	**21,483 (31.4%)**	**20,306 (29.2%)**	**18,421 (31.2%)**	**20,111 (30.0%)**	**48,351 (89.1%)**	**20,625 (86.0%)**	**33,690 (90.6%)**	**59,685 (88.5%)**
Total NR hits	**8,269 (12%)**	**7,786 (11%)**	**5,510 (9%)**	**6,172 (9%)**	**15,670 (29%)**	**6,682 (28%)**	**9,745 (26%)**	**17,256 (26%)**
Microbial	1,891	1,691	1,143	1,370	697	295	525	985
Eukaryota	6,357	6,080	4,338	4,781	14,900	6,355	9,191	16,194
Unassigned reads	**13,214 (19%)**	**12,529 (18%)**	**12,911 (22%)**	**13,939 (21%)**	**32,681 (60%)**	**13,943 (58%)**	**23,945 (64%)**	**42,429 (63%)**

Supplementary Table 3.9: Library characteristics of the metatranscriptomes evaluating rRNA depletion methods.

	CF1-D			CF1-F			CF4-B			CF4-C		
	None	**Ambion**	**Epicntr**	**None**	**Ambion**	**Epicntr**	**None**	**Ambion**	**Epicntr**	**None**	**Ambion**	**Epicntr**
Preprocessed reads	**2,088**	**26,591**	**1,991**	**40,876**	**11,692**	**25,238**	**19,728**	**11,170**	**32,737**	**31,791**	**25,813**	**36,172**
No. of bases	573,851	6,479,698	488,744	10,696,881	2,964,880	6,822,536	4,591,996	2,577,370	8,469,911	7,637,275	6,425,940	9,668,061
Mean read length	275	244	245	262	254	270	233	231	259	240	249	267
Total rRNA reads	**1,737 (83.2%)**	**4,355 (16.4%)**	**91 (4.6%)**	**29,499 (72.2%)**	**7,093 (60.7%)**	**17,267 (68.4%)**	**5,285 (26.8%)**	**8,643 (77.4%)**	**291 (0.9%)**	**16,371 (51.5%)**	**15,187 (58.8%)**	**1,761 (4.9%)**
Microbial rRNA	1,414 (67.7%)	66 (0.2%)	32 (1.6%)	19,978 (48.9%)	4,562 (39.0%)	12,035 (47.7%)	23 (0.1%)	6,540 (58.5%)	227 (0.7%)	6,916 (21.8%)	8,868 (34.4%)	1,076 (3.0%)
Eukaryota rRNA	323 (15.5 %)	4,288 (16.1%)	59 (3.0%)	9,520 (23.3%)	2,530 (21.6%)	5,232 (20.7%)	5,262 (26.7%)	2,193 (19.6%)	64 (0.2%)	9,455 (29.7%)	6,319 (24.5%)	683 (1.9%)
% rRNA removed*	**0%**	**80%**	**95%**	**0%**	**16%**	**5%**	**0%**	**-189%**	**97%**	**0%**	**-14%**	**91%**
Non-rRNA reads	**351 (16.8%)**	**22,236 (83.6%)**	**1,900 (95.4%)**	**11,377 (27.8%)**	**4,599 (39.3%)**	**7,971 (31.6%)**	**14,443 (73.2%)**	**2,527 (22.6%)**	**32,446 (99.1%)**	**15,420 (48.5%)**	**10,662 (41.2%)**	**34,411 (95.1%)**
Total NR hits	**102 (4.9%)**	**5964 (22.4%)**	**691 (34.7%)**	**3,327 (8.1%)**	**1,547 (13.2%)**	**2,857 (11.3%)**	**4,938 (25.0%)**	**722 (6.5%)**	**10,751 (32.8%)**	**5,905 (18.6%)**	**4,455 (17.3%)**	**15,766 (43.6%)**
Eukaryotic	74	5,244	407	2,790	1,262	2,524	4,614	528	10,227	4,553	2,891	8,274
Bacterial	26	674	283	520	282	312	287	191	471	1,326	1,554	7,442
Unassigned reads	**249 (11.9%)**	**16,272 (61.2%)**	**1,209 (60.7%)**	**8050 (19.7%)**	**3,052 (26.1%)**	**5,114 (20.3%)**	**9,505 (48.2%)**	**1,805 (16.2%)**	**21,695 (66.3%)**	**9,515 (29.9%)**	**6,171 (23.9%)**	**18,645 (51.5%)**

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample.

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Yersinia phage L-413C	67188	0.85	0.17	1.27	0.00	1.52	0.00	1.96	1.66
Xanthomonas phage OP2	52892	1.04	0.00	0.77	0.00	0.63	0.00	1.30	1.62
Vibrio phage VP882	70110	1.47	0.00	1.10	2.05	0.58	0.00	1.48	0.47
Vibrio phage nt-1 sensu lato	76312	0.00	0.13	0.24	0.00	0.00	0.00	0.22	0.13
Vibrio phage kappa	56692	0.80	0.00	0.00	0.00	0.00	0.00	1.96	1.48
Vibrio phage K139	64787	0.80	0.00	0.00	0.00	0.00	0.00	1.96	1.48
Tsukamurella phage TPA2	69059	4.35	0.51	3.56	0.00	4.03	6.06	3.58	3.49
Thermus phage P23-77	69378	3.16	0.00	1.83	3.01	3.14	0.00	2.44	1.60
Thermus phage P23-45	68090	0.83	0.06	0.72	0.00	1.07	0.00	0.90	1.04
Thermus phage IN93	68999	3.46	0.25	2.71	0.00	3.36	0.00	2.44	3.22
Synechococcus phage Syn5	51376	1.87	0.22	0.56	0.00	0.00	0.00	0.67	1.19
Synechococcus phage S-SSM5	64494	0.96	0.06	0.32	0.00	0.77	0.00	0.77	0.51
Synechococcus phage S-ShM2	58554	0.42	0.09	0.93	0.00	0.15	0.00	0.75	0.22
Synechococcus phage S-CBS2	68427	1.52	0.33	0.68	0.00	1.04	0.00	0.89	1.51
Synechococcus phage S-CBS1	110865	0.86	0.00	0.78	0.00	2.16	0.00	0.54	1.30
Stx2-converting phage 1717	69904	0.99	0.00	0.50	0.00	0.53	0.00	0.91	1.02
Streptomyces phage VWB	164602	3.71	0.25	3.97	0.00	4.01	0.00	3.70	3.73
Streptomyces phage phiSASD1	2404	3.21	0.27	2.15	0.00	2.74	0.00	1.98	2.48
Streptomyces phage phiC31	51478	1.31	0.00	1.36	0.00	1.51	0.00	1.41	1.28
Streptomyces phage mu1/6	56852	5.31	1.25	4.70	0.00	5.55	0.00	5.30	5.73
Streptococcus phage SM1	73453	0.00	3.50	0.00	0.00	0.00	0.00	0.00	0.51
Streptococcus phage Dp-1	52297	2.42	10.50	0.00	0.00	0.42	0.00	0.57	0.61
Stenotrophomonas phage S1	75811	1.42	0.13	1.66	0.00	2.14	0.00	2.65	1.19
Sodalis phage phiSG1	68952	1.34	0.00	0.00	0.00	1.19	0.00	0.34	0.65

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Sinorhizobium phage PBC5	53572	2.36	0.42	1.22	0.00	2.06	0.00	2.04	2.21
Shigella phage phiSboM-AG3	50968	0.30	0.03	0.76	0.00	0.16	0.00	0.93	0.46
Salmonella phage SFP10	42289	0.39	0.00	0.35	0.00	0.45	0.00	0.52	0.55
Roseobacter phage RDJL Phi 1	64807	1.47	0.08	0.61	0.00	1.15	0.00	1.34	0.92
Rhodothermus phage RM378	54512	0.51	0.20	0.22	0.00	0.00	0.00	0.40	0.16
Rhodobacter phage RcapMu	58471	2.18	0.00	2.44	0.00	3.05	0.00	1.26	2.20
Rhizobium phage 16-3	2244	4.10	0.46	3.22	0.00	4.42	0.00	3.06	3.44
Ralstonia phage RSM3	68569	0.00	0.00	4.73	0.00	1.65	0.00	1.55	0.00
Ralstonia phage RSM1	155445	0.00	0.00	5.55	0.00	0.00	0.00	1.54	0.00
Ralstonia phage RSL1	15664	1.63	0.06	1.06	0.00	0.99	0.00	0.96	1.43
Ralstonia phage RSB1	51621	1.71	0.12	1.60	0.00	1.96	0.00	1.54	0.76
Ralstonia phage phiRSA1	49136	1.89	0.00	2.13	0.00	2.98	0.00	2.48	1.80
Pseudomonas phage YuA	59798	2.68	0.29	1.60	0.00	2.34	0.00	2.09	2.54
Pseudomonas phage SN	69777	0.97	0.08	0.56	0.00	0.47	0.00	1.07	0.61
Pseudomonas phage phiKZ	70654	0.18	0.00	0.30	0.51	0.00	0.00	0.42	0.21
Pseudomonas phage phikF77	75931	1.30	0.12	1.82	0.00	1.87	0.00	1.37	1.74
Pseudomonas phage phiCTX	56276	1.35	0.15	2.02	0.00	1.65	0.00	1.11	1.62
Pseudomonas phage phi-2	57050	0.36	0.00	0.59	1.92	1.47	0.00	1.54	0.60
Pseudomonas phage phi15	59471	1.67	0.00	1.07	0.00	1.55	0.00	0.43	0.00
Pseudomonas phage PB1	52047	0.91	0.00	1.11	0.00	1.57	0.00	0.57	1.20
Pseudomonas phage PAJU2	68450	1.21	0.11	2.61	0.00	0.86	0.00	2.17	0.71
Pseudomonas phage MP38	153766	1.65	0.25	1.89	0.00	1.96	0.00	2.34	2.04
Pseudomonas phage MP22	155372	2.71	0.14	2.75	0.00	2.19	0.00	2.16	2.37
Pseudomonas phage M6	50913	1.84	0.21	2.16	0.00	1.70	0.00	1.51	2.19

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Pseudomonas phage LUZ19	156102	3.61	0.23	2.32	0.00	3.53	0.00	1.75	2.34
Pseudomonas phage LMA2	50550	1.57	0.08	1.82	0.00	2.14	0.00	0.92	1.17
Pseudomonas phage LKD16	64562	1.33	0.00	1.01	1.92	1.57	0.00	0.96	1.60
Pseudomonas phage LKA1	47057	1.31	0.24	1.03	0.00	1.58	0.00	1.21	1.48
Pseudomonas phage LIT1	41901	1.14	0.00	1.32	0.00	1.51	0.00	1.43	1.15
Pseudomonas phage F116	58037	3.24	0.23	3.37	1.51	3.27	0.00	3.08	3.46
Pseudomonas phage F10	52250	2.20	0.25	2.58	2.83	2.61	0.00	1.92	2.89
Pseudomonas phage EL	70797	0.55	0.08	0.14	0.65	0.49	4.33	0.33	0.32
Pseudomonas phage DMS3	52141	1.70	0.00	1.78	0.00	1.12	0.00	2.14	1.84
Pseudomonas phage D3112	59598	1.63	0.17	1.60	0.00	1.94	0.00	1.78	1.38
Pseudomonas phage D3	41441	1.21	0.09	1.35	0.00	1.68	0.00	1.22	1.75
Pseudomonas phage B3	64511	1.54	0.00	2.14	0.00	2.00	0.00	1.28	1.61
Pseudomonas phage 73	74483	0.89	0.23	1.18	0.00	0.53	0.00	1.54	0.90
Pseudomonas phage 201phi2-1	38234	0.00	0.03	0.10	0.00	0.38	0.00	0.47	0.13
Pseudomonas phage 14-1	2025	0.92	0.11	1.15	0.00	1.73	0.00	0.53	0.98
Prochlorococcus phage Syn33	46365	0.14	0.06	0.17	0.00	0.15	0.00	0.70	0.51
Prochlorococcus phage P-SSM2	26111	0.34	0.13	0.47	0.00	0.45	0.00	0.38	0.30
Planktothrix phage PaV-LD	8140	0.60	0.00	0.00	0.00	0.49	0.00	1.04	0.00
Phage phiJL001	34525	2.13	0.16	1.61	0.00	1.63	0.00	1.97	1.33
Pantoea phage LIMEzero	36798	1.75	0.12	2.18	1.92	0.00	0.00	1.89	1.62
Natrialba phage PhiCh1	38347	2.52	0.00	1.85	0.00	2.52	0.00	2.02	2.26
Myxococcus phage Mx8	37667	2.77	0.30	2.70	0.00	2.88	0.00	2.60	2.53
Mycobacterium phage Wildcat	33350	1.27	0.07	0.87	0.00	1.78	0.00	1.86	0.96
Mycobacterium phage Wee	32172	1.06	0.00	0.89	0.00	1.62	0.00	0.91	0.82

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Mycobacterium phage U2	36596	2.21	0.10	1.54	0.00	1.33	0.00	1.23	1.29
Mycobacterium phage TM4	40003	3.26	0.19	3.60	0.00	3.43	6.27	3.33	3.60
Mycobacterium phage Spud	36949	1.15	0.07	0.85	0.83	1.44	4.75	0.77	1.10
Mycobacterium phage Solon	41834	1.84	0.00	1.42	0.00	1.76	0.00	1.08	1.03
Mycobacterium phage ScottMcG	4680	1.24	0.11	0.93	0.83	1.38	0.00	0.92	1.12
Mycobacterium phage Rosebush	114768	3.46	0.31	2.03	2.45	2.65	0.00	3.10	3.07
Mycobacterium phage Rizal	26537	0.84	0.07	0.74	0.00	1.20	0.00	0.78	0.74
Mycobacterium phage Ramsey	55597	2.12	0.10	2.51	0.00	2.34	0.00	1.68	1.73
Mycobacterium phage Qyrzula	36892	2.42	0.26	2.10	1.48	2.29	0.00	2.26	2.35
Mycobacterium phage Pukovnik	40190	2.66	0.36	3.39	0.00	3.04	0.00	1.62	2.56
Mycobacterium phage Predator	39166	1.26	0.08	0.39	0.00	1.54	0.00	0.91	1.31
Mycobacterium phage Porky	43809	1.29	0.07	1.39	1.36	1.09	0.00	0.68	1.30
Mycobacterium phage PMC	174436	1.66	0.13	1.57	0.00	1.94	0.00	0.89	1.41
Mycobacterium phage PLot	22743	1.93	0.00	1.30	1.51	2.36	0.00	1.14	1.25
Mycobacterium phage Pipefish	77670	3.14	0.44	1.97	0.00	2.77	0.00	2.63	3.22
Mycobacterium phage Phlyer	39245	2.96	0.44	2.41	1.45	3.38	0.00	2.14	2.48
Mycobacterium phage Phaedrus	30889	2.76	0.41	2.71	1.47	3.26	0.00	2.16	2.61
Mycobacterium phage PG1	43033	1.54	0.13	1.20	0.00	1.43	0.00	0.97	1.24
Mycobacterium phage Peaches	31661	2.63	0.20	1.64	0.00	2.18	0.00	2.43	2.01
Mycobacterium phage Pacc40	59866	1.06	0.09	1.69	0.00	1.75	0.00	1.60	1.12
Mycobacterium phage Orion	137947	1.55	0.13	1.35	0.00	1.75	0.00	1.59	1.25
Mycobacterium phage Omega	37235	1.85	0.05	1.91	0.00	1.89	0.00	1.70	1.80
Mycobacterium phage Nigel	84576	3.19	0.40	2.96	0.00	3.55	0.00	2.05	2.97

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Mycobacterium phage Myrna	35372	3.08	0.24	2.87	1.07	3.35	0.00	2.89	3.07
Mycobacterium phage Lockley	36270	2.09	0.00	2.13	0.00	2.11	0.00	1.18	2.07
Mycobacterium phage LeBron	38764	1.00	0.00	1.13	0.00	0.33	0.00	1.21	1.14
Mycobacterium phage L5	11279	1.98	0.10	1.52	0.00	1.75	0.00	1.69	1.09
Mycobacterium phage Kostya	39937	1.12	0.15	1.26	1.37	1.14	0.00	0.95	0.98
Mycobacterium phage Konstantine	39792	0.44	0.08	0.71	0.00	1.25	0.00	1.13	0.72
Mycobacterium phage KBG	43769	1.67	0.10	1.10	0.00	1.35	0.00	1.13	1.42
Mycobacterium phage Jasper	37074	1.47	0.00	1.28	0.00	1.25	0.00	1.05	1.30
Mycobacterium phage Halo	180500	1.69	0.12	1.65	1.94	1.14	0.00	1.70	1.40
Mycobacterium phage Giles	176788	2.42	0.51	3.09	0.00	2.29	0.00	2.42	2.26
Mycobacterium phage Fruitloop	14927	1.30	0.11	1.25	1.61	1.39	0.00	1.30	1.63
Mycobacterium phage Faith1	5386	1.39	0.07	1.28	0.00	1.91	0.00	1.82	1.30
Mycobacterium phage ET08	42575	0.96	0.09	1.02	0.00	1.08	0.00	1.19	1.30
Mycobacterium phage DD5	42519	2.06	0.00	1.64	0.00	2.14	0.00	1.07	2.00
Mycobacterium phage D29	11624	1.61	0.21	2.61	0.00	2.15	0.00	2.21	1.83
Mycobacterium phage CrimD	33593	1.99	0.25	2.83	0.00	2.36	0.00	2.42	3.12
Mycobacterium phage Corndog	94800	2.77	0.31	2.34	0.00	3.04	0.00	2.15	2.33
Mycobacterium phage Cooper	36717	2.62	0.15	2.41	0.00	2.89	0.00	2.23	2.50
Mycobacterium phage Cjw1	63395	1.46	0.07	1.08	1.37	1.43	0.00	1.00	1.15
Mycobacterium phage Che9d	41593	1.34	0.09	1.73	0.00	1.43	0.00	1.83	1.81
Mycobacterium phage Che9c	48502	2.40	0.48	2.40	0.00	2.17	0.00	1.82	2.62
Mycobacterium phage Che8	45251	1.61	0.17	1.03	0.00	1.32	0.00	1.28	1.94
Mycobacterium phage Che12	44385	2.24	0.00	1.49	0.00	1.83	0.00	1.44	2.47
Mycobacterium phage Chah	39732	1.36	0.15	1.43	0.00	1.71	0.00	1.57	1.73

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Mycobacterium phage Catera	40751	1.13	0.06	0.83	0.83	1.47	0.00	0.73	0.95
Mycobacterium phage Cali	33693	1.87	0.18	1.22	0.00	1.91	0.00	0.98	1.62
Mycobacterium phage Bxz2	50625	2.28	0.29	1.26	0.00	1.70	0.00	1.99	2.21
Mycobacterium phage Bxz1	61765	1.13	0.07	0.85	0.00	1.52	0.00	0.91	1.19
Mycobacterium phage Bxb1	36524	1.56	0.33	0.90	1.76	1.91	0.00	1.51	1.49
Mycobacterium phage Brujita	157486	3.02	0.40	2.86	0.00	3.01	0.00	2.07	2.80
Mycobacterium phage BPs	107530	1.71	0.22	1.85	1.95	1.61	0.00	1.54	1.49
Mycobacterium phage Boomer	2460	1.58	0.09	1.54	1.62	1.99	0.00	0.84	1.77
Mycobacterium phage Bethlehem	70579	1.15	0.10	0.55	0.00	1.06	0.00	0.77	0.55
Mycobacterium phage Barnyard	38092	1.15	0.07	0.87	0.00	1.66	0.00	0.90	1.17
Mycobacterium phage Angelica	59199	2.55	0.09	2.22	0.00	2.02	0.00	2.19	2.28
Mycobacterium phage Angel	3625	1.79	0.16	1.96	1.96	1.04	0.00	1.79	1.50
Mycobacterium phage 244	58652	1.17	0.07	0.99	1.39	0.92	5.78	0.87	1.33
Microbacterium phage Min1	4529	2.52	0.49	2.97	0.00	3.53	0.00	2.52	3.03
Methanobacterium phage psiM2	4532	2.22	0.19	2.98	0.00	2.99	0.00	1.91	2.47
Klebsiella phage KP15	4530	0.88	0.12	0.58	0.76	1.14	0.00	0.80	0.82
Halorubrum phage HF2	37446	1.18	0.20	0.64	0.00	1.44	0.00	1.27	1.44
Halomonas phage phiHAP-1	35741	1.07	0.13	0.64	0.00	1.97	0.00	1.03	1.33
Haloarcula phage SH1	53373	5.13	0.98	3.98	0.00	5.09	7.04	3.94	4.80
Gordonia phage GTE5	36690	2.76	0.30	2.86	1.50	2.98	0.00	2.40	2.80
Gordonia phage GTE2	48674	1.78	0.00	1.28	1.86	1.60	0.00	1.14	1.39
Gordonia phage GRU1	37639	3.32	0.30	2.45	0.00	3.66	0.00	2.43	2.47
Escherichia phage HK639	54865	0.80	0.11	0.00	0.00	0.47	0.00	0.87	1.32
Enterobacteria phage TLS	39896	0.00	0.00	1.86	0.00	0.00	0.00	1.45	1.31

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

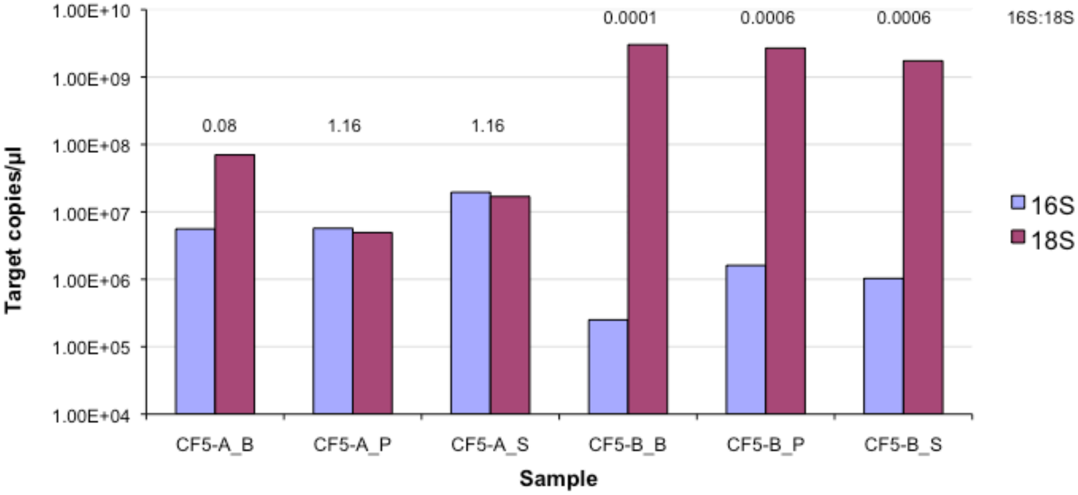
Phage	Genome Length	CF1-D	CF1-E	CF1-F	CF4-A	CF4-B	CF4-C	CF5-A	CF5-B
Enterobacteria phage TLS	39896	0.00	0.00	1.86	0.00	0.00	0.00	1.45	1.31
Enterobacteria phage RB43	37635	0.09	0.00	0.51	0.00	0.28	0.00	0.58	0.17
Enterobacteria phage P1	19870	0.18	0.06	3.52	1.18	0.26	0.00	1.01	0.21
Enterobacteria phage K1-5	47382	1.68	0.00	0.00	0.00	0.52	0.00	0.95	1.00
Enterobacteria phage Fels-2	57455	0.00	0.00	5.25	0.00	0.65	0.00	0.49	0.53
Deftia phage phiW-14	36748	1.46	0.10	2.03	0.82	1.75	0.00	1.09	1.89
Cyanophage PSS2	62714	0.36	0.00	0.27	1.08	0.61	0.00	0.71	0.63
Clostridium phage phiCTP1	52414	0.00	2.02	0.45	0.00	0.00	0.00	0.55	0.32
Clavibacter phage CMP1	42415	1.03	0.25	1.92	0.00	2.09	0.00	0.55	1.65
Burkholderia phage phiE255	47399	2.30	0.00	1.46	0.00	1.93	0.00	1.90	1.99
Burkholderia phage phiE202	48247	2.26	0.25	1.69	0.00	1.57	0.00	2.08	1.95
Burkholderia phage phiE125	48024	1.45	0.28	1.30	0.00	1.05	0.00	0.47	1.75
Burkholderia phage phiE12-2	63879	1.54	0.14	1.79	0.00	1.98	0.00	1.56	1.37
Burkholderia phage phi52237	44856	1.41	0.24	1.52	0.00	1.60	0.00	1.51	1.64
Burkholderia phage phi1026b	48177	0.53	0.10	1.49	0.00	0.43	0.00	1.32	1.15
Burkholderia phage KS9	72415	2.26	0.13	2.16	0.00	1.28	0.00	1.25	0.46
Burkholderia phage KS5	127065	1.28	0.00	1.97	0.00	1.13	0.00	1.31	1.76
Burkholderia phage KS14	42493	2.01	0.30	2.51	0.00	2.18	0.00	2.63	2.22
Burkholderia phage KS10	42663	0.98	0.00	1.11	0.00	2.10	0.00	1.67	1.36
Burkholderia phage KL3	42638	2.47	0.00	2.25	0.00	2.73	0.00	1.75	2.70
Burkholderia phage BcepNY3	12477	1.49	0.00	1.18	0.00	1.36	0.00	1.44	1.97
Burkholderia phage BcepNazgul	1181	1.65	0.09	1.33	1.63	1.41	0.00	1.17	1.67
Burkholderia phage BcepMu	7280	1.79	0.14	2.52	0.00	1.62	0.00	1.52	1.14
Burkholderia phage BcepIL02	5238	2.46	0.16	2.32	1.54	2.94	0.00	2.19	2.52
Burkholderia phage BcepGomr	39867	1.27	0.10	1.17	0.00	0.63	0.00	0.73	1.94

Supplementary Table 3.10: Normalized abundance value of phages that has >4 hits in at least one sample. (continued)

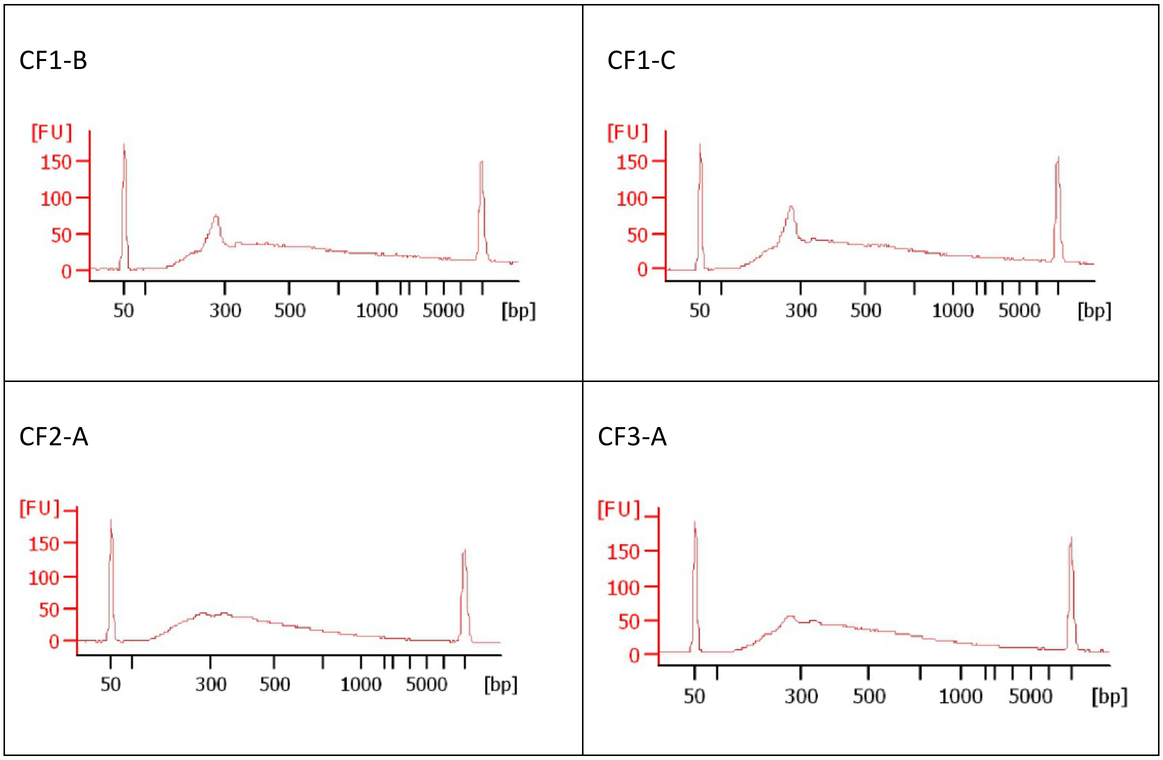
Phage	**Genome Length**	**CF1-D**	**CF1-E**	**CF1-F**	**CF4-A**	**CF4-B**	**CF4-C**	**CF5-A**	**CF5-B**
Burkholderia phage BcepC6B	4281	2.07	0.00	1.83	0.00	1.70	0.00	2.63	2.17
Burkholderia phage Bcep781	37456	2.06	0.21	1.67	0.00	1.93	0.00	2.01	1.15
Burkholderia phage Bcep43	134416	1.19	0.21	1.19	0.00	1.39	0.00	1.33	1.55
Burkholderia phage Bcep22	39077	2.63	0.50	1.90	1.53	3.03	0.00	1.92	2.76
Burkholderia phage Bcep176	218948	2.35	0.00	1.78	0.00	1.99	0.00	1.72	2.45
Burkholderia phage Bcep1	62337	1.54	0.11	1.37	0.00	1.50	0.00	1.32	1.48
Burkholderia ambifaria phage BcepF1	42271	0.89	0.27	1.04	1.41	1.04	0.00	0.73	0.90
Bordetella phage BPP-1	33985	2.66	0.34	2.00	0.00	3.25	0.00	1.54	1.59
Bordetella phage BMP-1	225268	2.66	0.34	1.99	0.00	3.25	0.00	1.53	1.49
Bordetella phage BIP-1	233234	2.66	0.34	1.99	0.00	3.25	0.00	1.81	1.58
Azospirillum phage Cd	173591	1.97	0.16	1.63	0.00	2.24	0.00	1.91	1.81
Archaeal BJ1 virus	172963	4.82	0.62	4.40	0.00	4.81	0.00	2.90	3.92
Aeromonas phage phiO18P	161475	1.29	0.00	2.76	0.00	0.64	0.00	1.39	1.46
Aeromonas phage Aeh1	58638	0.56	0.09	0.41	0.00	0.62	0.00	0.42	0.21
Aeromonas phage 65	39577	1.79	0.00	5.57	0.00	0.22	0.00	6.08	4.69
Actinoplanes phage phiAsp2	40449	4.88	0.49	4.04	0.00	4.83	0.00	3.87	4.04

Supplementary Table 3.11: Assembly characteristics of CF viromes. Each assembly contains all sequence reads from one patient, and the "contigs" file include singleton in this case. The contigs were subjected to BLASTx against NR for annotation. Sequences with both bacterial and viral annotations were examined manually and counted.

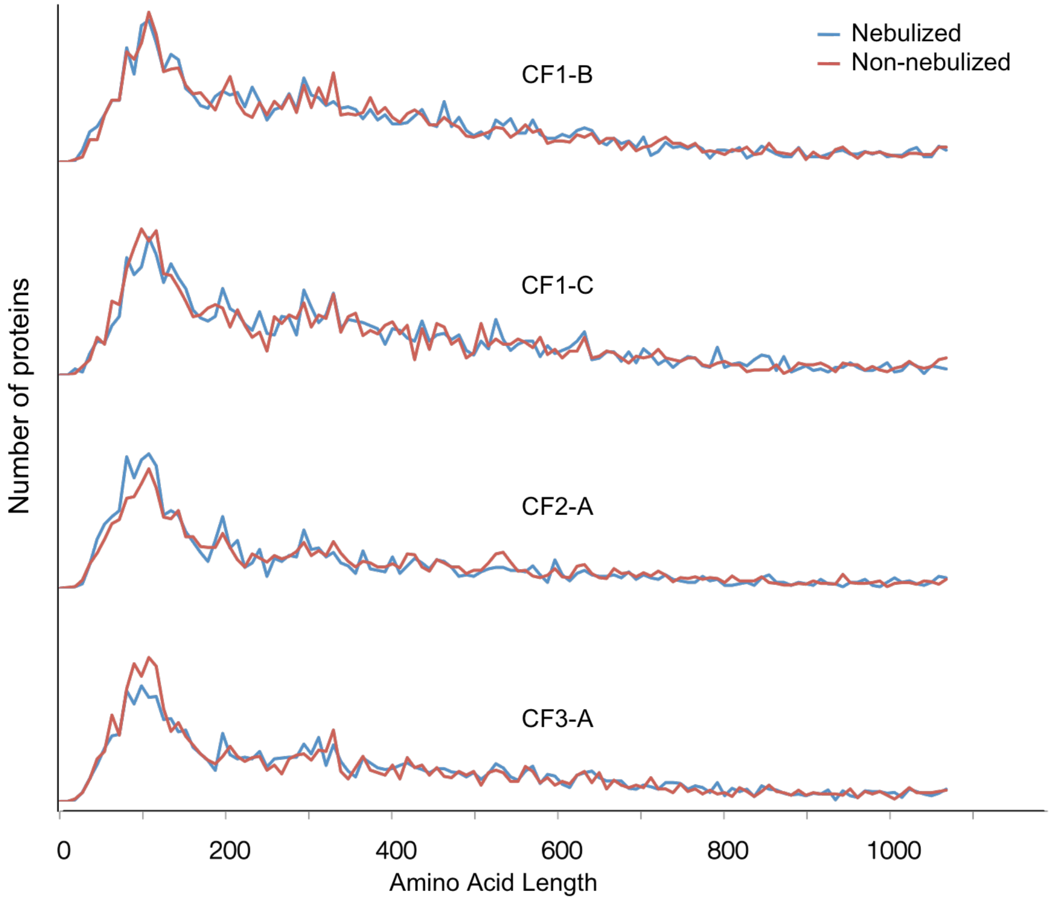
Sample	Total No. of Contigs	N50 of contigs	Largest contig size	Total No. of Seq with Bact/Phage annotations (%)
CF1	7,610	837	54,900	71 (0.9%)
CF4	3,068	789	6,416	0
CF5	10,869	698	31,895	3 (0.03%)



Supplementary Figure 3.1: The relative proportion of 16S and 18S rRNA gene based on quantitative PCR following the modified Breitenstein et al. method used to deplete host-associated (human) DNA in sample CF5-A and CF5-B. (B) Total DNA, (P) Pellet following hypotonic lysis and DNase treatment, (S) supernatant of the resuspended pellet.The ratio of 16S:18S increased following the hypotonic lysis and DNase treatment, indicated depletion of the 18S (human DNA) and the enrichment of 16S (microbial DNA).

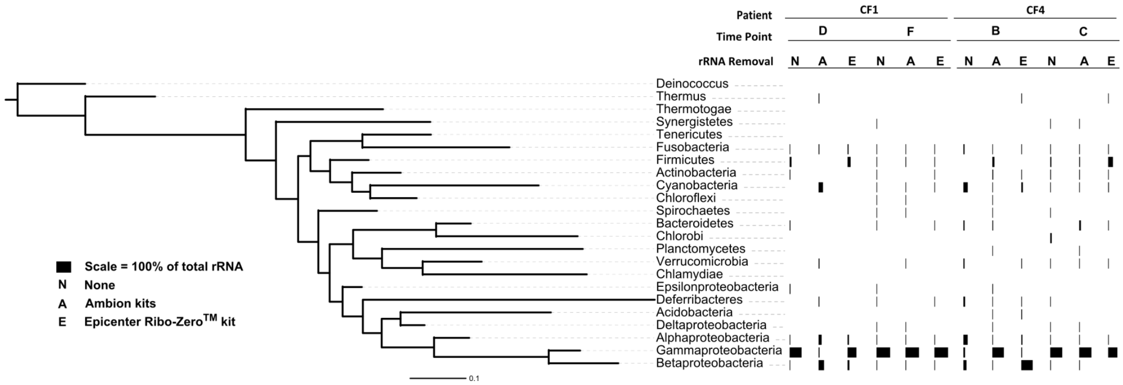


Supplementary Figure 3.2: cDNA electropherograms of metatranscriptomes generated to assess the effects of nebulization. The cDNA length profiles shown here were generated by the Agilent 2100 Bioanalyzer before nebulization.

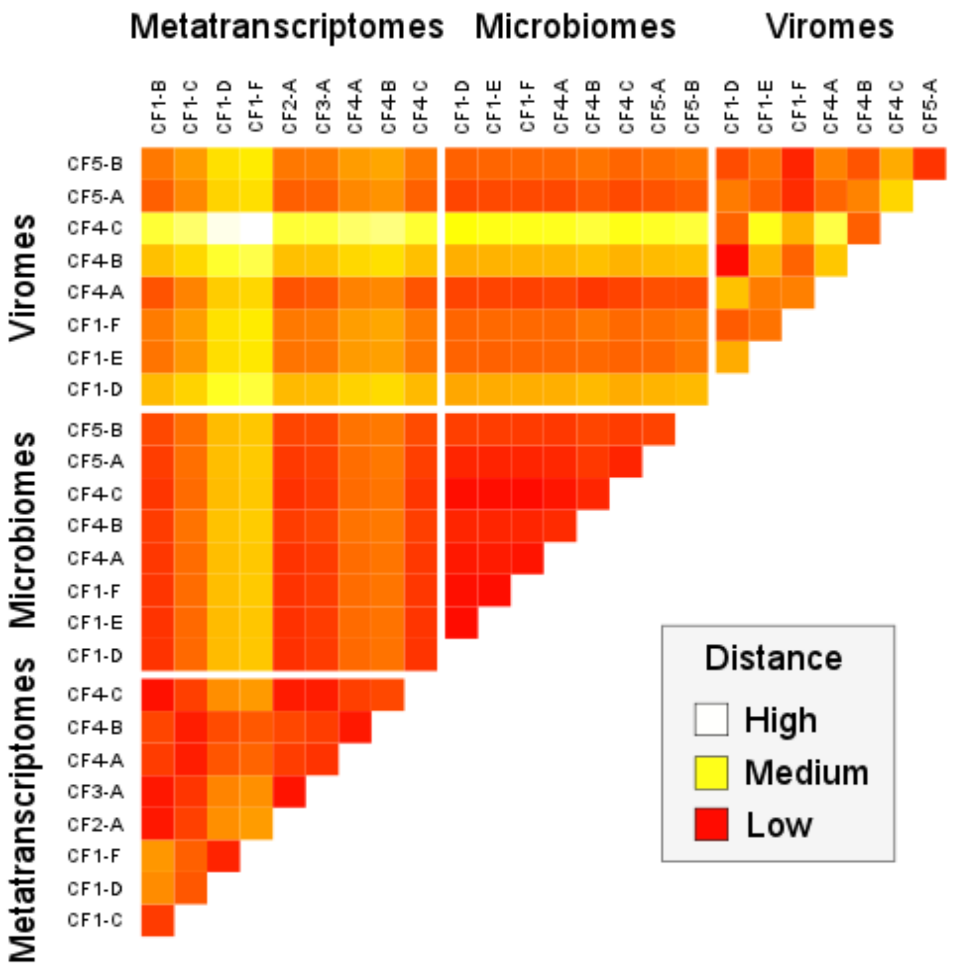


Protein Length (amino acids)

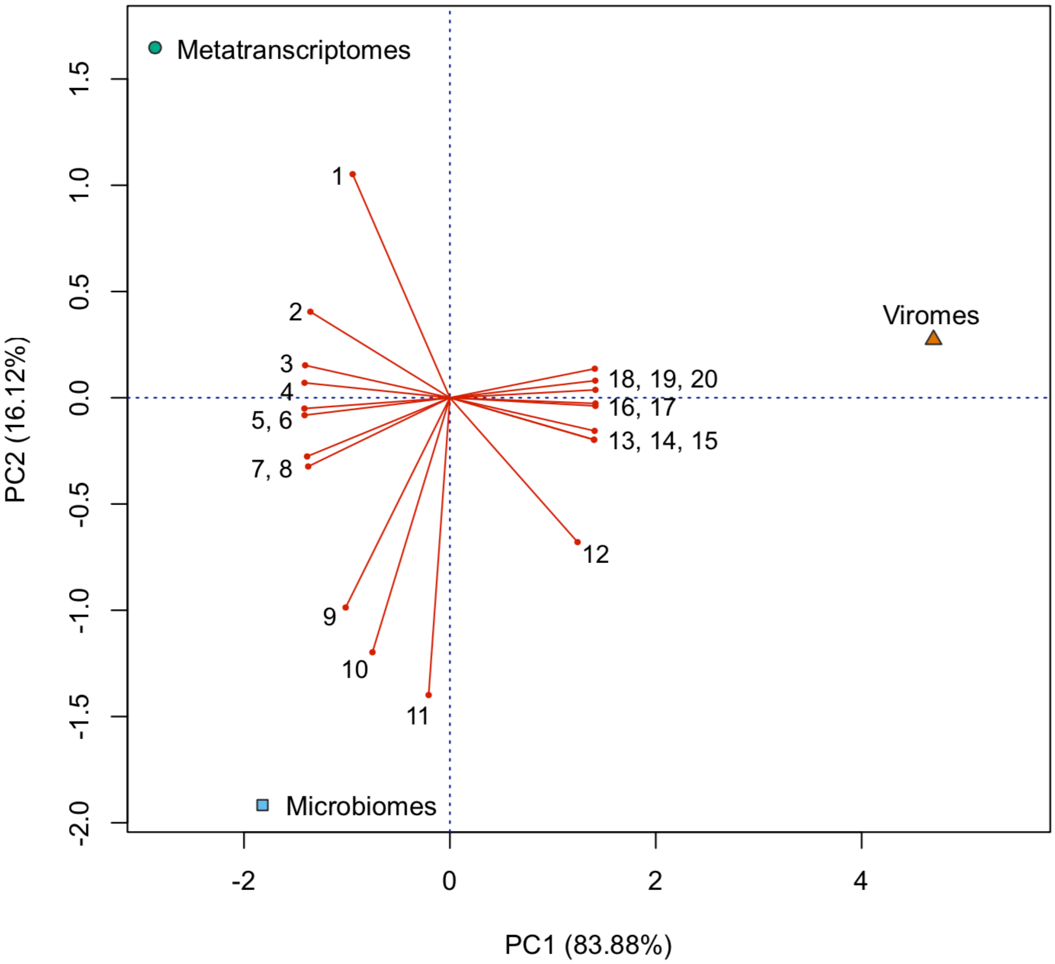
Supplementary Figure 3.3: Effect of nebulization on translated protein length. Protein lengths were predicted for cDNA transcripts by BLASTx against the NCBI non-redundant protein database. Protein length profiles for non-Chordata proteins were plotted with a cutoff at 1,200 amino acids.



Supplementary Figure 3.4: Relative abundance of microbial phyla/classes identified from rRNA reads.



Supplementary Figure 3.5: Similarities between the predicted metabolic profiles of the virome, microbiome, and metatranscriptome datasets. The heat map represents the Euclidian distance between datasets ranging from white (least similar) to red (most similar).



Supplementary Figure 3.6: Principle component analysis (PCA) of the top 20 metabolic pathways that displayed the greatest variance.

CHAPTER 4

Clinical Insights from Metagenomic Analysis of Cystic Fibrosis Sputum

Abstract

As DNA sequencing becomes faster and cheaper, genomics based approaches are being explored for personalized diagnoses and treatments. Here we provide a proof-of-principle for disease monitoring using personal metagenomic sequencing and traditional clinical microbiology, focusing on three adults with Cystic Fibrosis (CF). The CF lung is a dynamic environment that hosts a complex ecosystem comprised of bacteria, viruses, and fungi that can vary in space and time. Not surprisingly, the microbiome data from the induced sputum samples collected revealed a significant amount of species diversity not seen in routine clinical laboratory cultures. The relative abundances of several species changed as clinical treatment was altered, enabling identification of the climax and attack communities proposed in earlier work. All patient communities encoded a diversity of mechanisms to resist antibiotics, consistent with the multidrug resistant microbial communities commonly observed in CF patients. The metabolic potentials of these communities differed between the health status and recovery route of each patient. This pilot study thus provides an example of how metagenomic data might be used with clinical assessments for the development of treatments tailored to individual patients**.**

Introduction

A decade of advancements in sequencing technology and bioinformatics applications is shuttling in a new era of personalized medicine. Pathogens can be rapidly characterized during outbreaks (Köser et al. 2012; Underwood et al. 2013) and cancer patients can receive personalized diagnoses and treatments (Tran et al. 2013; Ross et al. 2013). Despite these significant advances, the technologies are yet to be used in routine clinical microbiology practice. Treating polymicrobial infections will also require a personalized approach, because the taxonomic identities and functional characteristics of microbial communities are often patient specific (Lim et al. 2012). Here we move towards this goal by using metagenomics to monitor complex pulmonary infections in patients with Cystic Fibrosis (CF).

Cystic fibrosis (CF) is a genetic disease affecting 70,000 individuals worldwide (Cystic Fibrosis Foundation: www.cff.org), and results from mutations in the gene that encodes the cystic fibrosis transmembrane conductance regulator (CFTR) (Kerem et al. 1989). These mutations result in altered trans-epithelial ion transport, leading to a dysfunctional mucus layer overlying epithelial cells in the respiratory and gastrointestinal tracts (Quinton 2010). In the lungs, the mucociliary clearance mechanism is impaired, resulting in chronic airway polymicrobial infections. The associated acute inflammatory and adaptive immune responses lead to a breakdown in the integrity of the airway wall, progressive gas exchange abnormalities and respiratory failure in many patients. Early in life, culture-based assessments indicate patients are usually infected with *Staphylococcus aureus*, *Haemophilus influenzae* and *Pseudomonas aeruginosa*. In more advanced stages of disease *P. aeruginosa* dominates, along with *Staphylococcus*, *Stenotrophomonas* and *Achromobacter* spp. Current treatments focus on controlling inflammation, the frequent use of broad-spectrum antibiotics, and physically clearing the airway biofilm by augmenting airway clearance.

Our working model describes two functional microbial communities in CF patients: *Climax* and *Attack* communities (Conrad et al. 2013). The *Climax* communities are typically bacterial and fungal populations that are stable over time and are inherently resistant to antibiotic therapy. They elicit prolonged innate and adaptive immune responses and are niche adapted. The *Attack* communities are predominantly newly acquired bacterial and viral populations that elicit strong innate immune responses and frequently trigger acute pulmonary exacerbations and are thus targets of therapy. *Attack* communities dominate earlier in CF airway disease when airway remodeling and damage is minimal (Conrad et al. 2013). In advanced stages of disease, the predominant *Climax* populations are thought to persist while the *Attack* communities fluctuate with exacerbation and treatment events. The *Climax* and *Attack* communities need not differ taxonomically, because it is the functional capabilities of the community that determine how it affects patient health. We hypothesize that community metabolic functions respond to perturbations, and that these responses can be used to identify the *Attack* and *Climax* communities within each patient.

Metagenomics sequences total community DNA from a particular source (e.g., sputum) to identify the taxonomic makeup and functional capabilities of the resident populations. It differs from community analyses that sequence only the 16S rRNA gene because all genes are sequenced, not just those used to identify taxonomy. Thus, information on community function is typically only obtainable with metagenomic sequencing, although some functions can be predicted based on taxonomy alone (e.g., methanogenesis). Neither approach requires culturing bacteria, and when applied to CF patients they have revealed microbial diversity and community complexity in airways to be unexpectedly high (LiPuma 2010; Lim et al. 2012). Furthermore, these sequence-based technologies have demonstrated that bacterial diversity decreases during treatment with broad-spectrum antibiotics, and as the disease advances over longer periods of time (Cox et al. 2010; Zhao et al. 2012). Although communities can differ between regions of the lung, and simultaneous samplings, even one occurrence of a gene encoding antibiotic resistance can have important implications for treatment.

Here we report a pilot study that focuses on obtaining a large amount of sequence data from microbial communities sampled longitudinally from the lungs of a few patients. The goal was to determine the types of information that can be obtained from such sequence data, and to frame this information in the context of clinical treatments and measured antibiotic resistance. As DNA sequencing becomes more affordable, and sequence analysis more efficient, this approach can be further developed to assist clinical decision-making and formulation of personalized therapies. Here we demonstrate an early attempt to use a metagenomic approach for probing changes in community function in individual patients over time to identify candidate changes that most drastically affect the patient. We discuss these results in the context of *Climax* and *Attack* communities for a thorough understanding of CF disease ecology.

Materials and Methods

Ten sputum samples and clinical data were collected from three adult CF patients at the Adult Cystic Fibrosis Clinic at the University of California San Diego Medical Center. Collections were made in accordance with the University of California Institutional Review Board (HRPP # 081500) and the San Diego State University Institutional Review Board (SDSU IRB 2121).

Study subjects were selected based on eligibility criteria that included all of the following: (i) a diagnosis of CF, i.e. two known mutations in the CFTR gene and/or an abnormally elevated sweat chloride test, (ii) an increase in respiratory symptoms associated with CF pulmonary exacerbations (see Supplementary Notes), and (iii) a drop in FEV_1 of at least 15% or more compared to their best FEV_1 in the past 12 months. Using these criteria 15 patients were initially recruited and screened for inclusion in this study, resulting in the collection and processing of 54 samples total. However, samples from patients who dropped out during the study period were not sequenced because we preferred to focus our limited resources on patients that had more complete longitudinal sampling and clinical information. Of those patients that remained, the three patients examined here are representative of different levels of underlying lung function and responses to treatment.

Treating physicians determined the selection and duration of antibiotic therapy and the frequency of follow up examinations. Sputum samples were obtained at the following clinical time points: (Ex) onset of an *exacerbation* prior to the initiation of antibiotics therapy; (Tr) within 24 hours prior to a change in antibiotic *therapy*; (Pt) within twelve hours of *post-treatment*, *i.e.* stopping antibiotic therapy; and (St) about 4 weeks after completing antibiotics when the patient was in the *stable* state, defined as achieving maintenance of respiratory symptoms without the need to alter their outpatient therapies. The Cystic Fibrosis Questionnaire-Revised (CFQR) evaluation and the UCSD Shortness of Breath (SOB) questionnaire were used during each collection.

During sample collection, sterile saline solution (60 ml) was used as a mouth rinse to minimize contamination by oral microbes. Sputum samples were then collected over a thirty-minute time period after the inhalation of four milliliters of 7% hypertonic saline via a Pari LC plus nebulizer. Samples were processed immediately, as described previously (Lim et al. 2012).

In brief, sputum samples were homogenized with syringe and then aliquoted for the separate isolation of viral particles and microbial cells. Viral samples (i.e., the virome) were diluted and treated with Dithiothreitol (DTT) to dissociate the mucus and then passed through a 0.45μm filter to remove large particles. Viral particles were then isolated and concentrated using cesium chloride density ultracentrifugation. Viral DNA was extracted using the formamide-CTAB/phenol:chloroform method (Lim et al. 2012). Microbial samples (i.e., the microbiome) were treated with β-mercaptoethanol to dissociate the mucus. Cells were repeatedly washed with sterilized deionized water to lyse human cells and then treated with DNase to remove extracellular DNA (e.g., human and biofilm-associated (Lim et al. 2012)). Microbial DNA was extracted using the Nucleospin® Tissue kit with the gram-positive variation that includes a lysozyme treatment.

All samples were sequenced using the Roche/454 pyrosequencing with GS-FLX Titanium chemistry. All datasets were preprocessed as previously described (Lim et al. 2012). Duplicates and reads of low quality were removed using PRINSEQ (Schmieder and Edwards 2011a). DeconSeq (Schmieder and Edwards 2011b) was used to screen for and exclude human-derived sequences from the microbiome data. The taxonomies of the resultant sequences were identified using a BLASTn search against the NCBI nucleotide (NT) database. Sequences were removed if they were assigned to the phylum *Chordata*, or any synthetic/vector sequence (Supplementary Table 4.1). Viromes were further annotated using a tBLASTx search against an in-house viral database. Abundance values were normalized based on the total number of reads per metagenome. All metagenomes were additionally annotated using the KEGG database (Kanehisa and Goto 2000) and analyzed using the HUMAnN pipeline (Abubucker et al. 2012). The normalized relative abundance values were used for subsequent principal component analysis (PCA). The antibiotic resistance potential of each microbiome was identified by comparing the data with (i) the antibiotic resistance database (ARDB) (Liu and Pop 2009) that contains 23,137 antibiotic resistant-associated sequences, using BLASTx with a threshold of 40% identity over at least 60% of the query sequence, and (ii) an up-to-date manually curated macrolide and aminoglycoside resistance database from UncovAR pipeline (Schmieder 2013). The abundance of each resistance annotation was normalized by the number of reads in each metagenome and weighted by the length of each gene and the total number of base pairs in the respective database.

For additional details on database generation, content, and BLAST parameters, see the Supplementary Note. All sequence data can be retrieved from NCBI SRA under accession number SRP009392.

Results

Sample and data collection

Sputum was induced and collected from three adult CF patients (median age of 36 years) following mouthwash with sterile saline solution (Table 4.1). Baseline FEV_1 values, defined by the best FEV_1 value for each patient in the past 12 months when the first sample was collected, ranged from 1.16L (27% predicted) to 4.15L (89% predicted). The first samples were collected when each patient presented a severe CF pulmonary exacerbation. The patients received standard care for an acute pulmonary exacerbation, including bronchodilators, chest physiotherapy and systemic antibiotics. The attending physician selected antibiotic type and duration of intravenous treatment. Up to four sputum samples were collected from each patient based on the clinical course that followed the initiation of their treatment, resulting in a total of 10 samples with matched exacerbation/post-treatment sample pairs from all three patients.

Clinical data were collected from the UCSD Adult CF Clinic (Tables 4.1 & 4.2). These data included culture-based microbiology assessments 100 days before and after collection of the first sample. The patient reported outcome surveys were collected at the time of sampling (see Materials and Methods). For each patient, these clinical data were combined with the metagenome data (Figure 4.1) to study individual cases; the data from all patients were also combined together for a comparative analysis.

Table 4.1: Information on patient samples

Patient (Gender)	*CFTR* Genotype Age	Sample Name	Health Status	Time Scale^{+}	FEV$_1$*	Treatment received
CF6 (Female)	ΔF508/Q372Q	CF6-A-Ex	Onset Exacerbation	Day 0	1.91 (57%)	Colistin, Ceftazidime
	39	CF6-B-Tr	Treatment	Day 12	2.03 (60%)	Ciprofloxacin, Aztreonam
		CF6-C-Pt	Post Treatment	Day 17	2.06 (61%)	Stopped all antibiotics
		CF6-D-St	Stable	Day 46	2.07 (61%)	Not on antibiotics
CF7 (Male)	ΔF508/ΔF508	CF7-A-Ex	Onset Exacerbation	Day 0	0.87 (21%)	Tobramycin, Ceftazidime, Trimethoprim/sulfamethoxazole
	36	CF7-B-Tr	Treatment	Day 20	0.80 (19%)	Piperacillin/Tazobactam, Trimethoprim/sulfamethoxazole
		CF7 C-Tr	Treatment Change	Day 27	0.82 (19%)	Piperacillin/Tazobactam, Trimethoprim/Sulfamethoxazole
		CF7-D-Pt	Post Treatment	Day 37	0.92 (22%)	Stopped antibiotics treatment, Prednisone was added on Day 33
CF8 (Male)	ΔF508/ΔF508	CF8-A-Ex	Onset Exacerbation	Day 0	3.39 (73%)	Meropenem, Tobramycin
	26	CF8-B-Pt	Post Treatment	Day17	4.15 (89%)	Meropenem, Tobramycin

$^{+}$ Time scale starts at Day 0 during the onset of exacerbation when the first sample was collected. The subsequent numbers indicate days after the first sample was collected.
* FEV$_1$ is measured as the forced expiratory volume in one second (% predicted).

Table 4.2: Bacterial culture data from hospital clinical lab. The timeline is corresponding to the date (Day 0) when the first sputum sample was taken during an exacerbation.

Patient (Gender)	Time Scale	Culture Positive organism	Antibiotic resistance profile
CF6 (Female)	Day -57	*Escherichia coli* (ESBL)	Ampicillin, Unasyn, Cefazolin, Ceftazidime, Cefotaxime, Cefipime, Cefuroxime, Ciprofloxacin, Gentamicin, Piperacillin/Tazobactam, Tobramycin.
	Day -6	*Escherichia coli* (ESBL)	Ampicillin, Unasyn, Cefazolin, Ceftazidime, Cefotaxime, Cefipime, Cefuroxime, Ciprofloxacin, Gentamicin, Piperacillin/Tazobactam, Trimethoprim/Sulfamethoxazole
	Day +82	*Escherichia coli* (ESBL)	Ampicillin, Unasyn, Cefazolin, Ceftazidime, Cefotaxime, Cefipime, Cefuroxime, Ciprofloxacin, Gentamicin, Piperacillin/Tazobactam, Trimethoprim/Sulfamethoxazole
CF7 (Male)	Day -38	*Pseudomonas aeruginosa* (mucoid)	Amikacin, Ciprofloxacin, Meropenem, Tobramycin
		Stenotrophomonas maltophilia	Amikacin, Ciprofloxacin, Gentamicin, Tobramycin
	Day -3	*Pseudomonas aeruginosa* (mucoid)	Amikacin, Ciprofloxacin, Gentamicin, Trimethoprim/Sulfamethaxazole
		Stenotrophomonas maltophilia	Amikacin, Ciprofloxacin, Gentamicin, Bactrim, Meropenem, Piperacillin, Tobramycin, Ceftazidime
	Day +13	*Pseudomonas aeruginosa*	Ciprofloxacin, Bactrim, Meropenem
	Day +86	*Stenotrophomonas maltophilia*	Ceftazidime
		Pseudomonas aeruginosa	Amikacin, Ciprofloxacin, Gentamicin, Tobramycin, Trimethoprim/Sulfamethoxazole
CF8 (Male)	Day -36	*Streptococcus* Group C	N/A
		Pseudomonas aeruginosa (mucoid) – strain 1	Amikacin, Gentamicin, Trimethoprim/Sulfamethoxazole
		Pseudomonas aeruginosa (mucoid) – strain 2	
	Day 0	*Streptococcus* Group C	N/A
		Pseudomonas aeruginosa (mucoid)	Amikacin, Ciprofloxacin, Gentamicin, Tobramycin, Trimethoprim/Sulfamethoxazole
	Day 10	*Stenotrophomonas* maltophilia	Ceftazidime
		Pseudomonas aeruginosa (mucoid)	

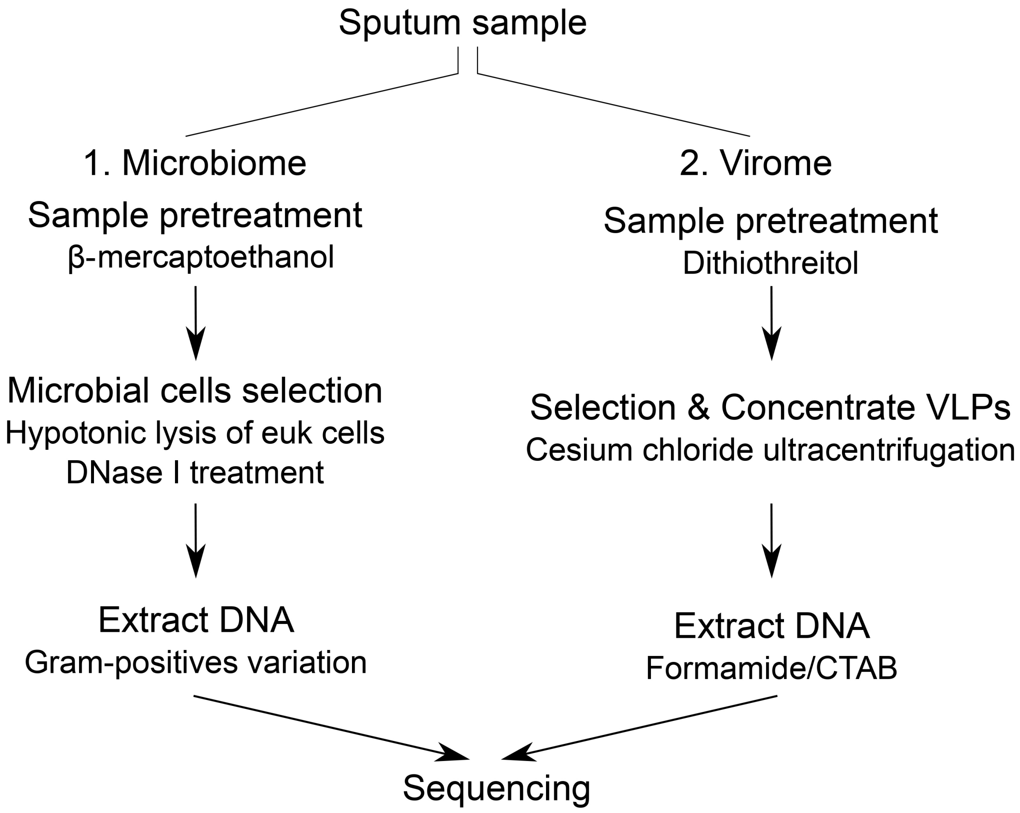


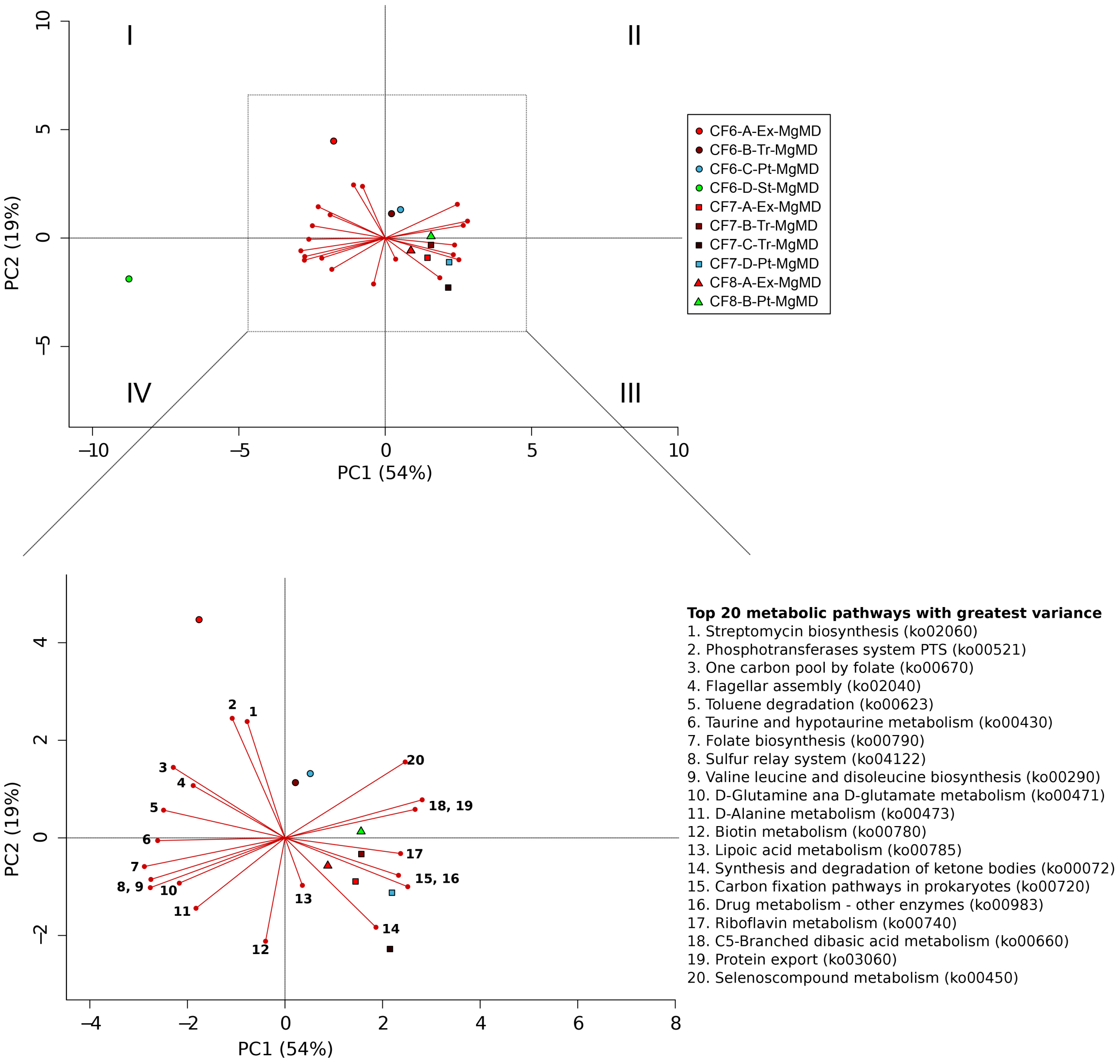
Figure 4.1: Workflow for the preparation of CF sputum samples for microbiome and virome sequencing.

Metabolism and resistance in CF microbiomes

Microbial metabolism and antibiotic resistance are compelling indicators of community function because they reveal which pathways facilitate microbial colonization and persistence in the CF lung. The metabolic pathways present in each microbiome were investigated with the goal of using this information to eventually develop tools for identifying important biomarkers of *Climax* and *Attack* communities. Because these communities are likely to be patient-specific, and therefore have important implications for personalized diagnosis and treatment, the differences in community metabolism were examined. A total of 222 metabolic pathways were identified from all patient datasets using BLASTx comparison against the KEGG database as described in the Materials and Methods. For pattern exploration purposes, principal component analysis (PCA) was used for dimension reduction. Twenty metabolic pathways with the greatest variance between patient microbiomes were ordered (see Methods) and used to identify the functions that varied the most across the microbiomes (Figure 4.2). Metabolic profiles are shown separately for each patient in Supplementary Figures 4.1-4.3.

As resistance against antibiotic treatment remains one of the main challenges in the treatment of CF pulmonary infections, the abundances of genes whose products or mutations are known to confer resistance to antibiotics were also determined (Figure 4.3A-B; Supplementary Table 4.3). One contribution of this pilot study will be to examine whether predicted antibiotic resistance profiles fluctuate through time, or whether a consistent increase (or decrease) in community resistance is observed. For clinicians to find use in metagenome data, it will be important to understand how quickly antibiotic resistance might change in the community.

Because the majority of viruses found in the CF lungs are bacteriophages (Willner et al. 2009; Willner and Furlan 2010), which are known to transfer genes between microbial hosts, the potential for exchange of antibiotic resistance genes was also evaluated using the viromes corresponding to each microbiome (Figure 4.3C; Supplementary Table 4.4). Bacterial contaminating sequences in the viromes were minimal, based on the low abundance of 16S ribosomal RNA genes in the viromes (< 0.05%).



Top 20 metabolic pathways with greatest variance
1. Streptomycin biosynthesis (ko02060)
2. Phosphotransferases system PTS (ko00521)
3. One carbon pool by folate (ko00670)
4. Flagellar assembly (ko02040)
5. Toluene degradation (ko00623)
6. Taurine and hypotaurine metabolism (ko00430)
7. Folate biosynthesis (ko00790)
8. Sulfur relay system (ko04122)
9. Valine leucine and disoleucine biosynthesis (ko00290)
10. D-Glutamine ana D-glutamate metabolism (ko00471)
11. D-Alanine metabolism (ko00473)
12. Biotin metabolism (ko00780)
13. Lipoic acid metabolism (ko00785)
14. Synthesis and degradation of ketone bodies (ko00072)
15. Carbon fixation pathways in prokaryotes (ko00720)
16. Drug metabolism - other enzymes (ko00983)
17. Riboflavin metabolism (ko00740)
18. C5-Branched dibasic acid metabolism (ko00660)
19. Protein export (ko03060)
20. Selenoscompound metabolism (ko00450)

Figure 4.2: A comparison of metabolic pathways between microbiomes using principal component analysis (PCA) of the twenty metabolic pathways that varied the most between microbiomes. The bottom panel presents a close-up view of the upper panel's squared region. Patient CF6 is represented by circles, patient CF7 is represented by squares, and patient CF8 is represented by diamonds. The colors inside the shapes represent the health status of each patient as shown in the figure legend.

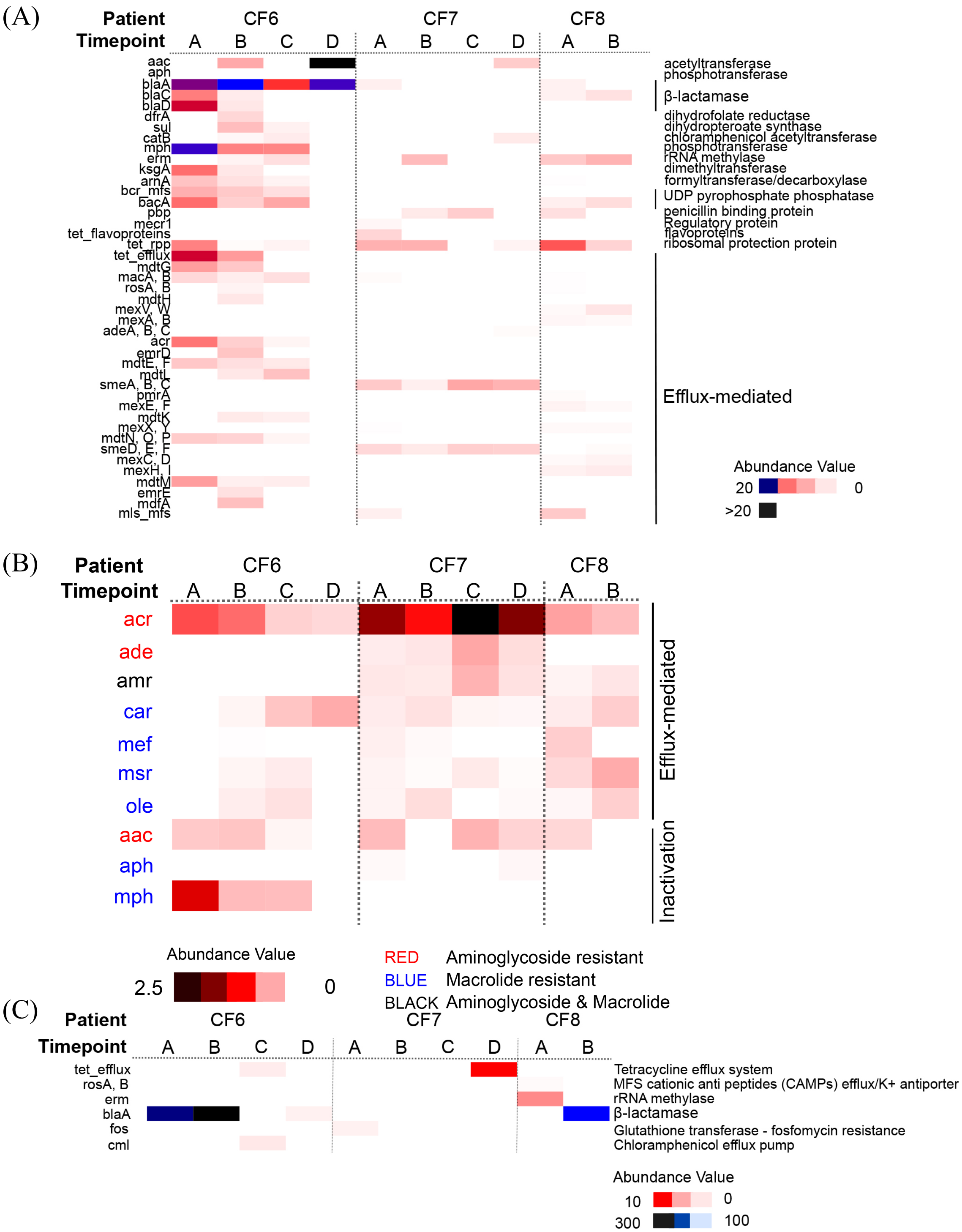


Figure 4.3: Abundances of antibiotic resistance genes based on (A) Antibiotic Resistance Database (ARDB) and (B) the program UncovAR that predicts resistance to aminoglycoside and macrolide antibiotics. (C) The antibiotic resistance gene profiles of the viromes based on BLASTx comparison against the ARDB. All abundances were normalized by metagenome and gene size, and weighted by database size.

Patient CF6

The clinical and sample information are presented in Table 4.1 and Figure 4.4.

The patient presented with dyspnea, increased cough and sputum production, and a 19% drop in her FEV_1 during her outpatient visit. The baseline FEV_1 was recorded as 2.36L (69% predicted). She initially started intravenous therapy that included eleven days of colistin and ceftazidime, but due to a suboptimal response, her therapy was then changed to five days of ciprofloxacin and aztreonam.

The first sample (CF6-A-Ex) was collected before the administration of intravenous antibiotics on Day 0. The second sample (CF6-B-Tr) was collected on Day 12 before the change in antibiotics (Table 4.1) .The third sample (CF6-C-Pt) was taken 5 days later (Day 17) when the patient's presenting symptoms resolved, and had an improved score on the UCSD Shortness of Breath questionnaire. At this point her FEV_1 improved to 2.06L (61% predicted), approximately 8% lower than her baseline FEV_1 value. A month following the end of therapy, the patient was clinically stable when the fourth sample (CF6-D-St) was collected. At this time, the patient-reported outcome measures reflected in the UCSD SOB and CFQR cumulative scores showed significant improvement even though the FEV_1 was essentially unchanged and remained below the baseline at 2.07L (61% predicted) (Supplementary Table 4.2). Importantly, there was a significant improvement in the respiratory domain of the CFQR evaluation (Supplementary Table 4.2). The patient was clinically defined as an "intermediate responder" based on her recovery and responses to therapy.

Clinical culturing revealed growth of the fungi *Candida albicans* and *Scedosporium apiospermum*, and extended spectrum β-lactamase (ESBL) *Escherichia* *coli* that exhibited sensitivities to various antibiotics during the study period (Table 4.2). Metagenomic analysis of the microbial communities (microbiomes) were consistent with the clinical culturing, showing a high abundance of *E coli* and a low abundance of *P. aeruginosa* (Figure 4.4). However, the microbiomes also included species that are not typically incorporated into clinical culturing protocols, such as *Streptococcus* spp. and *Rothia mucilaginosa*, largely because they are considered oral contaminants or benign respiratory microbes (Figure 4.4).

The relative abundance of *E. coli* decreased during the course of antibiotic therapy and then increased slightly when the patient's health was stable. In contrast, the relative abundances of *Streptococcus* spp. and *R. mucilaginosa* increased during the course of therapy. One month following the completion of intravenous antibiotic therapy, *R. mucilaginosa* and *E. coli* dominated the microbiome (Figure 4.4, time point D). *E. coli*, *Streptococcus* spp., and *R. mucilaginosa* were the numerically dominant bacteria in this patient, with *P. aeruginosa* detectable but at lower abundance at all time points (<0.1% in A, B, and C; 1.25% in D).

Historically and during the course of sampling, this patient was extensively exposed to a diversity of antibiotics, including: aminoglycosides, β-lactams, fluoroquinolones, macrolides and various polypeptides (Table 4.1). Hence, the microbial community was expected to have evolved resistance to these major groups of antibiotics. This was confirmed by clinical lab tests (Table 4.2; Figure 4.4), and predicted by the metagenomics analysis (Figure 4.3A; Supplementary Table 4.3).

Metagenome data suggested the microbes in this community encode a plethora of antibiotic resistance mechanisms: multidrug resistance efflux pumps, β-lactamases, and various enzymes that confer resistance (Supplementary Table 4.3). The abundances of two known resistance genes, *arnA* and *bla*, increased following treatment with colistin and ceftazidime, respectively. Two antifolate genes were present: *dfrA* encodes Group A drug-insensitive dihydrofolate reductase for trimethoprim resistance, and *sul* encodes sulfonamide-resistant dihydropteroate synthase for sulfonamide (sulfamethoxazole) resistance. Although the patient was not treated with any macrolides within the year prior to this study, various macrolide resistance mechanisms were found throughout the microbiomes (Figure 4.3A; Supplementary Table 4.3).

The PCA based on the potential microbial metabolic functions show that CF6 differs from the other patients considerably, especially during the exacerbation and stable states (Figure 4.2). The separation of the exacerbation sample (A) was largely driven by the presence of genes encoding streptomycin biosynthesis (ko00521) and phosphotransferases (ko02060), whereas separation of the stable sample (D) was driven by the presence of genes that encode taurine and hypotaurine metabolism (ko00430), folate biosynthesis (ko00790), a sulfur relay system (ko04122), D-glutamine and D-glutamate metabolism (ko00471), and valine, leucine and isoleucine biosynthesis (ko00290). The first and second principal components for treatment and post-treatment samples were very similar for all patients, indicating these communities had similar metabolic potentials at both time points.

Comparing different time points within patient CF6 (Supplementary Figure 4.1), drug metabolism was one of the main metabolic pathways that drove separation of the treatment (B) and post-treatment (C) samples. Separation of the stable sample was driven by several core metabolic pathways, which may indicate a “stable” *Climax* community in the patient’s lungs.

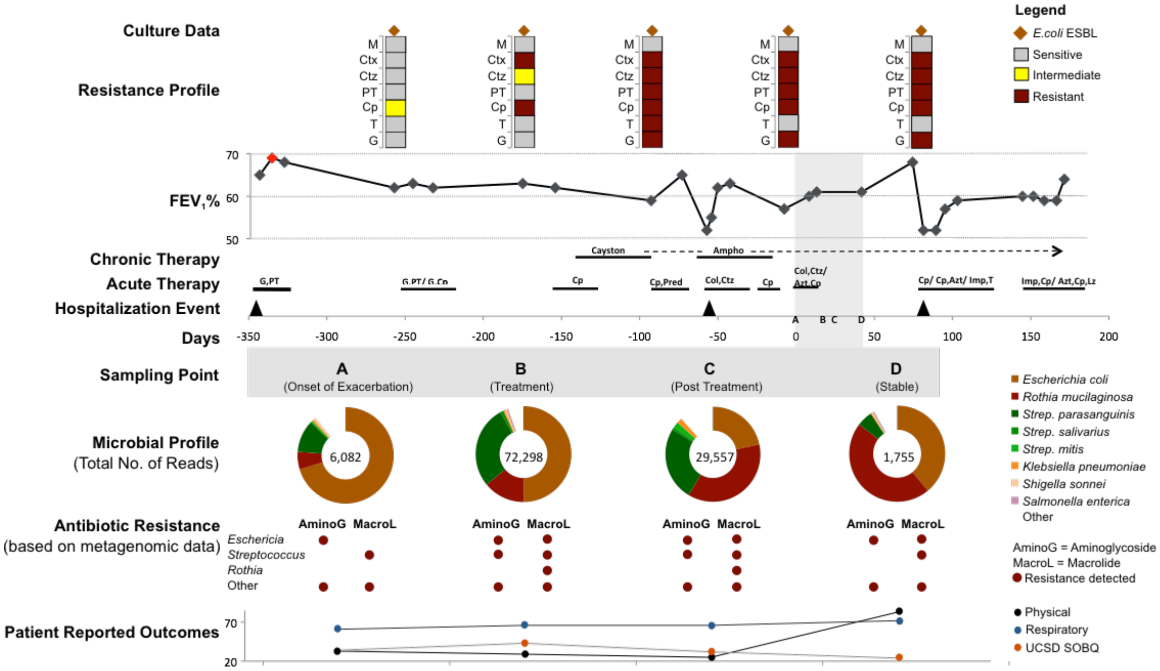


Figure 4.4: An overview of clinical and metagenomic data for patient CF6. This patient was clinically defined as an "intermediate responder". The FEV$_1$% is illustrated across a 550-day period and a red diamond indicates the baseline FEV$_1$%. The 'Resistance Profile' panel near the top shows the results from laboratory resistance tests on clinically-cultured microbes, obtained at the time point immediately below their placement. The first sample was collected during the onset of a clinically defined exacerbation at time point 0. The medications prescribed during each chronic and acute therapy are shown in the time line. The length of the black line corresponds to the duration of therapy. Acute therapy is represented by: M=Meropenem, Ctx=Ceftriaxone, Ctz=Ceftazidime, PT=Piperacillin/Tazobactam, Cp=Ciprofloxacin, T=Tobramycin, G=Gentamicin, Pred=Prednisone, Imp=Imipenem, Azt=Aztreonam, Lz=Linezolid, Col=Colistin, Ampho = Inhaled Amphotericin, Cayston = Inhaled Aztreonam

Patient CF7

The clinical and sample information are presented in Table 4.1 and Figure 4.5.

The baseline FEV_1 was recorded as 1.16L (27% predicted). This patient presented with increased dyspnea and sputum production as well as a 25% drop in his FEV_1, which is recorded as 0.87L (21% predicted), prompting treatment with different combinations of intravenous antibiotics including tobramycin, ceftazidime, piperacillin/tazobactam, and trimethoprim/sulfmethoxazole. The first sample (CF7-A-Ex) was collected before the administration of therapy. Sample CF7-B-Tr was collected 20 days after the treatment with tobramycin and ceftazidime, and sample CF7-C-Tr was collected 7 days after stopping ceftazidime and prior to the initiation of piperacillin/tazobactam and trimethoprim/sulfmethoxazole due to the lack of clinical and physiological response. The fourth sample (CF7-Pt) was taken 10 days later upon completion of treatment. However, the patient had not completely improved physiologically, showing no resolution of the initial respiratory symptoms and a worsening of shortness of breath (Supplementary Table 2). The FEV_1 improved to 0.92L (22% predicted), but remained more than 20% lower than the initial baseline FEV_1 value. The respiratory domain on the CFQR improved by a value greater than the MCID of 5 (Supplementary Table 4.2), suggesting that the patient might have improved slightly. The patient underwent a lung transplant about 3 months following the last sample collected.

Clinical culturing revealed growth of the fungi *C. albicans*, *Aspergillus fumigatus*. *S. maltophilia* and mucoid *P. aeruginosa* were also cultured, and had varying patterns of antibiotic susceptibility (Table 4.2; Figure 4.5). Metagenomic analysis at all time points showed an overall high relative abundance of *S. maltophilia* that ranged from 41%-90%, whereas *P. aeruginosa* was rare (<1%). *R. mucilaginosa*, *Rothia dentocariosa, Streptococcus* spp. and *Prevotella melaninogenica* were highly abundant during the onset of exacerbation and following the first unresponsive therapy. However, one month following the onset of exacerbation during which the patient was intensely treated with a combination of antibiotics, *S. maltophilia* repopulated the lung community and the overall bacterial diversity decreased. *S. maltophilia* is the key player within the microbial community in this patient and clinical testing indicated it was highly resistant to all major groups of antibiotics, including aminoglycosides, macrolides, β-lactamases, and fluoroquinolones (Figure 4.3A; Figure 4.5; Supplementary Table 4.3). The metagenome data suggested that the mechanisms of antibiotic resistance in this community were multidrug resistance efflux pumps, a protein that prevents tetracycline from inhibiting the ribosome, and various enzymes and transporters that confer resistance (Figure 4.3). As seen in CF6, even though macrolides are not reported in CF7's recent medical history, various macrolide-specific resistance mechanisms were found (Figure 4.3A; Supplementary Table 4.3). Further comparison of the data with aminoglycoside and macrolide resistance genes in UncovAR revealed that *S. maltophilia* likely relies on efflux pumps (e.g., *acr*) to purge antibiotics from the cell (Figure 4.3B).

PCA of metabolic pathways showed that the metabolic profiles of the microbial communities in CF7 were similar through time (Figure 4.2; Supplementary Figure 4.3). This limited change in the patient's microbial taxonomical and functional profiles was consistent with the patient's unresponsive clinical status. The data also suggests that a particular set of metabolic functions (Figure 4.3) may have been responsible for the persistence of his unresponsive *Climax* community. These metabolic functions include synthesis and degradation of ketone bodies (ko00072), carbon fixation pathways in prokaryotes (ko00720), drug metabolism pathways (ko00983), and riboflavin metabolism (ko00740). Further examination within the drug metabolism pathway revealed the presence of the arylamine N-acetyltransferase (NAT) gene involved in isoniazid metabolism, known to occur in *E. coli* (Chang and Chung 1998; Schomburg and Chang 2006). Isoniazid is commonly used to treat tuberculosis, but not within this patient according to his medical history. It is not known whether NAT is capable of metabolizing any drugs prescribed to patient CF7.

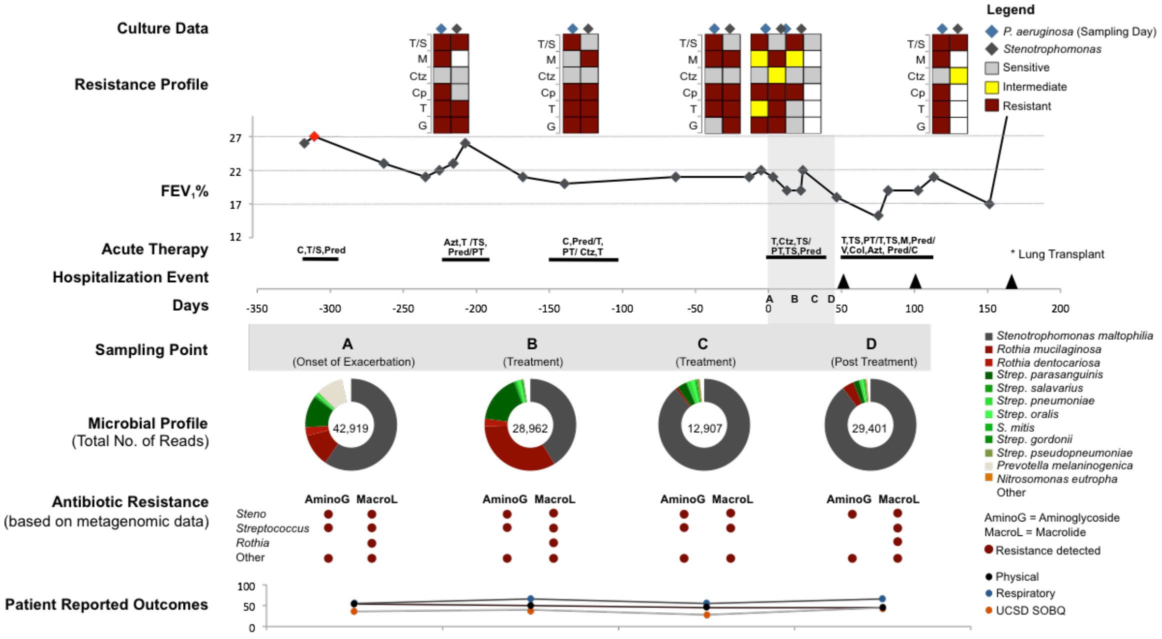


Figure 4.5: An overview of clinical and metagenomic data for patient CF7. This patient was clinically defined as a “non-responder”. The FEV1% is illustrated across a 520-day period and a red diamond indicates the baseline FEV1%; the spike at the end indicates the effect of lung transplantation. The ‘Resistance Profile’ panel near the top shows the results from laboratory resistance tests on clinically-cultured microbes, obtained at the time point immediately below their placement. The first sample was collected during the onset of a clinically defined exacerbation at time point 0. The medications prescribed during each chronic and acute therapy are shown in the time line. The length of the black line corresponds to the duration of therapy. Acute therapy is represented by: M=Meropenem, Ctx=Ceftriaxone, Ctz=Ceftazidime, PT=Piperacillin/Tazobactam, Cp=Ciprofloxacin, T=Tobramycin, G=Gentamicin, Pred=Prednisone, Imp=Imipenem, Azt=Aztreonam, Lz=Linezolid, Col=Colistin, Ampho = Inhaled Amphotericin, Cayston = Inhaled Aztreonam

Patient CF8

The clinical and sample information are presented in Table 4.1 and Figure 4.6.

The baseline FEV_1 was as 4.15L (89% predicted). The patient was admitted to the hospital for increased cough, dyspnea and sputum production with an 18% drop in his FEV_1 (Table 4.1). The patient was started on a combination of tobramycin and meropenem for a total of 16 days (Table 4.1). The first sample (CF8-A-Ex) was collected before the administration of therapy and the second sample (CF8-B-Pt) was collected 17 days later when the patient completed the treatment. At the end of therapy, the patient reported resolution of his initial respiratory symptoms, which was confirmed in his reported outcomes assessed by both the CFQR respiratory, and shortness of breath scores (Figure 4.6; Supplementary Table 4.2). The FEV_1 improved to baseline 4.15L (89% predicted). Based on the patient's recovery and responses to therapy, patient CF8 was considered a "responder".

Clinical culturing revealed growth of mucoid *P. aeruginosa*, *Streptococcus* group C, and *S. maltophilia* during the period when the samples were collected (Table 4.2). *P. aeruginosa* had varying antibiotic susceptibility patterns (Figure 4.6; Table 4.2). *Streptococcus* spp. was considered an oral contaminant in the clinical lab, and therefore its antibiotic susceptibilities were not tested. Metagenomic analysis showed a high diversity of bacteria in CF8's microbiomes, particularly in sample A (Figure 4.6). The most abundant bacteria were *P. aeruginosa, Streptococcus* spp., *Rothia* spp., and the anaerobes *Prevotella melaninogenica, Veillonella parvula*, and *Fusobacterium nucleatum.* In sample B that followed antibiotic treatment, most of the *Streptococcus* spp.

and anaerobes were eliminated and the microbiome was dominated by *P. aeruginosa*, *R. mucilaginosa*, and *Lactobacillus* spp.

The patient was treated with an aminoglycoside (tobramycin) and a β-lactam (meropenem) during the course of sampling (Table 4.1). Treatment with either a macrolide or fluoroquinolone was not reported during the 600-day medical history. However, the metagenomes data predicted resistance to several groups of antibiotics (Figure 4.3A; Figure 4.6; Supplementary Table 4.3) including aminoglycosides, β-lactams, fluoroquinolones and macrolides. Antibiotic resistance mechanisms detected in the microbiome data included β-lactamases, multidrug efflux pumps, the same ribosomal protection protein identified in CF7, and various enzymes that confer resistance (Supplementary Table 4.3). A high abundance of genes that confer resistance to tetracycline and macrolides were detected in the exacerbation sample (CF8-A-Ex) but their abundances decreased upon treatment with meropenem and tobramycin. However, the abundances of β-lactamase genes (conferring resistance to β-lactams), and those encoding multidrug resistance efflux pumps, increased post-treatment (Figure 4.3C).

PCA indicated that the metabolic pathways present in the communities found in both samples were quite similar. Interestingly the exacerbation sample was closer to the disease state sample from CF7, whereas the post-treatment sample was similar to the CF6 post-treatment sample, an intermediate responder. This “post-treatment” community was characterized by C5-branched dibasic acid metabolism, protein export, and selenocompound metabolism (Figure 4.2). The C5-branched dibasic acid metabolism belongs to the “carbohydrate metabolism” superclass, and is known to provide an alternative source of carbon and energy. Comparison of the top 20 representative metabolic pathways that differed between the two CF8 samples (Supplementary Figure 4.3) indicated that folate biosynthesis, glycan degradation, and nathphalene and dioxin degradation were responsible for distinguishing the metagenomes from these samples.

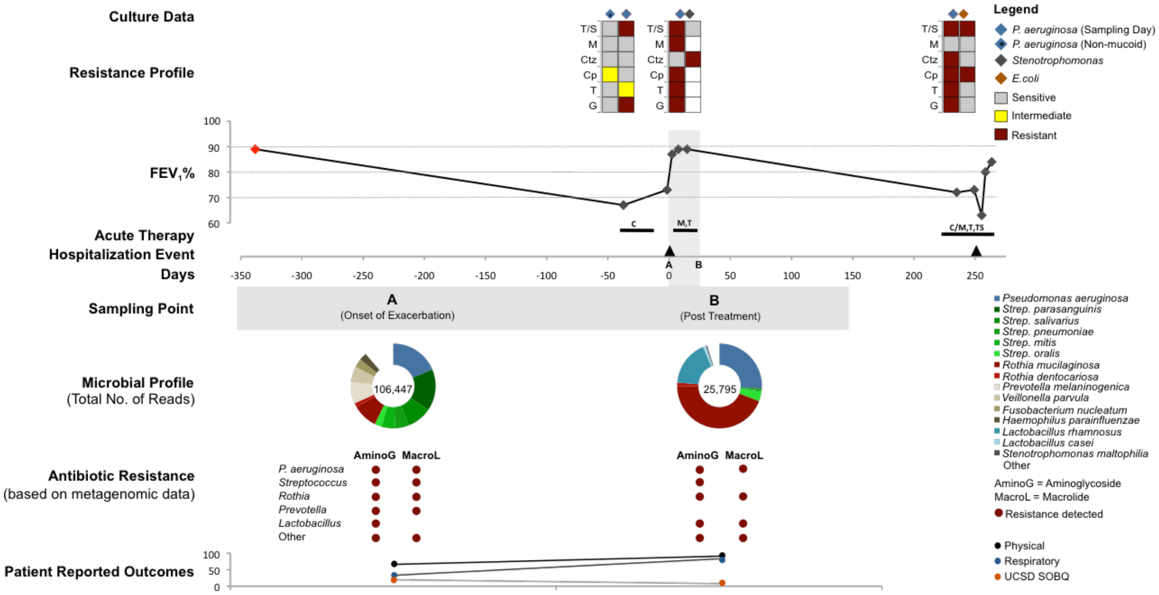


Figure 4.6: An overview of clinical and metagenomic data in patient CF8. This patient was clinically defined as a "responder". The FEV$_1$% was illustrated across a 600-day period and a red diamond indicates the baseline FEV$_1$%. The 'Resistance Profile' panel near the top shows the results from laboratory resistance tests on clinically-cultured microbes, obtained at the time point immediately below their placement. The first sample was collected during the onset of a clinically defined exacerbation at time point 0. The medications prescribed during each chronic and acute therapy are shown in the time line. The length of the black line corresponds to the duration of therapy. Acute therapy is represented by: M=Meropenem, Ctx=Ceftriaxone, Ctz=Ceftazidime, PT=Piperacillin/Tazobactam, Cp=Ciprofloxacin, T=Tobramycin, G=Gentamicin, Pred=Prednisone, Imp=Imipenem, Azt=Aztreonam, Lz=Linezolid, Col=Colistin, Ampho = Inhaled Amphotericin, Cayston = Inhaled Aztreonam

Discussion

The CF lung is a complex ecosystem hosting a wide range of interacting bacteria, viruses and fungi that collectively alter host immune responses. This dynamic ecosystem drives short and long-term clinical outcomes of CF patients. To survive, these airway microbes must adapt their intermediary metabolism to the available resources to resist therapy and the host immune responses.

Combining clinical information with metagenomic analysis (Figure 4.4 – 4.6) has provided valuable insights into the potential use of sequencing in clinical settings. The patients were chosen from a larger cohort based on their distinct responses to treatments and levels of underlying lung function. CF6 was characterized by moderate lung disease but responded to the therapeutic plan, CF7 was characterized by severe end-stage disease and did not respond to therapy, and CF8 was characterized by mild pulmonary disease and completely responded to therapy. In case report format, we specifically demonstrated:

1) **Each patient hosts a unique polymicrobial community**. Metagenomics detected a high level of species diversity and community dynamicity not seen in routine cultures. Semi-quantitative measurements of individual species showed that their relative abundances fluctuated temporally. Whether these fluctuations are due to sampling or community dynamics remains to be determined.
2) **Community metabolism differed between patients, and within a patient over time.** Predictions of community metabolism suggested that this too is dynamic, changing over time and variable between patients that differed in their health status.

Although we cannot attribute these fluctuations to patient characteristics alone, as they may be due to sampling, it is clear that there is abundant variation for subsequent studies to investigate.

3) **An unappreciated diversity of genes encoding antibiotic resistance pathways was detected from the metagenomes in all patients** That these genes are often found in bacteriophage genomes makes it likely that they can be transferred horizontally between community members. Clinical decision-making may benefit from such information, to understand whether it is best to target individual bacteria or individual genes/functions.

Each of these points is discussed in detail below.

Unique polymicrobial communities

The results showed that the numerically dominant bacteria varied considerably between patients. Patient CF6 was represented mainly by *E. coli, R. mucilaginosa, S. parasanguinis,* and *Kleibsella pneumoniae*; patient CF7 was represented by *S. maltophilia, R. mucilaginosa,* and *Streptococcus* spp.; while patient CF8 was represented by *P. aeruginosa* and *R. mucilaginosa*. A previous study showed that lung bacteria are most likely acquired from the patient's living environment, and that the microbial community fluctuates in respond to therapeutic perturbations (Lim et al. 2012). The results presented here are consistent with these previous findings and extends the list of microbes known to be associated with the CF lungs. For example, patient CF8 had a high abundance of *Lactobacillus* spp. commonly found in the oral and gastrointestinal (GI) tract; its presence in the lung of immunocompromised individuals has been associated with life-threatening pulmonary cases (Jones et al. 1994). *Lactobacillus rhamnosus* can be introduced to the GI tract through yogurt and other dairy products (Holzapfel et al. 1998). Even though it is beneficial in most cases, it has also been associated with endocarditis (Avlami et al. 2001), pulmonary abscess and pleuritis (Shoji et al. 2010). *Lactobacillus casei* was found to have protective role in the lung of a mouse model during *S. pneumoniae* infection (Haro et al. 2009). In CF8, the presence of *Lactobacillus* spp. and the significant reduction of *Streptococcus* spp. following treatment may indicate a protective role for *Lactobacillus* sp. in *Streptococcus* infection. This is not surprising as a small study by Bruzzese *et al.* (Bruzzese et al. 2004) found that the treatment with probiotic *L. rhamnosus* GG (LGG) decreased the level of intestinal inflammation markers and rectal nitric oxide production in children with CF compared placebo-controlled group. A separate study further showed that children with CF treated with LGG showed a reduction of pulmonary exacerbations (Bruzzese et al. 2007).

The dynamics of the microbial communities within each patient support the *Climax* and *Attack* model previously described in an ecological view of the CF airways (Conrad et al. 2013). The main players (the *Climax* community) persist across time even though their abundances change with perturbations, while the *Attack* community is transient and dynamic. Every patient presented a complex lung microbial ecosystem consisting of distinct *Climax* and *Attack* communities.

Recently, 16S rRNA gene surveys have been suggested for routine clinical use (Salipante et al. 2013). However, 16S sequence data have a limited ability to resolve taxonomy to the species level (Filkins et al. 2012), and may introduce biases during the primer-binding step of PCR (Cai et al. 2013). For example, different species of *Pseudomonas* or *Streptococcus* could not be distinguished, and *Staphylococcus aureus* that was detected by culturing was not detected by sequencing (Filkins et al. 2012; Salipante et al. 2013).

Community metabolism

Functional information gleaned from the metagenomic data showed that the metabolic potentials of these communities were distinct, helping to determine whether a community should be labeled as *Climax* or *Attack*. The PCA of the top 20 most variable functions provided a preliminary view on the *Attack*-associated metabolic potentials (Figure 4.2, quadrants 1 and 3), and the *Climax*-associated metabolic potentials that render the community resistant to treatment and enable persistence through perturbations (Figure 4.2, quadrants 2, 3, and 4). It is important to note that the groupings of these samples are not mutually exclusive. The positive loadings of the first and second principal components may be representing metabolic potentials important for both recovery and response to treatment. The taxonomical and functional profiles of patient CF7 did not considerably change through time, which is consistent with this patient's unresponsiveness to treatment and his unchanged health status. This suggests the metabolic functions of the climax community were associated with persistence, and that changes in the *Attack* community were associated with exacerbation and declines in lung function.

Antibiotic resistance

Of the many challenges facing the CF community, the evolution of antibiotic resistance is one of the most pressing concerns. The presence of antibiotic resistance genes that encode resistance against major groups of antibiotics (e.g., aminoglycosides, macrolides, β-lactamases, and fluoroquinolones) suggests these communities may be capable of rapid genetic adaptation to resist perturbations and stresses imposed by treatment. This rapid adaptation would be at least partially fueled by horizontal gene transfer via phages and plasmids. Active multidrug efflux mechanisms are known to be one of the major determinants of antibiotic resistance in many CF pathogens including *P. aeruginosa*, *Burkholderia cepacia*, and *S. maltophilia* (Nikaido 1996; Alonso and Martínez 1997; Zhang et al. 2000). Our data (Figure 4.3A) support this observation as more than 50% of the antibiotic resistance genes identified were predicted to encode efflux-mediated resistance mechanisms. β-lactamases were the most abundant genes identified. Their high abundance in the viromes also suggested that phage-mediated spread of β-lactamases could occur within the community (Figure 4.3C). Similar to *P. aeruginosa*, *S. maltophilia* is highly resistant to antibiotics due to the presence of various intrinsic and acquired resistance mechanisms that include β-lactamases, penicillinase, cephalosporinase, aminoglycoside acetyl-transferase (*aac*), efflux pumps, and biofilm formation (Avison et al. 2001; Di Bonaventura et al. 2004; Falagas et al. 2009). An up-to-date database containing comprehensive annotations for aminoglycoside and macrolide resistance mechanisms facilitated the detection of resistance genes present in the metagenomic data, especially across the *S. maltophilia*-rich microbiomes of patient CF7. With this manually-curated database and data analysis framework, we have demonstrated the potential uses of metagenomics for the identification and monitoring of antibiotic resistance in clinical microbiology (Schmieder and Edwards 2012; Schmieder 2013).

The antibiotic resistance profiles of each patient were dynamic. In several cases, community members appeared to loose resistance to particular antibiotics, reflected in both clinical and metagenomic measurements. In the case of CF6, *E. coli* appeared to lose its resistance to Tobramycin, and *Streptococcus* and *Rothia* were predicted to be less resistant to aminoglycosides and macrolides, respectively (Figure 4.4). It remains to be determined whether these "losses" are due to fluctuations in the number of cells sampled or truly present, or the transfer of these resistance genes between community members. For example, a decrease in the number of *Streptococcus* cells sampled could explain the predicted loss in the presence of *Streptococcus*-associated antibiotic resistance genes in the later stable sample (D). But such a direct relationship could not explain the decrease in *Rothia* associated macrolide resistance genes co-occurring with an increase in *Rothia* abundance. In this case, it is possible that these genes are monitoring a separate, non-*Rothia*, community member.

The sequencing of the virome portion of CF lungs suggests also that antibiotic resistance is not likely confined to one bacterial species, for the genes conferring such resistances can be shuttled back and forth between microbes. In any case, having information on the predicted resistance of the whole community is perhaps one of the most useful pieces of information extracted from metagenome sequencing. If a patient's lung metagenome suggests resistance to macrolides, despite not having been prescribed such antibiotics, this can prove to be vital information for clinicians to prescribe appropriate antibiotic therapy. Our data provide a minimal estimate of the resistance potential for the community, and additional sequence coverage could be used to identify resistance genes present at lower abundance. Such rare genes are also important because they could come to dominate as the community composition is altered by antibiotic treatment.

Study limitations

This study illustrates the use of metagenomics to monitor microbial communities in the clinical setting, which may eventually help clinicians in their daily efforts to improve the lifes of CF patients. Because this metagenomic approach is at an early stage of development, the number of patients and samples presented in this study is relatively small (n=3 patients; 10 sputum samples) due to the amount of effort required for each sample. This limited sampling restricts our ability to determine whether the observed fluctuations in community composition, metabolism, and antibiotic resistance are truly occurring over time, or whether they are due to variability associated with sampling sputum, which may not consistently originate from the same region of the lung but from different regions that harbor their own persistent communities. Although larger studies will be needed to sort this out, this study has highlighted some of the most important issues to be solved prior to introducing personalized metagenomics into the clinic.

Another potential limitation concerns the ongoing controversy surrounding the extent of oropharyngeal contamination of sputum samples. Mouth wash and rinsing of the oral cavity using sterile saline solution prior to sputum induction is a National Institutes of Health (NIH)-recommended standard protocol for obtaining a minimally contaminated sputum sample (National Institute of Health). Previous studies showed significant evidence that induced CF sputum samples are strongly indicative of the lung environment and only minimally contaminated with mouth microbes (Rogers et al. 2006; Goddard et al. 2012; Fodor et al. 2012). Although we cannot rule out that some of our sputum samples were contaminated with oral microbes, the presence of such microbes in the oral cavity suggests they too could colonize CF lungs, and should therefore be considered as members of the community. This is particularly important for tracking antibiotic resistance genes, because even if an oral microbe has little chance of surviving in the lung environment, its genes may be transferred to those microbes that thrive in the lung environment.

Concluding remarks and study significance

The primary significance of this study is the combined use of metagenomic sequencing and clinical microbiology for monitoring polymicrobial infections in individual patients. This shotgun metagenomics approach not only provides accurate species level (sometimes strain level) taxonomic assignments, it also provides functional information at the gene level, e.g., the presence of potential antibiotic resistant genes and mutation-induced resistance mechanisms (Schmieder and Edwards 2012). In addition, the reconstruction of whole genomes is possible (Lim et al. 2013) and this can potentially provide important molecular information that is necessary for infection control (Dunne et al. 2012). Validation and normalization of the metagenomic data would also improve quantification of microbes, and the downstream clinical interpretation and therapeutic strategies. Other concerns and specific examples are also reviewed in (Dunne et al. 2012). Sample preparation, methodology, and bioinformatics will continue to improve these efforts, eventually leading to real-time monitoring of microbial communities in CF patients.

Medical diagnosis is a multidimensional process that includes physical assessment of the patient by physicians and nurses, nonspecific screening tests, monitoring of the efficacy of selected treatments, and the collection of specimens for biomedical laboratory processing. Nowadays, it is becoming increasingly possible to complement this information with sequence data. Real-time pathogen sequencing has been suggested to control pathogen outbreaks as current methods are slow and offer limited resolution (Dunne et al. 2012). As a proof-of-principle, this study presents the value of coupling metagenomics with clinical findings, helping to move us closer to molecular diagnoses. Diagrams such as those shown in Figures 4.4 – 4.6 would be instrumental in condensing vast amounts of data into clinically useful tools for tracking patient disease progression, corresponding treatments, and microbial community responses to those treatments.

The advancement in sequencing technologies and their decreasing cost is bringing us closer to diagnoses and treatments that are augmented by genomics technologies. To be relevant for clinical applications, the workflow is only possible with the aid of robotics and automation, and the turnaround times can be scaled to within 48 hours, which is over two times faster than a conventional culture-based procedure that takes 3-5 days for CF samples. Such timely information could affect clinical management of the patients. Of course, the quality of the data is very dependent on the choice of sequencing technology and data analysis pipeline. However, the optimization of upstream robotics, the further development of bioinformatics tools, and increasing computing power will continually move the field towards this goal. Although the implementation of metagenomic analysis as part of clinical diagnostic tool would be accompanied by the challenges of data interpretation by health care professionals, the consistency and accuracy of the technologies, and navigating the complexities of administrative policy, we see invaluable therapeutic potential in the real-time monitoring of microbial communities and their capabilities to resist treatment efforts.

Acknowledgments

This work was supported by the National Institute of Health (1 R01 GM095384-01) awarded to Forest Rohwer.

Chapter 4, in full, is published in the Journal of Clinical Microbiology. Yan Wei Lim, Jose S. Evangelista III, Robert Schmieder, Barbara Bailey, Matthew Haynes, Mike Furlan, Heather Maughan, Robert Edwards, Forest Rohwer, and Douglas Conrad; 2013. The dissertation author was the primary investigator and author of this paper. Article link: http://jcm.asm.org/content/52/2/425.full

References

Abubucker S, Segata N, Goll J, Schubert A, Rodriguez-Mueller B, Zucker J, team tHMPMR, Schloss P, Gevers D, Mitreva M, Huttenhower C (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLos Comput Biol. doi: 10:10.1371/journal.pcbi.1002358

Alonso A, Martínez JL (1997) Multiple antibiotic resistance in Stenotrophomonas maltophilia. Antimicrob Agents Chemother 41:1140–1142.

Avison MB, Higgins CS, von Heldreich CJ, Bennett PM, Walsh TR (2001) Plasmid location and molecular heterogeneity of the L1 and L2 beta-lactamase genes of Stenotrophomonas maltophilia. Antimicrob Agents Chemother 45:413–419. doi: 10.1128/AAC.45.2.413-419.2001

Avlami A, Kordossis T, Vrizidis N, Sipsas NV (2001) Lactobacillus rhamnosus endocarditis complicating colonoscopy. J Infect 42:283–285. doi: 10.1053/jinf.2001.0793

Bruzzese E, Raia V, Gaudiello G, Polito G, Buccigrossi V, Formicola V, Guarino A (2004) Intestinal inflammation is a frequent feature of cystic fibrosis and is reduced by probiotic administration. Aliment Pharmacol Ther 20:813–819. doi: 10.1111/j.1365-2036.2004.02174.x

Bruzzese E, Raia V, Spagnuolo MI, Volpicelli M, De Marco G, Maiuri L, Guarino A (2007) Effect of Lactobacillus GG supplementation on pulmonary exacerbations in patients with cystic fibrosis: a pilot study. Clin Nutr Edinb Scotl 26:322–328. doi: 10.1016/j.clnu.2007.01.004

Cai L, Ye L, Tong AHY, Lok S, Zhang T (2013) Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. PLoS ONE 8:e53649. doi: 10.1371/journal.pone.0053649

Chang FC, Chung JG (1998) Evidence for arylamine N-acetyltransferase activity in the Escherichia coli. Curr Microbiol 36:125–130.

Conrad D, Haynes M, Salamon P, Rainey PB, Youle M, Rohwer F (2013) Cystic fibrosis therapy: a community ecology perspective. Am J Respir Cell Mol Biol 48:150–156. doi: 10.1165/rcmb.2012-0059PS

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044. doi: 10.1371/journal.pone.0011044

Di Bonaventura G, Spedicato I, D'Antonio D, Robuffo I, Piccolomini R (2004) Biofilm formation by Stenotrophomonas maltophilia: modulation by quinolones, trimethoprim-sulfamethoxazole, and ceftazidime. Antimicrob Agents Chemother 48:151–160.

Dunne WM Jr, Westblade LF, Ford B (2012) Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol 31:1719–1726. doi: 10.1007/s10096-012-1641-7

Falagas ME, Kastoris AC, Vouloumanou EK, Rafailidis PI, Kapaskelis AM, Dimopoulos G (2009) Attributable mortality of Stenotrophomonas maltophilia infections: a systematic review of the literature. Future Microbiol 4:1103–1109. doi: 10.2217/fmb.09.84

Filkins LM, Hampton TH, Gifford AH, Gross MJ, Hogan DA, Sogin ML, Morrison HG, Paster BJ, O'Toole GA (2012) Prevalence of Streptococci and Increased Polymicrobial Diversity Associated with Cystic Fibrosis Patient Stability. J Bacteriol 194:4709–4717. doi: 10.1128/JB.00566-12

Fodor AA, Klem ER, Gilpin DF, Elborn JS, Boucher RC, Tunney MM, Wolfgang MC (2012) The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. PLoS ONE 7:e45001. doi: 10.1371/journal.pone.0045001

Goddard AF, Staudinger BJ, Dowd SE, Joshi-Datar A, Wolcott RD, Aitken ML, Fligner CL, Singh PK (2012) Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci 109:13769–13774. doi: 10.1073/pnas.1107435109

Haro C, Villena J, Zelaya H, Alvarez S, Agüero G (2009) Lactobacillus casei modulates the inflammation-coagulation interaction in a pneumococcal pneumonia experimental model. J Inflamm 6:28. doi: 10.1186/1476-9255-6-28

Holzapfel WH, Haberer P, Snel J, Schillinger U, Huis in't Veld JH (1998) Overview of gut flora and probiotics. Int J Food Microbiol 41:85–101.

Jones SD, Fullerton DA, Zamora MR, Badesch DB, Campbell DN, Grover FL (1994) Transmission of Lactobacillus pneumonia by a transplanted lung. Ann Thorac Surg 58:887–889.

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27 –30. doi: 10.1093/nar/28.1.27

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the Cystic Fibrosis gene: Genetic analysis. Science 245:1073–1080. doi: 10.1126/science.2570460

Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–2275. doi: 10.1056/NEJMoa1109910

Lim YW, Schmieder R, Haynes M, Furlan M, Matthews TD, Whiteson K, Poole SJ, Hayes CS, Low DA, Maughan H, Edwards R, Conrad D, Rohwer F (2013) Mechanistic model of Rothia mucilaginosa adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data. PLOS ONE 8:e64285. doi: 10.1371/journal.pone.0064285

Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. J Cyst Fibros Off J Eur Cyst Fibros Soc. doi: 10.1016/j.jcf.2012.07.009

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clin Microbiol Rev 23:299 –323. doi: 10.1128/CMR.00068-09

Liu B, Pop M (2009) ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res 37:D443–D447. doi: 10.1093/nar/gkn656

National Institute of Health Critical care therapy and respiratory care section.

Nikaido H (1996) Multidrug efflux pumps of gram-negative bacteria. J Bacteriol 178:5853–5859.

Quinton PM (2010) Role of epithelial $HCO3^-$ transport in mucin secretion: lessons from cystic fibrosis. Am J Physiol Cell Physiol 299:C1222–1233. doi: 10.1152/ajpcell.00362.2010

Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Kehagia V, Connett GJ, Bruce KD (2006) Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with Cystic Fibrosis. J Clin Microbiol 44:2601–2604. doi: <p>10.1128/JCM.02282-05</p>

Ross JS, Ali SM, Wang K, Palmer G, Yelensky R, Lipson D, Miller VA, Zajchowski D, Shawver LK, Stephens PJ (2013) Comprehensive genomic profiling of epithelial ovarian cancer by next generation sequencing-based diagnostic assay reveals new routes to targeted therapies. Gynecol Oncol 130:554–559. doi: 10.1016/j.ygyno.2013.06.019

Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG (2013) Rapid 16S rRNA Next-Generation Sequencing of Polymicrobial Clinical Samples for Diagnosis of Complex Bacterial Infections. PLoS ONE 8:e65226. doi: 10.1371/journal.pone.0065226

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6:e17288. doi: 10.1371/journal.pone.0017288

Schmieder R, Edwards R (2012) Insights into antibiotic resistance through metagenomic approaches. Future Microbiol 7:73–89. doi: 10.2217/fmb.11.135

Schmieder RA (2013) A framework for identifying antibiotic resistance in the human microbiome.

Schomburg PD, Chang DA (2006) Arylamine N-acetyltransferase. Springer Handb Enzym 243–258.

Shoji H, Yoshida K, Niki Y (2010) Lung abscess and pleuritis caused by Lactobacillus rhamnosus in an immunocompetent patient. J Infect Chemother Off J Jpn Soc Chemother 16:45–48. doi: 10.1007/s10156-009-0004-5

Tran B, Brown AMK, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, Stein LD, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE (2013) Feasibility of real time next generation sequencing of cancer genes linked to drug response: Results from a clinical trial. Int J Cancer 132:1547–1555. doi: 10.1002/ijc.27817

Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J (2013) Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol 51:232–237. doi: 10.1128/JCM.01696-12

Willner D, Furlan M (2010) Deciphering the role of phage in the cystic fibrosis airway. Virulence 1:309–313. doi: 10.4161/viru.1.4.12071

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in Cystic Fibrosis and non-Cystic Fibrosis individuals. PloS One 4:e7370. doi: 10.1371/journal.pone.0007370

Zhang L, Li XZ, Poole K (2000) Multiple antibiotic resistance in Stenotrophomonas maltophilia: involvement of a multidrug efflux system. Antimicrob Agents Chemother 44:287–293.

Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, Young VB, LiPuma JJ (2012) Decade-long bacterial community dynamics in Cystic Fibrosis airways. Proc Natl Acad Sci 109:5809–5814. doi: 10.1073/pnas.1120577109

Appendix for Chapter 4

Supplementary Notes

Protocol defined exacerbation

A protocol-defined exacerbation is an event when the patient:

1) Meets four or more of the Fuchs criteria (Fuchs et al. 1994). Fuchs criteria are most commonly used to evaluate exacerbation events in adult CF individuals. It has also been extensively used in several clinical trials to define a CF exacerbation. We have added the missed days of school/work as an additional criterion.

2) The patient must have a physiologic drop in FEV_1 of at least 15% or more compared to the best FEV_1 in the previous 12 months.

Patient is assigned as clinical responder when the patient shows:

3) Improvement in all of the Fuchs symptoms/signs that declined from the baseline value at the onset of the exacerbation.

4) Improvement in FEV_1 to within 90% and 250 cc of their greatest FEV_1 achieved in the past 12 months.

5) Improvement in the subjective CFQR and UCSD SOBQ surveys. These are validated disease specific questionnaires. A minimally clinically important difference of greater than 5 (MCID>5) was used as a cut off for level of significance. All others who do not meet these criteria were considered as non-responders.

Sample description, collection and processing

Each sample was given a unique patient ID (CF<number>) followed by the time point when the sample was collected (represented by a letter, <A-Z>). A patient’s clinical status, designated by the clinician, was based on the commonly used Fuch's criteria, lung function tests, and the patient's reported outcome. Sputum samples were collected in the clinic during the patient's visit. The sample was then syringe-homogenized and split into 5 aliquots for metagenomic (2) and metatranscriptomic (1) studies, culturing (1), and storage (1). Immediately after homogenization, each homogenate for microbial metatranscriptomic study was transferred to a 15 ml falcon tube containing 1 volume of 0.1 mm zirconia beads and 3 volumes of Trizol LS and was immediately vortexed for 10 minutes at medium speed to mechanically lyse microbial cells while maintaining the RNA intact. All samples were then transported on ice to the lab.

Detailed standard protocols can be downloaded from www.coralandphage.com. The protocols include initial sample pre-processing and pre-treatment prior to microbial cell and virus enrichment. Due to the large files describing every method, and possible deviations from the SOP, the website allows researchers to choose the sample type and procedure to be done, and automatically generates an appropriate corresponding workflow.

Bioinformatics

All samples were sequenced using the GS-FLX Titanium chemistry system. Multiplexed SFF sequence data files were separated according to their unique identifiers, and FASTA formatted sequences and corresponding quality scores were extracted using the GS-SFF tools software package (Roche: Brandord, CT).

Data Preprocessing

All datasets were preprocessed using PRINSEQ to remove low quality reads, reads shorter than 60 bp, duplicate reads, and low complexity reads. The command used was:

```
perl prinseq-lite.pl -verbose -log -fastq file.fastq -derep 1245 -lc_method entropy -lc_threshold 50 -trim_qual_right 15 -trim_qual_left 15 -trim_qual_type mean -trim_qual_rule lt -trim_qual_rule lt -trim_qual_window 2 -trim_tail_left 5 -trim_tail_right 5 -min_len 60 -min_qual_mean 15 -ns_max_p 1 -rm_header
```

Viral and microbial metagenomes were further processed using DeconSeq to remove all human-like sequences with at least 90% query length coverage and 90% identity. This was done using the web version available at http://edwards.sdsu.edu/deconseq

Reference Databases

The NCBI non-redundant protein (hereafter NR) database (version Feb 14, 2012) was downloaded from the NCBI FTP server: ftp://ftp.ncbi.nih.gov/blast/db/

The NCBI non-redundant nucleotide (hereafter NT) database (version Feb 14, 2012) was downloaded from the NCBI FTP server: ftp://ftp.ncbi.nih.gov/blast/db/

The viral database (created Feb 23, 2012) includes: 4,019 unique viral genomes downloaded from the NCBI FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/ and ftp://ftp.ncbi.nih.gov/refseq/release/viral/

The Anitbiotic Resistance Genes Database (ARDB) (version 1.1, July 3, 2009) includes 23,137 antibiotic resistant-associated protein sequences downloaded from http://ardb.cbcb.umd.edu. In order to normalize the number of hits against the size of the metagenomes (total number of reads per metagenome), as well as the total number of base pairs in the database, we assume an average of 1000 bp per gene.

Database searches

Database searches were performed using the BLAST program. Unless specified, the default command-line options were used. Fine-tuning of the options based on the characteristics of the input data may yield better performance and/or results. Analysis of the BLAST output was performed using in-house Perl scripts. BLAST version 2.2.24 was downloaded from: ftp://ftp.ncbi.nih.gov/blast/executables/release/LATEST/

Data Analysis

rRNA-like and non-rRNA reads were identified from the preprocessed metatranscriptomes using BLASTn against the SILVA database (threshold of 50% query coverage and 75% alignment identity). Non-rRNA reads were annotated using BLASTx against the NCBI non-redundant protein database (threshold of 40% identity over at least 60% of the query sequence).

The preprocessed metagenomes were annotated using BLASTn against the NCBI nucleotide database (threshold of 40% identity over at least 60% of the query sequence). Sequences assigned to the phylum Chordata and to vector or synthetic sequences were identified and removed. Virome sequences were then compared against the viral database containing 4,019 unique viral genome sequences using a tBLASTx search (threshold of 40% identity over at least 60% of the query sequence).

The best hit designation was assigned to the alignment with the highest coverage, identity and score values within the specified thresholds. For BLASTx against NR, if there were multiple amino acid alignments (within the top 50 BLAST hits) against the same database sequence without overlap in both the query and database sequence, and within the length of the query sequence, the combined coverage, identity and score values were calculated for each query sequence to account for possible frame-shifts.

Taxonomic and functional assignments

The query sequence taxonomy and/or function were assigned based on the best matching database sequence(s). If there were multiple best hits with the same coverage, identity and score values that belonged to different taxa, or the matching database sequence belonged to different taxa, then the taxonomies/functions were randomly assigned using 100,000 bootstraps. This approach is similar to assigning an equal fraction to all possible taxa, but additionally provides the standard deviation for each assigned mean value. Query sequences with no BLAST hits and those unassigned due to the defined threshold were classified as “unassigned” or “unknown”. The diversity of microbiomes was calculated based on the number of bacterial species identified in the datasets.

All metagenomes were additionally annotated using the KEGG database (Kanehisa and Goto 2000) and analyzed using the HUMAnN pipeline (Abubucker et al. 2012). The normalized relative abundance values were used for subsequent analysis. The top 20 pathways that vary the most between microbiomes were used for the principal component analysis (PCA) illustrated in Figure 4.2, and the analysis was done using the R package “bpca”.

References

Abubucker S, Segata N, Goll J, Schubert A, Rodriguez-Mueller B, Zucker J, team tHMPMR, Schloss P, Gevers D, Mitreva M, Huttenhower C (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLos Computational Biology. doi: 10:10.1371/journal.pcbi.1002358

Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash MI, Ramsey BW, Rosenstein BJ, Smith AI, Wohl ME (1994) Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with Cystic Fibrosis. The New England Journal of Medicine 331:637–642.

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28:27 –30. doi: 10.1093/nar/28.1.27

Supplementary Table 4.1: Library information of the viral and microbial metagenomes. The number of reads after host removal indicates the number of reads used for all downstream analysis.

Name	Viral Metagenomes			Microbial Metagenomes		
	% CVS[1]	# Processed Reads[2]	% Unknown[3]	% CVS[1]	# Processed Reads[2]	% Unknown[4]
CF6-A-Ex	61%	36,302	98.11%	95%	5,957	1%
CF6-B-Tr	79%	22,864	97.67%	58%	71,825	5%
CF6-C-Pt	31%	38,246	94.78%	38%	29,360	18%
CF6-D-St	17%	64,465	94.90%	96%	1,714	2%
CF7-A-Ex	42%	50,282	98.68%	23%	42,765	10%
CF7-B-Tr	36%	134,937	90.91%	23%	28,840	10%
CF7-C-Tr	24%	60,984	82.64%	82%	12,836	2%
CF7-D-Pt	24%	124,498	67.85%	75%	29,282	3%
CF8-A-Ex	52%	51,083	98.62%	30%	106,166	18%
CF8-B-Pt[5]	58%	47,501	97.82%	78%	8,706	6%

[1] Percentage of total high-quality reads that matched to Chordata (assuming these are host contamination), vector and synthetic (CVS) sequences. [2]Number of high-quality reads minus the number of reads matched to CVS sequences based on DeconSeq (microbial metagenomes only) and BLASTn against the nucleotide (NT) database. [3] Percentage of total sequences that do not have significant hits against the viral genome database using tBLASTx. [4] Percentage of total sequences that do not have significant hits against the NT database using BLASTn. [5] This library contains 21% of contaminating vector sequences.

Supplementary Table 4.2: Questionnaires score

Patient (Gender)	Time Scale	CFQR [1]	Respiratory Score[2] (MCID±5)	UCSD SOBQ [3] (MCID±5)	FEV_1[4]
CF6 (Female)	Day 0	625	61	34	1.91 (57%)
	Day 12	603	66	43	2.03 (60%)
	Day 17	626	66	32	2.06 (61%)
	Day 46	897	72	24	2.07 (61%)
CF7 (Male)	Day 0	907	55	36	0.87 (21%)
	Day 20	823	66	40	0.80 (19%)
	Day 27	883	55	29	0.82 (19%)
	Day 37	848	66	45	0.92 (22%)
CF8 (Male)	Day 0	686	33	20	3.39 (73%)
	Day 17	933	83	8	4.15 (89%)

[1] Cumulative score of CFQR (Cystic Fibrosis Questionnaire - Revised) measuring the 12 domains representing patients' quality of life. The 12 domains include physical, role, vitality, emotion, social, body image, eating, treatment burden, health perceptions, and symptom scales based on weight, respiratory and digestion. [2] Respiratory score based on the CFQR. [3] SOBQ: Shortness of Breath Questionnaire. [4] FEV_1 is measured as the forced expiratory volume in one second (% predicted).

Supplementary Table 4.3: Antibiotic resistance profiles predicted by BLASTx comparison of microbiome data to the Antibiotic Resistance Database (ARDB).

Sample	Resistant genes	Resistance class	Resistance types
CF6-A	acr, bla-a, bla-c, bla-d, tet efflux, mph, mdtef, mdtnop, tet-rpp, ksga, bcr-mfs, baca, macab, arna, mdtm.	Multidrug resistance efflux pump. Macrolide-specific efflux system. Class A beta-lactamase. Class C beta-lactamase. Class D beta-lactamase. Macrolide phosphotransferase. Tetracycline efflux pump. Ribosomal protection protein. Undecaprenyl pyrophosphate phosphatase.	Tetracyclin. Acriflavin. Aminoglycoside. Beta-lactam. Glycylcycline. Macrolide. Ceftazidime (and other cephalosporins). Puromycin. Doxorubicin. Erythromycin. Cloxacillin. Penicillin. Deoxycholate. Fosfomycin. Kasugamycin. Polymycin. Bacitracin. Chloramphenicol. Norfloxacin.
CF6-B	bla-a, bla-c, bla-d, acr, mdtnop, mdtef, tet-efflux, sul, mph, bcr mfs, mdfa, emrd, mdtg, macab, arna, baca, rosab, aac, mdtk, mdth, mdtl, ksga, dfra, mdtm, erm, tet-rpp, emre, catb	Multidrug resistance efflux pump. Class A beta-lactamase. Tetracycline efflux pump. Sulfonamide-resistant dihydropteroate synthase. Macrolide phosphotransferase. Macrolide-specific efflux system. Aminoglycoside N-acetyltransferase. Undecaprenyl pyrophosphate phosphatase. Class C beta-lactamase. Class D beta-lactamase. Group A drug-insensitive dihydrofolate. rRNA adenine N-6-methyltransferase. Ribosomal protection protein. Group B chloramphenicol acetyltransferase.	Tetracyclin. Acriflavin. Aminoglycoside. Beta-lactam. Glycylcycline. Macrolide. Ceftazidime (and other cephalosporins). Puromycin. Doxorubicin. Erythromycin. Cloxacillin. Penicillin. Deoxycholate. Fosfomycin. Kasugamycin. Polymycin. Bacitracin. Chloramphenicol. Norfloxacin. Sulfonamide. Amikacin. Dibekacin. Isepamicin. Netilmicin. Sisomicin. Tobramycin. Enoxacin. Norfloxacin. Trimethoprim. Lincosamide. Streptogramin-b.

Supplementary Table 4.3: Antibiotic resistance profiles predicted by BLASTx comparison of microbiome data to the Antibiotic Resistance Database (ARDB). (continue)

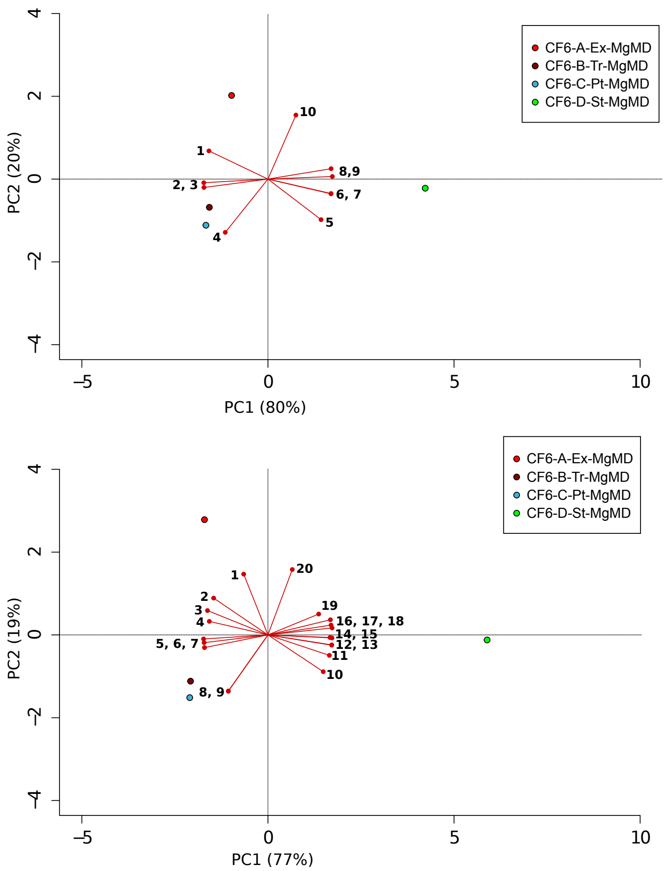
Sample	Resistant genes	Resistance class	Resistance types
CF6-C	bla-a, acr, mdtnop, mdtk, mdtef, sul, mph, bcr mfs, macab, arna, baca, mdtl, dfra, mdtm, erm, tet-rpp, catb	Multidrug resistance efflux pump. Class A beta-lactamase. Macrolide-specific efflux system. Sulfonamide-resistant dihydropteroate synthase. Macrolide phosphotransferase. Undecaprenyl pyrophosphate phosphatase. rRNA adenine N-6-methyltransferase. Ribosomal protection protein. Group B chloramphenicol acetyltransferase.	Ceftazidime (and other cephalosporins). Macrolide. Chloramphenicol. Doxorubicin. Erythromycin. Bacitracin. Acriflavin. Puromycin. T-chloride. Aminoglycoside. Beta-lactam. Glycylcycline. Enoxacin. Norfloxacin. Sulfonamide. Lincosamide. Streptogramin-b, Polymyxin. Tetracycline.
CF6-D	aac, bla-a.	Aminoglycoside N-acetyltransferase. Class A-beta-lactamase.	Amikacin. Dibekacin. Isepamicin. Netilmicin. Sisomicin. Tobramycin. Ceftazidime (and other cephalosporins). Monobactam. Penicillin.
CF7-A	smeabc, smedef, tet-rpp, bla-a, macab, mecr1, tet-flavo, mexxy, mls-mfs.	Multidrug resistance efflux pump. Ribosomal protection protein. Class A beta-lactamase. Macrolide specific efflux system. Methicillin-resistance regulatory protein for mecA. Flavoproteins. Macrolide-Lincosamide-Streptogramin B efflux pump.	Fluoroquinolone. Tetracycline. Macrolide. Methicillin. Aminoglycoside. Glycylcycline. Cephalosporin.
CF7-B	tet-rpp, smedef, smeabc, pbp, erm.	Multidrug resistance efflux pump. Ribosomal protection protein. rRNA adenine N-6-methyltransferase.	Fluoroquinolone. Tetracycline. Penicillin. Lincosamide. Macrolide. Streptogramin-b.

Supplementary Table 4.3: Antibiotic resistance profiles predicted by BLASTx comparison of microbiome data to the Antibiotic Resistance Database (ARDB). (continue)

Sample	Resistant genes	Resistance class	Resistance types
CF7-C	smeabc, smedef, pbp	Multidrug resistance efflux pump.	Fluoroquinolone. Penicillin.
CF7-D	smabc, smedef, tet-rpp, aac, adeabc, catb	Multidrug resistance efflux pump. Ribosomal protection protein. Aminoglycoside N-acetyltransferase. Group B chloramphenicol acetyltransferase.	Fluoroquinolone. Aminoglycoside. Chloramphenicol. Tetracycline. Amikacin. Dibekacin. Isepamicin. Netilmicin. Sisomicin. Tobramycin.
CF8-A	tet-rpp, pbp, mexhi, mexef, mls mfs, mexab, erm, mexxy, mexcd, mexvw, bla-a, bla-c, baca, macab, arna, rosab, pmra.	Multidrug resistance efflux pump. Ribosomal protection protein. Macrolide-Lincosamide-Streptogramin B efflux pump. rRNA adenine N-6-methyltransferase. Class A beta-lactamase. Class C beta-lactamase. Undecaprenyl pyrophosphate phosphatase.	Tetracycline. Penicillin. Macrolide. Chloramphenicol. Fluoroquinolone. Aminoglycoside. Beta-lactam. Tigecycline. Lincosamide. Streptpgramin-b. Erythromycin. Roxithromycin. Cephalosporin. Bacitracin. Norfloxacin. Ciprofloxacin. Polymyxin. Fosmidomycin.
CF8-B	mexvw, mexhi, tet-rpp, mexcd, erm, mexab, baca, mexxy, bla-c, mexef, smedef.	Multidrug resistance efflux pump. Ribosomal protection protein. rRNA adenine N-6-methyltransferase. Class C beta-lactamase. Undecaprenyl pyrophosphate phosphatase.	Tetracycline. Lincosamide. Macrolide. Streptogramin-b, Erythromycin. Fluoroquinolone. Glycylcycline. Roxythromycin. Aminoglycoside. Beta-lactam. Tigecycline. Cephalosporin. Chloramphenicol. Bacitracin.

Supplementary Table 4.4: Antibiotic resistance profile assigned by BLASTx comparison against the Antibiotic Resistance Database (ARDB) based on the viromes.

Sample	Resistant genes	Resistance class	Resistance types
CF6-A	blaA	Class A β-lactamase.	Cephalosporins. Penicillin
CF6-B	blaA	Class A β-lactamase	Cephalosporins. Penicillin
CF6-C	tet_efflux, cml	Major facilitator superfamily transporter: tetracycline efflux pump, chloramphenicol efflux pump	Tetracycline, Chloramphenicol
CF6-D	blaA	Class A β-lactamase.	Cephalosporins, Penicillin
CF7-A	fos	Glutathione transferase	Fosfomycin
CF7-B	Not-detected	-	-
CF7-C	Not-detected	-	-
CF7-D	tet_efflux	Major facilitator superfamily transporter, tetracycline efflux pump	Tetracycline
CF8-A	erm, rosab	rRNA adenine N-6-methyltransferase. Efflux pump/potassium antiporter system	Lincosamide, Streptpgramin-b, Macrolide, Fosmidomycin.
CF8-B	blaA	Class A β-lactamase.	Cephalosporin, Penicillin



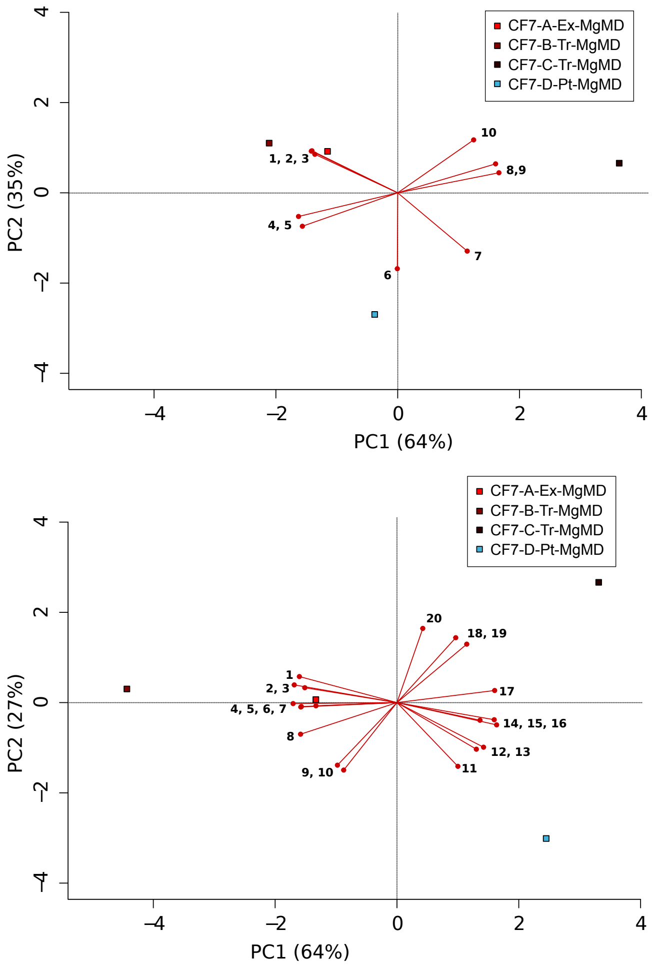
Top 10 metabolic pathways with greatest variance

1. Selenocompound metabolism (ko00450)
2. C5-Branched dibasic acid metabolism (ko00660)
3. Protein export (ko03060)
4. Drug metabolism - other enzymes (ko00983)
5. D-Glutamine ana D-glutamate metabolism (ko00471)
6. Valine leucine and disoleucine biosynthesis (ko00290)
7. Sulfur relay system (ko04122)
8. Folate biosynthesis (ko00790)
9. D-Alanine metabolism (ko00473)
10. Flagellar assembly (ko02040)

Top 20 metabolic pathways with greatest variance

1. Streptomycin biosynthesis (ko02060)
2. Bacterial chemotaxis (ko02030)
3. Selenocompound metabolism (ko00450)
4. Chloroalkane and chloroalkene degradation (ko00625)
5. Citrate cycle-TCA cycle (ko00020)
6. C5-Branched dibasic acid metabolism (ko00660)
7. Protein export (ko03060)
8. Carbon fixation pathways in prokaryotes (ko00720)
9. Drug metabolism - other enzymes (ko00983)
10. D-Glutamine ana D-glutamate metabolism (ko00471)
11. Taurine and hypotaurine metabolism (ko00430)
12. Valine leucine and disoleucine biosynthesis (ko00290)
13. Sulfur relay system (ko04122)
14. Biotin metabolism (ko00780)
15. Peptidoglycan biosynthesis (ko00550)
16. Folate biosynthesis (ko00790)
17. Ubiquinone and other terpenoid-quinone biosynthesis (ko00130)
18. D-Alanine metabolism (ko00473)
19. Toluene degradation (ko00623)
20. Flagellar assembly (ko02040)

Supplementary Figure 4.1: Principal component analysis (PCA) showing the (a) top 10 and (b) top 20 metabolic pathways that displayed the greatest variance within patient CF6 samples.

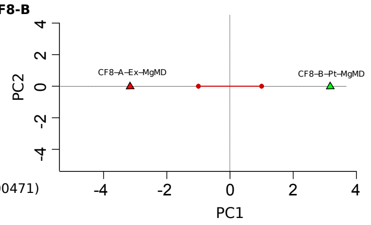


Top 10 metabolic pathways with greatest variance
1. Phosphotransferase system-PTS (ko02060)
2. Sulfur relay system (ko04122)
3. D-Glutamine and D-glutamate metabolism (ko00471)
4. Naphthalene degradation (ko00626)
5. Beta-alanine metabolism (ko00410)
6. Fatty acid biosynthesis (ko00061)
7. Biotin metabolism (ko00780)
8. D-alanine metabolism (ko00473)
9. Riboflavin metabolism (ko00740)
10. Lipoic acid metabolism (ko00785)

Top 20 metabolic pathways with greatest variance
1. Aminoacyl tRNA biosynthesis (ko00970)
2. Fructose and mannose metabolism (ko00051)
3. Pantothenate and CoA biosynthesis (ko00770)
4. Phosphotransferase system-PTS (ko02060)
5. Sulfur relay system (ko04122)
6. D-Glutamine and D-glutamate metabolism (ko00471)
7. Ribosome (ko03010)
8. Base excision repair (ko03410)
9. Naphthalene degradation (ko00626)
10. Beta-alanine metabolism (ko00410)
11. Fatty acid biosynthesis (ko00061)
12. Limonene and pinene degradation (ko00903)
13. Valine leucine and disoleucine degradation (ko00280)
14. Chlorocyclohexane and chlorobenzene degradation (ko00361)
15. Lipopolysaccharide biosynthesis (ko00540)
16. Biotin metabolism (ko00780)
17. Styrene degradation (ko00643)
18. D-alanine metabolism (ko00473)
19. Riboflavin metabolism (ko00740)
20. Lipoic acid metabolism (ko00785)

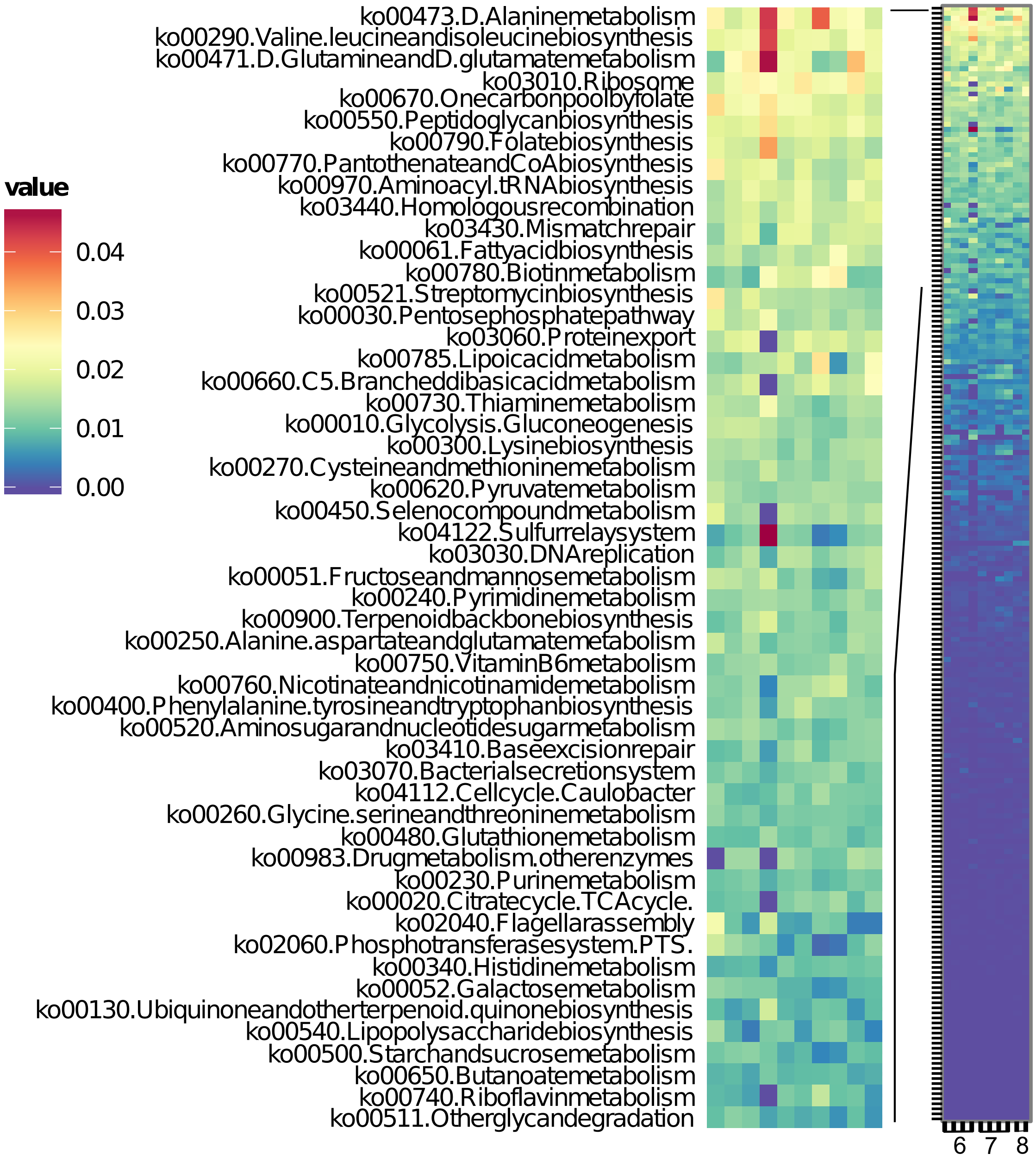
Supplementary Figure 4.2: Principal component analysis (PCA) showing the (a) top 10 and (b) top 20 metabolic pathways that displayed the greatest variance within patient CF7 samples.

Metabolic pathways seperating CF8-A from CF8-B
1. Folate biosynthesis (ko00790)
2. Peptidoglycan biosynthesis (ko00550)
3. Other glycan degradation (ko00511)
4. Amino acyl-tRNA biosynthesis (ko00970)
5. Dioxin degradation (ko00621)
6. Lipopolysaccharide biosynthesis (ko00540)
7. Riboflavin metabolism (ko00740)
8. D-Alanine metabolism (ko00473)
9. Naphthalene degradation (ko00626)
10. Ribosome (ko03010)
11. D-Glutamine ana D-glutamate metabolism (ko00471)



Metabolic pathways separating CF8-B from CF8-A
12. Fructose and mannose metabolism (ko00051)
13. Ubiquinone and other terpenoid-quinone biosynthesis (ko00130)
14. Sulfur metabolism (ko00920)
15. Phosphotransferase system PTS (ko02060)
16. Citrate cycle - TCA cycle (ko00020)
17. Geraniol degradation (ko00281)
18. Toluene degradation (ko00623)
19. C5-Branched dibasic acid metabolism (ko00660
20. Lipoic acid metabolism (ko00785)

Supplementary Figure 4.3: Principal component analysis (PCA) showing the top 20 metabolic pathways that displayed the greatest variance within patient CF8 samples.



Supplementary Figure 4.4: Functional characterization of the microbiomes based on the normalized metabolic pathway abundances in each sample. The Y-axis shows only the top 45 orthologous groups sorted by the total mean value of each orthologous group.

Chapter 5

Mechanistic model of *Rothia mucilaginosa* adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data

Abstract

The impaired mucociliary clearance in individuals with Cystic Fibrosis (CF) enables opportunistic pathogens to colonize CF lungs. Here we show that *Rothia mucilaginosa* is a common CF opportunist that was in present in 83% of our patient cohort, almost as prevalent as *Pseudomonas aeruginosa* (89%). Sequencing of lung microbial metagenomes identified unique *R. mucilaginosa* strains in each patient, presumably due to evolution within the lung. The *de novo* assembly of a near-complete *R. mucilaginosa* (CF1E) genome illuminated a number of potential physiological adaptations to the CF lung, including antibiotic resistance, utilization of extracellular lactate, and modification of the type I restriction-modification system. Metabolic characteristics predicted from the metagenomes suggested *R. mucilaginosa* have adapted to live within the microaerophilic surface of the mucus layer in CF lungs. The results also highlight the remarkable evolutionary and ecological similarities of many CF pathogens; further examination of these similarities has the potential to guide patient care and treatment.

Introduction

Cystic fibrosis (CF) is a genetic disease caused by mutation of the cystic fibrosis transmembrane conductance regulator (CFTR) gene (Kerem et al. 1989). In CF lungs, the defective CFTR protein affects trans-epithelial ion transport and consequently leads to the accumulation of thick and static mucus. The resultant hypoxic microenvironment encourages the colonization of opportunistic microbes, viruses, and fungi (reviewed in (LiPuma 2010)), causing acute and chronic infection. A few of the most commonly isolated pathogens are *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Haemophilus influenzae,* and *Burkholderia cepacia*. However, an increasing number of microbial species have been detected in the CF airway using culture-independent methods such as metagenomic sequencing (Rogers et al. 2004; Harris et al. 2007; Bittar et al. 2008; Cox et al. 2010; Guss et al. 2011). Metagenomics is a powerful approach that has been used to successfully characterize the microbial and viral communities in CF individuals (Willner et al. 2009; Willner and Furlan 2010; Willner et al. 2011b; Willner et al. 2012; Lim et al. 2012). These types of studies have illuminated the complexity of microbial and viral communities, captured the vast diversity of functions encoded by these organisms, and have been used to trace the evolution of whole genomes (Narasingarao et al. 2012; Iverson et al. 2012).

Previous sequencing of CF metagenomes revealed the presence of *Rothia mucilaginosa* at relatively high abundances in most patients (Lim et al. 2012). *R. mucilaginosa* was first isolated from milk in 1900 as *Micrococcus mucilaginosus* (Migula 1900). It was later re-isolated and further studied by Bergan *et al.* in 1970 (Bergan et al. 1970), and renamed *Stomatococcus mucilaginosus* in 1982 based on its 16S rDNA and biochemical characteristics (Bergan and Kocur 1982). A recent study comparing *S. mucilaginosus* to *Rothia dentocariosa* and another unknown species (later known as *Rothia nasimurium*) led to the reclassification of *S. mucilaginosus* as *R. mucilaginosa* (Collins et al. 2000).

R. mucilaginosa is an encapsulated, Gram-positive non-motile coccus (arranged in clusters) belonging to the phylum Actinobacteria. It has variable catalase activity, reduces nitrate, and hydrolyses aesculin (Bergan and Kocur 1982; Collins et al. 2000; Doel et al. 2005). It is a facultative anaerobe commonly found in the human oral cavity and upper respiratory tract (Bergan et al. 1970; Olsen et al. 2009; Guglielmetti et al. 2010), and occasionally the gastrointestinal tract (Wang et al. 2005), small intestinal epithelial lining (Ou et al. 2009), tongue (Kazor et al. 2003; Preza et al. 2009), teeth (Nyvad and Kilian 1987; K. Philip et al. 2009), colostrum (Jiménez et al. 2008), breast milk (Delgado et al. 2008), and dental plaques (G.H. 1969; Ready et al. 2004). Although *R. mucilaginosa* is commonly regarded as normal flora of the oral cavity and upper respiratory tract, its association with a wide range of diseases (Supplementary Table 5.1) highlights its potential as an opportunistic pathogen, especially in immuno-compromised patients (Stackebrandt 2006).

At the genus level, *Rothia* has been reported by Tunney *et al.* (Tunney et al. 2008) as an aerobic species that can be isolated from CF sputum and pediatric bronchoalveolar lavage (BAL) samples. It has also been detected under anaerobic culturing conditions and via 16S rRNA gene surveys (Tunney et al. 2008). Typically *R. dentocariosa* was the main species identified in these studies (van der Gast et al. 2011). In addition, Bittar *et al.* (Bittar et al. 2008) and Guss *et al.* (Guss et al. 2011) have characterized *R. mucilaginosa* as a “newly” emerging CF pathogen. Even so, *R. mucilaginosa* is usually treated as part of the normal oral microbiota in the clinical lab. As a result, the presence of *R. mucilaginosa* in CF lungs may be under-reported and the significance of infection is underestimated.

Here we confirm that *R. mucilaginosa* is present and metabolically active in the lungs of CF patients. Comparisons with a non-CF reference genome revealed the presence of unique *R. mucilaginosa* strains in each patient. A near-complete genome was reconstructed from the metagenomic reads of one patient; comparison of these sequence data with a non-CF reference genome enabled the identification of unique genomic features that may have facilitated adaptation to the lung environment.

Materials and Methods

Microbial metagenome data: Induced sputum samples were collected from CF volunteers at the Adult CF Clinic (San Diego, CA, United States) by expectoration. All collection was approved by the University of California Institutional Review Board (HRPP 081500) and San Diego State University Institutional Review Board (SDSU IRB#2121). Written informed consent was provided by study participants and/or their legal guardians. Fresh CF sputum samples were processed as described in (Lim et al. 2012). In brief, sputum samples were homogenized, bacterial cells were pelleted by centrifugation, and pellets were repeatedly washed and then treated with DNase to remove human DNA prior to extraction of bacterial DNA.

Sequence read processing. A total of 18 microbiomes were previously sequenced using Roche-454 GSFLX (Lim et al. 2012). The data were downloaded from NCBI sequence read archive (Accession # SRP009392). Reads that were duplicates or of low quality were removed using PRINSEQ (Schmieder and Edwards 2011a), and those that matched human-derived sequences were removed using DeconSeq (Schmieder and Edwards 2011b). Sequence reads with similarity to the phylum Chordata and to vector or synthetic sequences were identified by BLASTn against NCBI nucleotide database (threshold of 40% identity over at least 60% query coverage), and removed from the metagenomes. A detailed description of sample processing and preliminary analyses of these datasets has been published (Lim et al. 2012).

BWA mapping of the metagenomes: The processed metagenomic reads were mapped to the *Rothia mucilaginosa* DY-18 (GI: 283457089) reference genome using a modified version of BWA-SW 0.5.9. The coverage values based on the reference mapping are shown in Supplementary Table 5.4.

De novo *assembly and scaffolding:* The metagenomic reads from CF1E were *de novo* assembled using the Newbler software version 2.6 with ≥35bp overlap and ≥95% identity. All resultant contigs were aligned to the reference genome (*R. mucilaginosa* DY-18) using nucmer with its -maxmatch option (using all anchor matches regardless of their uniqueness). This option will allow repetitive or multi-copy sequences (e.g., rRNA operons) to assemble into a single contig, enabling that contig to be subsequently mapped to more than one genomic region. All alignments were examined manually. Full length contigs were ordered based on their coordinates on the reference alignment, and this ordering was used with an in-house Perl script to build the final scaffold containing 181 contigs.

Genome annotation: The CF1E scaffold was annotated using the RAST web annotation service (Aziz et al. 2008) with the latest FIGfams version 57 (Genome ID: 43675.9). In order to allow a direct comparison, the reference genomes of *R. mucilaginosa*, DY-18 and *R. mucilaginosa* M508 (downloaded from the Genome OnLine Database) were also annotated using the same pipeline. CRISPR loci were identified using CRISPRFinder (Grissa et al. 2007). The spacers between the repeats were extracted and compared to the virome sequenced from the same sample (downloaded from the NCBI; SRX090639) (Lim et al. 2012), and other viromes in mymgdb (Schmieder and Edwards).

Rothia-targeted 16S PCR of lung sections from explanted lungs. DNA was extracted from 5-6 homogenized lung tissues from explanted lungs of four transplant patients using the Macherey-Nagel Nucleospin Tissue Kit (Macherey-Nagel, Bethlehem, PA) with the Gram-positive variation that included an overnight proteinase K digestion. Extracted DNA was amplified using Actinobacteria-targeted PCR primers (Rothia_1F: 5’-GGGACATTCCACGTTTTCCG-3’, Rothia_1R: 5’-TCCTATGAGTCCCCACCATT-3’) that encompass a 322bp region of the 16S rRNA gene including the hypervariable regions 6–7. Two of the four patients were positive for Actinobacteria; right lower and lingular (left) lobes for Lung 9, and lingular lobe for Lung 7 (Supporting Information S3, Supplementary Figure 5.5). The PCR products were purified and sequenced. Sequencing of the three partial 16S gene fragments indicated *Rothia* was present in lungs from one of the four CF patients.

Results and Discussion

Mutations in the CFTR locus affect proper ion transport in lung epithelial cells, impairing the clearance of mucus in the airways and encouraging microbial colonization and persistence. Until recently, most laboratory and clinical microbiology only focused on a few pathogens, particularly *P. aeruginosa*. This "one mutation, one pathogen" model of the CF ecosystem is being replaced by a "polyphysiology, polymicrobial" view that is expected to improve treatment until gene-therapy is able to fix the underlying genetic cause.

An important step in understanding the CF lung ecosystem, with the ultimate goal of eliminating microbes or altering their pathogenicity, is to determine which microbes are present and the ways in which their survival depends on local chemistry. An initial step in this direction is to use metagenomic and metatranscriptomic data previously generated from microbes and viruses present in CF sputum samples (Lim et al. 2012). Metagenomic data provide information on which microbes and viruses are present, and their metabolic capabilities, while metatranscriptomic data provide information on which organisms are metabolically active (Lim et al. 2012). Such metabolism data will enable predictions of local lung chemistry that may impose patient-specific selective pressures on the microbes. Data from eighteen microbiomes existing in six CF patients with different health statuses were analyzed here (Supporting Information S1; Supplementary Table 5.2).

The metagenomic data showed that 15/18 (83%) sputa contained *R. mucilaginosa* and 16/18 (89%) sputa contained the common CF pathogen *P. aeruginosa*. Although *P. aeruginosa* was present in a greater number of samples, its abundance was lower than that of *R. mucilaginosa* in 11 of the 14 samples where these species co-existed. Both species abundances ranged from 1% to 62% (Figure 5.1). The relative percentages of these two opportunistic pathogens varied between patients and within the same patient as their health status changed. The results show no obvious pattern of synergy or competition between the two pathogens.

A health status of ‘Ex’ (for exacerbation) indicates a stark decline in lung function that is typically treated with intravenous antibiotics. Thus, between a health status of ‘Ex’ and ‘Tr’, patients will have been given antibiotics in addition to those that are often prescribed as continued therapy. The abundance data in Figure 5.1 indicate that these exacerbation-associated antibiotic treatments did little to permanently exclude *R. mucilaginosa* from CF lung communities. Patients CF1, CF4, CF6, CF7, and CF8 all had appreciable abundances of *R. mucilaginosa* by the last sampling time point. Only the *R. mucilaginosa* population in CF5 did not recover from antibiotic treatment by the last sampling time point; however, as this patient was only followed for 21 days (compared to 17-58 for the other patients), it is possible *R. mucilaginosa* could still rebound from antibiotic treatment. These results indicate that *R. mucilaginosa* is able to survive the typical CF antibiotic treatment, as is the main CF pathogen *P. aeruginosa*.

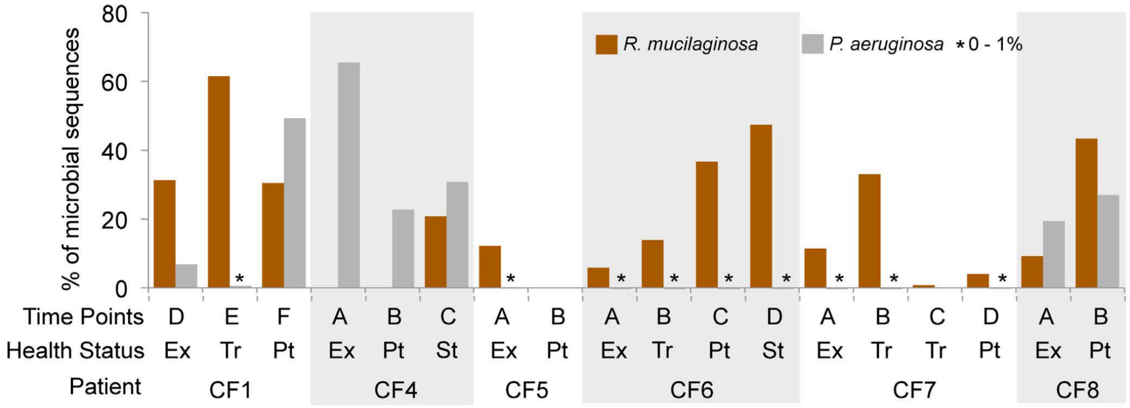


Figure 5.1: Prevalence of *Rothia mucilaginosa*, and the prototypical pathogen *Pseudomonas aeruginosa*, in eighteen microbiomes from six CF patients. Patients were sampled at times of differing health status (Supplementary Table 5.1). Ex: Exacerbation; Tr: On treatment; Pt: Post treatment; St: Stable; * present in <1% of the microbiome.

R. mucilaginosa is present in CF lung explants and metabolically active. The presence of *R. mucilaginosa* DNA in sputum samples (as detected by metagenome sequencing) could be explained by its abundance in the oral cavity and subsequent contamination of the sputum during collection. However, this is unlikely for several reasons. Previous studies have indicated little contamination of sampled sputa with oral inhabitants (Rogers et al. 2006; Goddard et al. 2012), and the presence of *Rothia* has been confirmed in CF lungs (Fodor et al. 2012). Examination of lung tissue samples was the best way to definitively determine the presence of *R. mucilaginosa* in CF lungs from our cohort. Between 5 and 6 lung tissue sections from explanted lungs of each of four transplant patients were screened for *Rothia*-related microbes using 16S rDNA targeted PCR and sequencing. One out of the four patients was positive for *Rothia* (Supporting Information S3, Supplementary Figure 5.5), indicating this bacterium is indeed present within lung airways. Unfortunately this patient was not available for the metagenome sequencing. The *R. mucilaginosa* population present in the oral cavity may serve as a reservoir and "stepping stone" for lower respiratory infection, as described in many respiratory chronic infections such as CF and chronic obstructive pulmonary disease (Gomes-Filho et al. 2010).

The presence of *R. mucilaginosa* DNA in sputum or lung tissues does not necessarily indicate this bacterium is metabolically active in the lung environment. Examination of a metatranscriptome dataset indicated that mRNAs and rRNAs are being produced by *Rothia* species (Supporting Information S2; Supplementary Figure 5.2-5.3), which suggests this bacterium is metabolically active in the CF lung.

Genetic differences of R. mucilaginosa between patients. Longitudinal studies of *P. aeruginosa* (Oliver et al. 2000), *Burkholderia dolosa* (Lieberman et al. 2011), and *Staphylococcus aureus* (Goerke and Wolz 2010) within and between CF patients have shown evolutionary adaption to the CF lung. Here, we define adaptation as a process where mutations that alter pathogen behavior (in this case, metabolism) become fixed in response to specific environmental pressures, e.g. the availability of nutrients, oxygen, or redox potential. The power of the metagenomic data is in its ability to uncover the genetic mutations underlying these adaptations, that occur over long periods of selection. Characterizing these mutations thus enables us to infer which selection pressures are strongest in the CF lung, whether they be the dynamic lung physiology, immune system surveillance, and/or antibiotic treatment.

We found evidence for similar evolutionary adaptation in *R. mucilaginosa*. The metagenomic sequences from each sample were mapped separately against the reference genome *R. mucilaginosa* DY-18, (GI: 283457089; originally isolated from persistent apical periodontitis lesions (Yamane et al. 2010). As shown in Figure 5.2, the mapped sequences reveal gaps where portions of the reference genome sequence were not covered by metagenomic reads (i.e., were absent) in the CF-derived datasets (gap patterns > 5 kbp shown in Figure 5.2; Supplementary Table 5.3). The out-group in Figure 5.2 is due to low coverage of *R. mucilaginosa* reads in the metagenomes of these patient samples (<1X coverage; Supplementary Table 5.4). Most of the gaps occurred in regions of low GC content (Figure 5.2), which most likely represent genes acquired by DY-18 via horizontal gene transfer (Lawrence and Ochman 1998).

The gap patterns were most different between patients, indicating unique *R. mucilaginosa* strains exist in each patient. Within each patient, differences in gap patterns between time points were less numerous, but their existence indicates that the genome of *R. mucilaginosa* has been evolving independently in each patient. Combined with similar findings for *P. aeruginosa* (Oliver et al. 2000) and *S. aureus* (Goerke and Wolz 2010), this suggests that essentially every CF patient harbors a unique strain of *R. mucilaginosa* that evolves in the lung. If each strain also has a unique antibiotic resistance profile, then CF treatment will need to be tailored to the particular strain present in each patient.

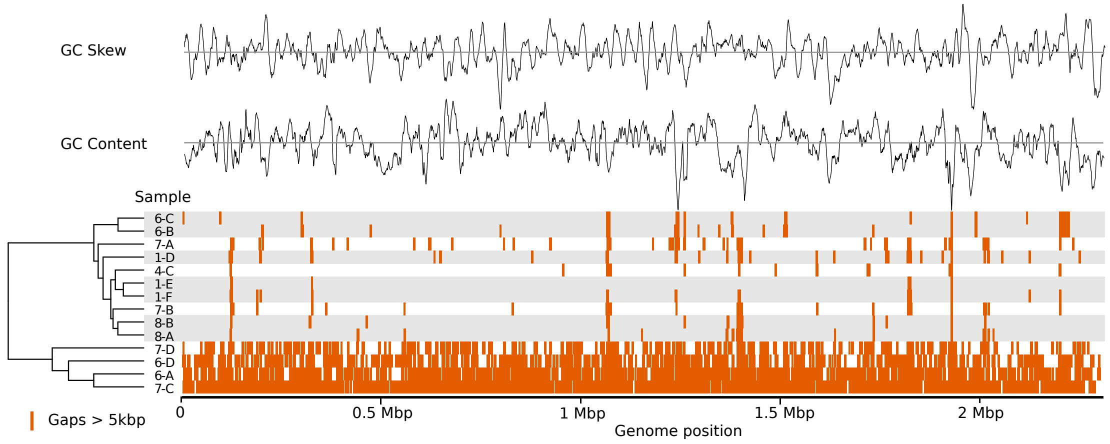


Figure 5.2: Hierarchical clustering of the sample based on gap patterns, which correspond to regions of the reference genome *R. mucilaginosa* DY-18 that were not represented by any metagenomic reads. BWA mapping was used and gaps were identified in a 1 kbp stepwise window. Only gaps ≥5 kbp were plotted. Exact coordinates and annotations for the gaps are in Supplementary Table 5.3. The clade composed of 7-D, 6-D, 6-A, and 7-C appears to be lacking the majority of the reference genome, due to the low sequence coverage of *R. mucilaginosa* in these metagenomes.

Characteristics of CF1E genome scaffold. The metagenome from CF1E had over 40,000 reads mapping to the reference genome, indicating enough data may be present to reconstruct the full genome of the *R. mucilaginosa* strain present. All CF1E metagenomic reads were assembled *de novo* into 996 contigs with a N50 value (weighted median value of all contigs) of 11,178bp. Contigs were aligned against the reference genome *R. mucilaginosa* DY-18 using nucmer (Kurtz et al. 2004), resulting in one single scaffold built from 181 contigs with an 8.8-fold average sequencing depth (Figure 5.3). The CF1E *R. mucilaginosa* genome scaffold was then annotated using the RAST server (Genome ID: 43675.9) and compared to DY-18 that had been re-annotated using the same pipeline.

The CF1E genome scaffold consists of one circular chromosome of 2,278,618 bp with a GC content of 59.6%. Only large indels are reported here and SNPs were not examined. No large rearrangements were detected between CF1E and the reference genome DY-18. Phylogenetic analysis of the 16S rDNA loci indicated CF1E and the reference strain DY-18 are close relatives (Supporting Information S3; Supplementary Figure 5.4), which is consistent with their average pairwise nucleotide identity of 85%. The sequence reads were relatively equally distributed across the genome, except at the multi-copy rRNA genes and in the highly conserved *rhs* region (Figure 5.3).

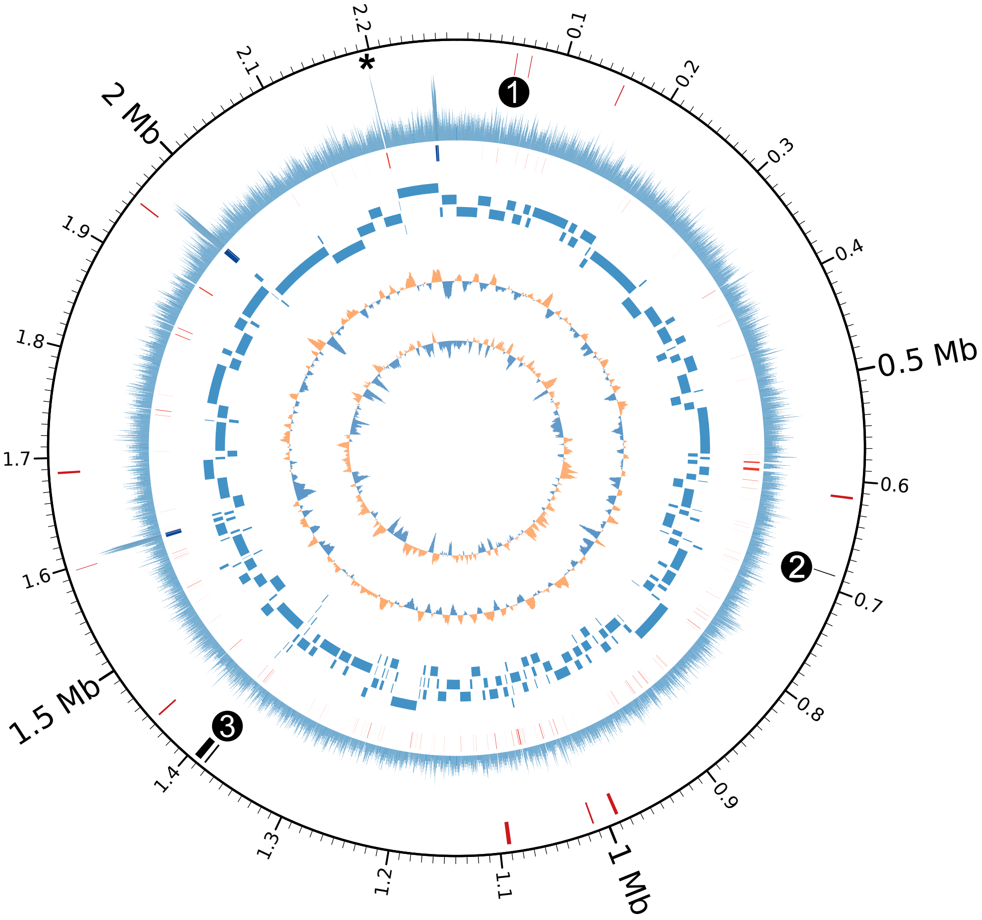


Figure 5.3: Circular representation of *R. mucilaginosa* CF1E draft genome. Genome coordinates are given in Mbp. From outside to inside, the circles represent: (i) Fragments missing in the DY-18 reference (red); CRISPR region and phage-associated genes as 1) phage lysin, 2) phage shock protein, and 3) CRISPR elements, (ii) Coverage of the genome up to the scale of 50X; * marked Region V containing a genome fragment with an average coverage of 38 X (peaks at 48X) at the region of rearrangement hotspot (rhs) elements (iii) Gaps in the scaffold (red); rRNA operons (blue) (iv) Contig order and size; (v) GC skew; (vi) GC content deviation.

The high coverage regions – rRNA operons and rhs elements. The rRNA operons assembled into one single contig (contig173) and had an average coverage depth of approximately 3.5 times the average depth of coverage for the rest of the scaffold (i.e. 31X versus 8.8X). The sequence from this contig was used to fill in three gaps that were predicted to correspond to the rRNA operons, based on alignment with the reference genome (Figure 5.3).

The rearrangement hot spot (*rhs*) gene region (Figure 5.3: Region V marked *) also had a high average coverage of 39X. The primary structure of Rhs proteins consists of an N-terminal domain, a "core" domain, a hyperconserved domain, and a DPxGL motif followed by a C-terminus that varies between strains and species (Jackson et al. 2009). Previous studies have shown that *rhs* genes play a role in competition between strains or species, similar to the contact-dependent growth inhibition (CDI) system (Poole et al. 2011). The variable C-termini of Rhs proteins have toxin activities, and the small genes that typically follow *rhs* genes are thought to encode proteins that provide immunity to the toxins. Kung *et al.* (2012) showed that the *rhs-CT* in *P. aeruginosa* delivers toxins to eukaryotic cells, activating the inflammasome (Kung et al. 2012). The high coverage of the conserved *rhs* region suggests that *rhs* is present in high abundance in the CF microbial community. It is possible that the *rhs* system is widely used by CF microbes for (i) cell-to-cell interactions and communication, particularly for biofilm formation, (ii) direct antagonistic effects on the growth or viability of competitors, and/or (iii) attacking cells of the host-immune system. Additional experimental studies are needed to further assess these possibilities.

The high coverage *rhs* region in the CF1E genome scaffold included an *rhs* gene sequence related to one of the two *rhs* genes of DY-18 (RMDY18_19250). However, there is an apparent gap in the scaffold sequence, beginning 24 amino acids upstream of the DPxGL motif of the encoded Rhs protein. In the DY-18 reference genome, this gap corresponds to the coding sequence for the toxic C-terminal region of Rhs, and the beginning of the gene encoding the RhsI immunity protein. Assuming the presence of multiple *rhs-CT/rhsI* modules in the metagenome, assembling this region will be challenging.

Functional annotation of the R. mucilaginosa genome scaffold. RAST predicted 1,739 gene products belonging to 248 function subsystems (Supplementary Table 5.5). The most abundant functions included biosynthesis and degradation of amino acids and derivatives, protein metabolism, cofactor/vitamin/prosthetic group/pigment biosynthesis and metabolism, and carbohydrate metabolism (Supplementary Table 5.6). Thirty-seven ORFs present in the DY-18 genome and absent in the CF1E scaffold are listed in Table 5.1 (DY-18 specific). Genes only present in CF1E are listed in Table 5.2 (CF1E specific). Genome regions specific to only one of the two strains ranged from multiple kbp (mostly in gene coding regions) to a few nucleotides in non-coding regions. Additional analyses were performed on several of the genomic regions unique to CF1E; regions were chosen for their potential influences on CF-lung specific evolution of niche utilization and antibiotic resistance.

(i) L-lactate dehydrogenases (LDHs): The CF1E scaffold had a cytochrome c-dependent LDH (EC 1.1.2.3) in addition to the expected NAD (P)-dependent LDH (EC 1.1.1.27). The nucleotide sequence of LDH (EC 1.1.2.3) was 80% identical to the LDH

of *R. dentocariosa* ATCC 17931. Lactate is secreted by the human host, and produced by many CF-associated microbes (e.g., *Staphylococcus* and *Streptococcus* spp.) through fermentation (De Backer et al. 1997). Lactate has been detected in the CF sputum at a mean concentration of 3 mM, and higher concentrations have been correlated with lower lung function (Bensel et al. 2011). Because LDHs enable cells to use lactate as an energy source for growth and reproduction, they are considered as virulence factors. For example, utilization of lactate by *Neisseria* spp. (reviewed in Smith et al. 2007) enhances their rate of O_2 metabolism (Britigan et al. 1988).

R. mucilaginosa is a facultative anaerobe. The presence of both types of LDH may allow cells to respond to micro-changes in oxygen and nutrient availability by utilizing different metabolic pathways. This would indicate that the primary niche of *R. mucilaginosa* is the microaerophilic environment at the epithelial surfaces in the mucus plug, which also contains lactate and oxygen from the cells and blood, respectively. A cytochrome c-dependent LDH could allow *R. mucilaginosa* to utilize extracellular L-lactate with cytochrome c as the terminal oxidase (Lederer 1974) under aerobic conditions, producing pyruvate and hydrogen peroxide (H_2O_2) (Garvie 1980). Pyruvate could serve as a food reservoir for fermentative bacteria (e.g., *R. mucilaginosa* in the CF lung (Price-Whelan et al. 2007)) while also inhibiting the glucose uptake rate of competing bacteria (Brown and Whiteley 2007). The production of H_2O_2 could also serve to inhibit the growth of other organisms, or be used by microbes with catalase activity to yield water that is scarce in the dehydrated CF lung environment (Potter et al. 1967; Tarran et al. 2001). Under anaerobic conditions, NAD-dependent LDH allows the organism to undergo fermentation through the reduction of pyruvate to lactate (reviewed in (Garvie 1980)**.**

(ii) Antibiotic resistant genes: An additional copy of a gene encoding the macrolide export ATP-binding/permease protein MacB was found into the CF1E scaffold. Sequence alignments showed that the two MacB-encoding genes are only 12% identical at the nucleotide level, indicating that one MacB was acquired horizontally and did not originate by gene duplication. The protein sequence of the acquired MacB matched a hypothetical protein in *R. mucilaginosa* M508 and MacB from *R. mucilaginosa* ATCC 25296 (E-value: 0). The predicted amino acid sequences showed specific hits to the family comprising the MJ0796 ATP-binding cassette (CD03255), followed by a MacB-like periplasmic core domain (PFAM 12704) and FtsX-like permease family (PFAM02687) domain.

In addition, modulator of drug activity B (MdaB) was present in the CF1E scaffold. Overexpression of MdaB has been shown to confer resistance against tetracycline and adriamycin in *E. coli* (Adams and Jia 2006). In addition to this gene, the genome of *R. mucilaginosa* in CF1E encoded drug resistance transporters (EmrB/QacA subfamily), multidrug resistance transporters (Bcr/CflA family), and a glycopeptide antibiotic resistance protein. These diverse strategies for antibiotic resistance may underlie *R. mucilaginosa's* ability to survive antibiotic treatments (Figure 1).

(iii) Type I restriction modification: The type I restriction modification (R-M) system is a mechanism to protect against foreign nucleic acids via non site specific endonucleases (Murray 2000). There are three subunits: M (Modification/Methyltransferase), S (Specificity) and R (Restriction). The M and S

subunits are responsible for recognizing self and non-self, while the R subunit performs the cleavage. The S subunit contains two target recognition domains that are important for restriction specificity and modification of the complex activity. Mapping of metagenomic reads to the reference DY-18 genome (Figure 5.2, Supplementary Table 5.3) showed that only the CF1E metagenome had this Type I R-M system region, whereas the other metagenomes had gaps of 7-9 kbp around this region of the DY-18 genome. The CF1E scaffold likely encodes an S subunit with different sequence specificity, as this subunit is only 37% identical (nucleotides) or 41% identical (protein) to the DY-18 copy (Supplementary Table 5.7 and Supplementary Figure 5.1). This is of interest because Type I R-M systems have been modified during the adaptation of *P. aeruginosa* and *Burkholderia cenocepacia* to the CF lung. For example, the Type I R-M of *P. aeruginosa* Liverpool epidemic strain (LES) colonizing CF patients was shown to carry a different regulatory specificity (M-subunit) in comparison to strain PA01 (Smart et al. 2006). In addition, the expression of type I R-M was greatly increased in *B. cenocepacia* in the presence of sub-inhibitory concentrations of antibiotics (Sass et al. 2011). Together these observations suggest that modification of type I R-M system could be a general mechanism for adaptation to the CF lung.

(iv) Phage lysin: Phage lysins are anti-bacterial agents often used in bacterial competition, and have also been associated with the release of cellular components to the extracellular medium during biofilm formation (Whitchurch et al. 2002; Carrolo et al. 2010). One copy of the phage lysin gene was present in CF1E, but this did not have any appreciable nucleotide similarity to any genes in phage or bacteria. However, bioinformatic analysis of the predicted amino acid sequence revealed its similarity to the

N-acetylmuramoyl-L-alanine amidase of *R. mucilaginosa* ATCC strain 25296 (E-value: 10^{-150}), a hypothetical protein of *R. mucilaginosa* M508 (E-value: 10^{-148}), and an amidase-5 domain similar to pneumococcal bacteriophage Dp-1 (E-value: 6.88 X 10^{-42}). Phage lysins are commonly found in prophages (Schmitz et al. 2010). However, no prophages were detected in the CF1E genome scaffold based on PhiSpy (Akhter et al. 2012). Although it is currently unclear what, if any, advantage is offered by this phage lysin in the *R. mucilaginosa* genome, this lysin could provide an alternative strategy for microbial competition.

(v) Clusters of interspaced short palindromic repeats (CRISPRs): CRISPRs are characterized by stretches of short sequence repeats that flank short "spacer" sequences composed of viral or plasmid DNA. Four CRISPR elements were identified in CF1E; these were all ~4 kbp downstream of the Cas1 CRISPR-associated gene. The length of these CRISPRs ranged from 253 bp to 1,316 bp (Supplementary Table 5.8). All CRISPRs contained the same direct repeat sequence of 36 bp. The spacers in each CRISPR element (collectively referred to as a 'spacer set') ranged in copy number from 3 to 17, and their sizes ranged from 33 bp to 88 bp. Two of the spacer sets code for hypothetical proteins while the other two sets are unknown (Supplementary Table 5.8). A total of 48 spacer sequences were extracted from the four spacer sets; these spacers were compared to the CF1E virome sequences, but no similarities were found.

Table 5.1: Genomic regions present in the DY-18 reference genome but missing from the CF1E draft genome. See details in Supplementary Table 5.11.

Region starting coordinate (on DY-18)	Region size (bp)	Protein(s) annotated on the genomic fragment
44,199	697	Predicted ARSR subfamily of helix-turn-helix bacterial transcription regulatory protein CDS
119,455	7,752	Conserved hypothetical protein, putative cell filamentation protein CDS
138,835	470	Predicted nucleic acid-binding protein CDS; exopolysaccharide biosynthesis protein related to N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase CDS
149,644	1,290	UBA/THIF-type NAD/FAD binding fold CDS
319,311	5,105	ATP-binding protein of ABC transporter
620,953	3,564	Putative glutamate transporter permease protein CDS; ABC-type amino acid transport system, permease component CDS; Glutamate binding protein CDS; COG1126: ABC-type polar amino acid transport system, ATPase component CDS
1,410,305	915	Putative integral membrane protein CDS
1,817,923	2,084	Pyruvate oxidase [ubiquinone, cytochrome] (EC 1.2.2.2) CDS
1,924,698	994	Cell wall surface anchor family protein CDS
2,084,618	5,926	Amino acid ABC transporter, periplasmic amino acid-binding protein CDS; Cystathionine gamma-lyase (EC 4.4.1.1) CDS; O-acetylhomoserine sulfhydrylase (EC 2.5.1.49) CDS
Total of 20		Hypothetical proteins*

Table 5.2: Predicted protein-coding sequences present in the CF1E scaffold annotated by RAST, but missing from the DY-18 reference (list excludes hypothetical proteins). See details in Supplementary Table 5.12.

Inserted position (based on DY-18)	Fragment length (bp)	Protein annotated on the inserted fragment
53,748	3,174	Phage lysin, N-acetylmuramoyl-L-alanine amidase CDS
65,967	913	Modulator of drug activity B
70,542	1,534	Putative DNA-binding protein
156,926	1,615	Predicted nucleic acid-binding protein CDS
239,949	1,690	Putative hydrolase CDS
613,830	2,685	Acyltransferase 3 CDS
808,821	318	Mobile element protein CDS
1,013,020	1,866	2-oxoglutarate/malate translocator CDS
1,671,769	3,475	Macrolide export ATP-binding/permease protein MacB (EC 3.6.3.-) CDS
1,884,458	1,840	Mobile element protein CDS
2,044,067	1,547	L-lactate dehydrogenase (EC 1.1.2.3) CDS
Total of 21		Hypothetical proteins*

Phages are an important source of genes in microbial communities. The CRISPRs found in *R. mucilaginosa* CF1E may correspond to previously attacking phages and plasmids that these cells were able to resist. In order to identify these phage perpetrators, spacer sequences were compared against all host-associated and environmental viromes in MyMgDB (Schmieder and Edwards). One of the spacers was identified in two human oral cavity viromes (Willner et al. 2011a), whereas none of the spacers were similar to sequences from other environmental viromes (Supplementary Table 5.9). The results suggest these bacteria may have been exposed to phages found in the oral cavity, which suggests cells may have existed in this environment prior to opportunistic infection of the CF lungs. Because these spacer sequences did not match phages in the virome sequenced from the same sample, the phages to which *R. mucilaginosa* is resistant are not present, or are below the detection limit, in this sample. However, if temperate phages dominate in the CF lung (Willner et al. 2011a) as in the human gut virome (Reyes et al. 2010), this result is expected because the virome would largely composed of free-living viruses. However it is also possible that these CRISPRs do not protect the cells against phage infection, but are involved in a CRISPR-dependent modulation of biofilm formation, as described previously in *P. aeruginosa* (reviewed in Palmer and Whiteley 2011). Biofilm formation has been shown to be important for persistent bacterial infection of CF lungs, as well as an overall decline in lung function. Therefore, the role of these CRISPRs in CF1E and other CF lung isolates' pathogenesis should be explored further.

Conclusions

The metagenomic and genomic analyses presented here suggest that *R. mucilaginosa* is a common inhabitant of CF lungs, and that it evolves and adapts to each patient's lung environment over the course of a persistent infection. Genomic analysis of CF1E highlighted many potential adaptations: multiple genes encoding L-lactate dehydrogenases (LDHs) that could enable utilization of lactate, many multi-drug efflux pumps for antibiotic resistance, and the modification of *rhs* elements and the type I restriction system. Alterations of the type I restriction system has the potential to influence horizontal transfer of genes. The CF1E genomic sequence indicates extensive phage-host interactions, including the acquisition of a phage lysin and changing CRISPR elements.

Based on these potential metabolic adaptations, we hypothesize that *R. mucilaginosa* lives in the microaerophilic surface of the viscous mucus layer that is characteristic of CF airways (Figure 5.4). Under this hypothesis, cytochrome c-dependent LDH would enable *R. mucilaginosa* to use extracellular lactate. However, this process would require oxygen, which is more readily available at the surface of the mucus layer (e.g., from the blood). As the oxygen level is depleted, metabolism could be supported by fermentation and anaerobic respiration with nitrate as an alternative electron acceptor, as observed in *P. aeruginosa* (Hoffman et al. 2010). Persistence in low oxygen environments would also allow for evasion of antibiotics and ROS activity. In addition, *R. mucilaginosa* carries a low-pH induced ferrous ion (Fe^{2+}) transporter along with heme and hemin uptake and utilization systems. Co-occurring CF pathogens including *P. aeruginosa* and *S. maltophilia* are known to synthesize redox active phenazines that are able to reduce Fe^{3+} to Fe^{2+} (Dietrich et al. 2008; Wang et al. 2011) potentially giving *R. mucilaginosa* access to Fe^{2+} in the low pH sputum where the ferrous ion transporter is induced.

The results presented here highlight the similar evolutionary trajectories and ecological niches of several species of bacteria that colonize the CF lung. These similarities are remarkable because each bacterial species starts with different genetic material: *P. aeruginosa* has a relatively large genome (> 6 Mbp), whereas *R. mucilaginosa* has only a 2 Mbp genome (Supplementary Table 5.10). These findings suggest that obtaining strain specific genome data can illuminate patient-specific bacterial inhabitants of CF patients. This specific information enables predictions to be made regarding the bacteria's physiological adaptations in each patient, which would further enable physicians to optimize antibiotic treatments.

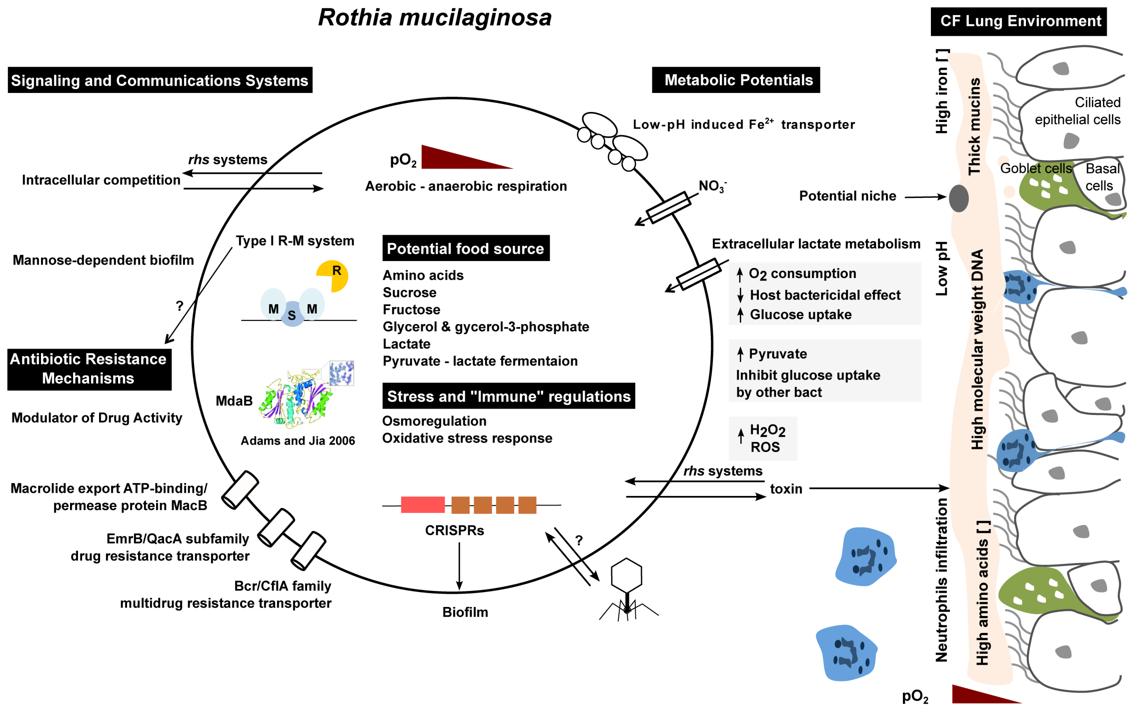

Rothia mucilaginosa
Signaling and Communications Systems
rhs systems
Intracellular competition
Mannose-dependent biofilm
?
Antibiotic Resistance Mechanisms
Modulator of Drug Activity
Type I R-M system
R
M
S
M
MdaB
Adams and Jia 2006
Macrolide export ATP-binding/ permease protein MacB
EmrB/QacA subfamily drug resistance transporter
Bcr/CflA family multidrug resistance transporter
pO2
Aerobic - anaerobic respiration
Potential food source
Amino acids
Sucrose
Fructose
Glycerol & gycerol-3-phosphate
Lactate
Pyruvate - lactate fermentaion
Stress and "Immune" regulations
Osmoregulation
Oxidative stress response
CRISPRs
Biofilm
?
Metabolic Potentials
Low-pH induced Fe2+ transporter
NO3-
Extracellular lactate metabolism
↑ O2 consumption
↓ Host bactericidal effect
↑ Glucose uptake
↑ Pyruvate
Inhibit glucose uptake by other bact
↑ H2O2
ROS
rhs systems
toxin
Potential niche
CF Lung Environment
High iron [ ]
Thick mucins
Low pH
High molecular weight DNA
Neutrophils infiltration
High amino acids [ ]
Ciliated epithelial cells
Goblet cells
Basal cells
pO2

Figure 5.4: Hypothesized adaptions of *R. mucilaginosa* to the CF lung environment. The model is based on the comparison between the reference genome DY-18 and the reconstructed genome CF1E. rhs: rearrangement hot spot; Type I R-M: Type I restriction modification; MdaB: Modular of drug activity B; ROS: reactive oxygen species; CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats.

Acknowledgments

This work was supported by the National Institute of Health (1 R01 GM095384-01) awarded to Forest Rohwer.

Chapter 5, in full, is published in the PLoS ONE journal. Yan Wei Lim, Robert Schmieder, Matthew Haynes, Mike Furlan, T. David Matthews, Katrine Whiteson, Stephen J. Poole, Christopher S. Hayes, David A. Low, Heather Maughan, Robert Edwards, Douglas Conrad, and Forest Rohwer; 2013. The dissertation author was the primary investigator and author of this paper. Article Link: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064285

References

Adams MA, Jia Z (2006) Modulator of drug activity B from Escherichia coli: crystal structure of a prokaryotic homologue of DT-diaphorase. J Mol Biol 359:455–465. doi: 10.1016/j.jmb.2006.03.053

Akhter S, Aziz RK, Edwards RA (2012) PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-based and composition-based strategies.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. doi: 10.1186/1471-2164-9-75

Bensel T, Stotz M, Borneff-Lipp M, Wollschläger B, Wienke A, Taccetti G, Campana S, Meyer KC, Jensen PØ, Lechner U, Ulrich M, Döring G, Worlitzsch D (2011) Lactate in cystic fibrosis sputum. Journal of Cystic Fibrosis 10:37–44. doi: 10.1016/j.jcf.2010.09.004

Bergan T, Bøvre K, Hovig B (1970) Priority of Micrococcus mucilaginosus Migula 1900 over Staphylococcus salivarius: Andrewes and Gordon 1907 with proposal of a neotype strain. International Journal of Systematic Bacteriology 20:107–113. doi: 10.1099/00207713-20-1-107

Bergan T, Kocur M (1982) NOTES: Stomatococcus mucilaginosus gen.nov., sp.nov., ep. rev., a member of the family Micrococcaceae. International Journal of Systematic Bacteriology 32:374–377. doi: 10.1099/00207713-32-3-374

Bittar F, Richet H, Dubus J-C, Reynaud-Gaubert M, Stremler N, Sarles J, Raoult D, Rolain J-M (2008) Molecular detection of multiple emerging pathogens in sputa from Cystic Fibrosis patients. PLoS ONE. doi: 10.1371/journal.pone.0002908

Britigan BE, Klapper D, Svendsen T, Cohen MS (1988) Phagocyte-derived lactate stimulates oxygen consumption by Neisseria gonorrhoeae. An unrecognized aspect of the oxygen metabolism of phagocytosis. J Clin Invest 81:318–324.

Brown SA, Whiteley M (2007) A novel exclusion mechanism for carbon resource partitioning in Aggregatibacter actinomycetemcomitans. J Bacteriol 189:6407–6414. doi: 10.1128/JB.00554-07

Carrolo M, Frias MJ, Pinto FR, Melo-Cristino J, Ramirez M (2010) Prophage spontaneous activation promotes DNA release enhancing biofilm formation in

Streptococcus pneumoniae. PLoS ONE 5:e15678. doi: 10.1371/journal.pone.0015678

Collins MD, Hutson RA, B√•verud V, Falsen E (2000) Characterization of a Rothia-like organism from a mouse: description of Rothia nasimurium sp. nov. and reclassification of Stomatococcus mucilaginosus as Rothia mucilaginosa comb. nov. International Journal of Systematic and Evolutionary Microbiology 50:1247–1251.

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044. doi: 10.1371/journal.pone.0011044

De Backer D, Creteur J, Zhang H, Norrenberg M, Vincent J-L (1997) Lactate production by the lungs in acute lung injury. Am J Respir Crit Care Med 156:1099–1104.

Delgado S, Arroyo R, Martín R, Rodríguez JM (2008) PCR-DGGE assessment of the bacterial diversity of breast milk in women with lactational infectious mastitis. BMC Infectious Diseases 8:51. doi: 10.1186/1471-2334-8-51

Dietrich LEP, Teal TK, Price-Whelan A, Newman DK (2008) Redox-active antibiotics control gene expression and community behavior in divergent bacteria. Science 321:1203–1206. doi: 10.1126/science.1160619

Doel JJ, Benjamin N, Hector MP, Rogers M, Allaker RP (2005) Evaluation of bacterial nitrate reduction in the human oral cavity. European Journal of Oral Sciences 113:14–19. doi: 10.1111/j.1600-0722.2004.00184.x

Fodor AA, Klem ER, Gilpin DF, Elborn JS, Boucher RC, Tunney MM, Wolfgang MC (2012) The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. PLoS ONE 7:e45001. doi: 10.1371/journal.pone.0045001

G.H. B (1969) The components of the cell walls and extracellular slime of four strains of Staphylococcus salivarius isolated from human dental plaque. Archives of Oral Biology 14:685–697. doi: 10.1016/0003-9969(69)90190-3

Garvie EI (1980) Bacterial lactate dehydrogenases. Microbiol Rev 44:106–139.

Goddard AF, Staudinger BJ, Dowd SE, Joshi-Datar A, Wolcott RD, Aitken ML, Fligner CL, Singh PK (2012) Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. PNAS 109:13769–13774. doi: 10.1073/pnas.1107435109

Goerke C, Wolz C (2010) Adaptation of Staphylococcus aureus to the cystic fibrosis lung. Int J Med Microbiol 300:520–525. doi: 10.1016/j.ijmm.2010.08.003

Gomes-Filho IS, Passos JS, Seixas da Cruz S (2010) Respiratory disease and the role of oral bacteria. J Oral Microbiol. doi: 10.3402/jom.v2i0.5811

Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. NAR 35:W52–W57. doi: 10.1093/nar/gkm360

Guglielmetti S, Taverniti V, Minuzzo M, Arioli S, Stuknyte M, Karp M, Mora D (2010) Oral bacteria as potential probiotics for the pharyngeal mucosa. Applied and Environmental Microbiology 76:3948–3958. doi: 10.1128/AEM.00109-10

Guss AM, Roeselers G, Newton ILG, Young CR, Klepac-Ceraj V, Lory S, Cavanaugh CM (2011) Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. ISME J 5:20–29.

Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurso FJ, Pace NR (2007) Molecular identification of bacteria in bronchoalveolar lavage fluid from children with Cystic Fibrosis. PNAS 104:20529–20533. doi: 10.1073/pnas.0709804104

Hoffman LR, Richardson AR, Houston LS, Kulasekara HD, Martens-Habbena W, Klausen M, Burns JL, Stahl DA, Hassett DJ, Fang FC, Miller SI (2010) Nutrient availability as a mechanism for selection of antibiotic tolerant Pseudomonas aeruginosa within the CF airway. PLoS Pathog 6:e1000712. doi: 10.1371/journal.ppat.1000712

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590. doi: 10.1126/science.1212665

Jackson AP, Thomas GH, Parkhill J, Thomson NR (2009) Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC Genomics 10:584. doi: 10.1186/1471-2164-10-584

Jiménez E, Delgado S, Fernández L, García N, Albújar M, Gómez A, Rodríguez JM (2008) Assessment of the bacterial diversity of human colostrum and screening of staphylococcal and enterococcal populations for potential virulence factors. Research in Microbiology 159:595–601. doi: 10.1016/j.resmic.2008.09.001

K. Philip, W. Y. Teoh, S. Muniandy, Yaakob H (2009) Pathogenic bacteria predominate in the oral cavity of Malaysian subjects.

Kazor CE, Mitchell PM, Lee AM, Stokes LN, Loesche WJ, Dewhirst FE, Paster BJ (2003) Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. Journal of Clinical Microbiology 41:558–563. doi: 10.1128/JCM.41.2.558-563.2003

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the Cystic Fibrosis gene: Genetic analysis. Science 245:1073–1080. doi: 10.1126/science.2570460

Kung VL, Khare S, Stehlik C, Bacon EM, Hughes AJ, Hauser AR (2012) An rhs gene of Pseudomonas aeruginosa encodes a virulence protein that activates the inflammasome. PNAS 109:1275–1280. doi: 10.1073/pnas.1109285109

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. doi: 10.1186/gb-2004-5-2-r12

Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. PNAS 95:9413–9417.

Lederer F (1974) On the first steps of lactate oxidation by bakers' yeast L-(+)-lactate dehydrogenase (Cytochrome b2). European Journal of Biochemistry 46:393–399. doi: 10.1111/j.1432-1033.1974.tb03632.x

Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Jr MRD, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nature Genetics 43:1275–1280. doi: 10.1038/ng.997

Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. J Cyst Fibros. doi: 10.1016/j.jcf.2012.07.009

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clinical Microbiology Reviews 23:299 –323. doi: 10.1128/CMR.00068-09

Migula W (1900) System der bakterien: bd. Specielle systematik der bakterien. G. Fischer

Murray NE (2000) Type I restriction systems: Sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev 64:412–434.

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6:81–93. doi: 10.1038/ismej.2011.78

Nyvad B, Kilian M (1987) Microbiology of the early colonization of human enamel and root surfaces in vivo. European Journal of Oral Sciences 95:369–380. doi: 10.1111/j.1600-0722.1987.tb01627.x

Oliver A, Cantón R, Campo P, Baquero F, Blázquez J (2000) High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science 288:1251–1253. doi: 10.1126/science.288.5469.1251

Olsen I, Preza D, Aas JA, Paster BJ (2009) Cultivated and not-yet-cultivated bacteria in oral biofilms. Microbial Ecology in Health and Disease 21:65–71. doi: 10.1080/08910600902907509

Ou G, Hedberg M, Horstedt P, Baranov V, Forsberg G, Drobni M, Sandstrom O, Wai SN, Johansson I, Hammarstrom M-L, Hernell O, Hammarstrom S (2009) Proximal small intestinal microbiota and identification of rod-shaped bacteria associated with childhood celiac disease.

Palmer KL, Whiteley M (2011) DMS3-42: the secret to CRISPR-dependent biofilm inhibition in Pseudomonas aeruginosa. J Bacteriol 193:3431–3432. doi: 10.1128/JB.05066-11

Poole SJ, Diner EJ, Aoki SK, Braaten BA, t' Kint de Roodenbeke C, Low DA, Hayes CS (2011) Identification of functional toxin/immunity genes linked to Contact-Dependent Growth Inhibition (CDI) and Rearrangement Hotspot (Rhs) systems. PLoS Genet 7:e1002217. doi: 10.1371/journal.pgen.1002217

Potter JL, Matthews LW, Spector S, Lemm J (1967) Studies on pulmonary secretions. II. Osmolality and the ionic environment of pulmonary secretions from patients with cystic fibrosis, bronchiectasis, and laryngectomy. Am Rev Respir Dis 96:83–87.

Preza D, Olsen I, Willumsen T, Grinde B, Paster BJ (2009) Diversity and site-specificity of the oral microflora in the elderly. European Journal of Clinical Microbiology & Infectious Diseases 28:1033–1040. doi: 10.1007/s10096-009-0743-3

Price-Whelan A, Dietrich LEP, Newman DK (2007) Pyocyanin alters redox homeostasis and carbon flux through central metabolic pathways in Pseudomonas aeruginosa PA14. J Bacteriol 189:6372–6381. doi: 10.1128/JB.00505-07

Ready D, Lancaster H, Qureshi F, Bedi R, Mullany P, Wilson M (2004) Effect of amoxicillin use on oral microbiota in young children. Antimicrobial Agents and Chemotherapy 48:2883–2887. doi: 10.1128/AAC.48.8.2883-2887.2004

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466:334–338. doi: 10.1038/nature09199

Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Bruce KD (2004) Characterization of bacterial community diversity in Cystic Fibrosis lung infections by use of 16S ribosomal DNA terminal restriction fragment length polymorphism profiling. J Clin Microbiol 42:5176–5183. doi: <p>10.1128/JCM.42.11.5176-5183.2004</p>

Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Kehagia V, Connett GJ, Bruce KD (2006) Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with Cystic Fibrosis. J Clin Microbiol 44:2601–2604. doi: <p>10.1128/JCM.02282-05</p>

Sass A, Marchbank A, Tullis E, LiPuma JJ, Mahenthiralingam E (2011) Spontaneous and evolutionary changes in the antibiotic resistance of Burkholderia cenocepacia observed by global gene expression analysis. BMC Genomics 12:373. doi: 10.1186/1471-2164-12-373

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6:e17288. doi: 10.1371/journal.pone.0017288

Schmieder R, Edwards RA MyMGDB. http://edwards.sdsu.edu/cgi-bin/mymgdb.

Schmitz JE, Schuch R, Fischetti VA (2010) Identifying active phage lysins through functional viral metagenomics. Appl Environ Microbiol 76:7181–7187. doi: 10.1128/AEM.00732-10

Smart CHM, Walshaw MJ, Hart CA, Winstanley C (2006) Use of suppression subtractive hybridization to examine the accessory genome of the Liverpool cystic fibrosis epidemic strain of Pseudomonas aeruginosa. J Med Microbiol 55:677–688. doi: 10.1099/jmm.0.46461-0

Smith H, Tang CM, Exley RM (2007) Effect of host lactate on gonococci and meningococci: New concepts on the role of metabolites in pathogenicity. Infect Immun 75:4190–4198. doi: 10.1128/IAI.00117-07

Stackebrandt E (2006) The Genus Stomatococcus: Rothia mucilaginosa, basonym Stomatococcus mucilaginosus. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds) The Prokaryotes. Springer New York, pp 975–982

Tarran R, Grubb BR, Gatzy JT, Davis CW, Boucher RC (2001) The relative roles of passive surface forces and active ion transport in the modulation of airway surface liquid volume and composition. J Gen Physiol 118:223–236. doi: 10.1085/jgp.118.2.223

Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS (2008) Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 177:995–1001. doi: 10.1164/rccm.200708-1151OC

Van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW, Carroll MP, Parkhill J, Bruce KD (2011) Partitioning core and satellite taxa from within Cystic Fibrosis lung bacterial communities. ISME J 5:780–791.

Wang M, Ahrné S, Jeppsson B, Molin G (2005) Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. FEMS Microbiology Ecology 54:219–231. doi: 10.1016/j.femsec.2005.03.012

Wang Y, Wilks JC, Danhorn T, Ramos I, Croal L, Newman DK (2011) Phenazine-1-carboxylic acid promotes bacterial biofilm development via ferrous iron acquisition. J Bacteriol 193:3606–3617. doi: 10.1128/JB.00396-11

Whitchurch CB, Tolker-Nielsen T, Ragas PC, Mattick JS (2002) Extracellular DNA required for bacterial biofilm formation. Science 295:1487. doi: 10.1126/science.295.5559.1487

Willner D, Furlan M (2010) Deciphering the role of phage in the cystic fibrosis airway. Virulence 1:309–313. doi: 10.4161/viru.1.4.12071

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in Cystic Fibrosis and non-Cystic Fibrosis individuals. PLOS ONE 4:e7370. doi: 10.1371/journal.pone.0007370

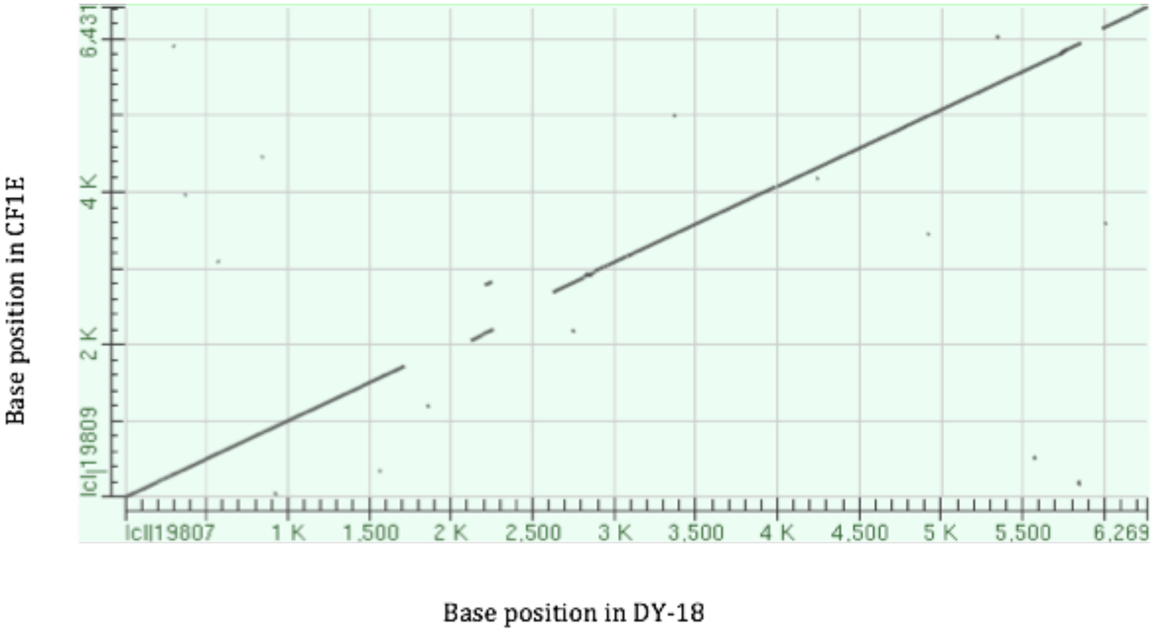
Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, Angly FE, McDole T, Mariella RP, Rohwer F, Haynes M (2011a) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci U S A 108:4547–4553. doi: 10.1073/pnas.1000089107

Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, Lim YW, Rainey PB, Schmieder R, Youle M, Conrad D, Rohwer F (2012) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. Am J Respir Cell Mol Biol 46:127–131. doi: 10.1165/rcmb.2011-0253OC

Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, Rohwer F, Conrad D (2011b) Spatial distribution of microbial communities in the cystic fibrosis lung. The ISME Journal 6:471–474. doi: 10.1038/ismej.2011.104

Yamane K, Nambu T, Yamanaka T, Mashimo C, Sugimori C, Leung K-P, Fukushima H (2010) Complete genome sequence of Rothia mucilaginosa DY-18: A clinical isolate with dense meshwork-like structures from a persistent apical periodontitis lesion. Sequencing. doi: 10.1155/2010/457236

Appendix for Chapter 5



Supplementary Figure 5.1: Dot Plot matrix view of the alignment of CF1E type I restriction modification system (subunit M, R, S) against DY-18.

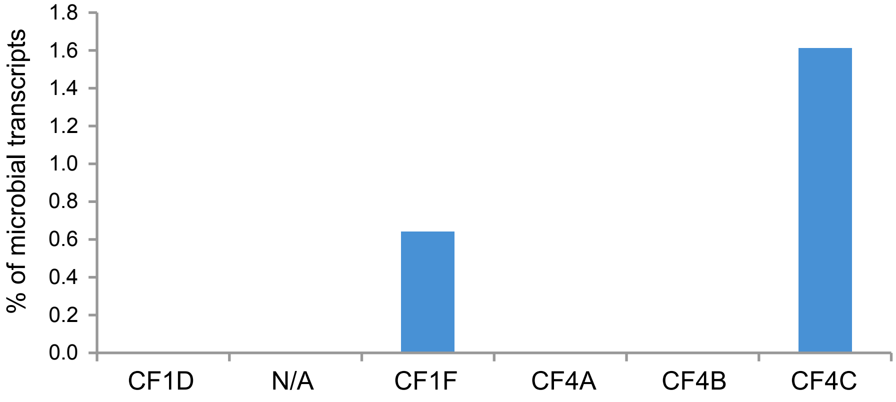
Supporting Information

(S1) Samples information

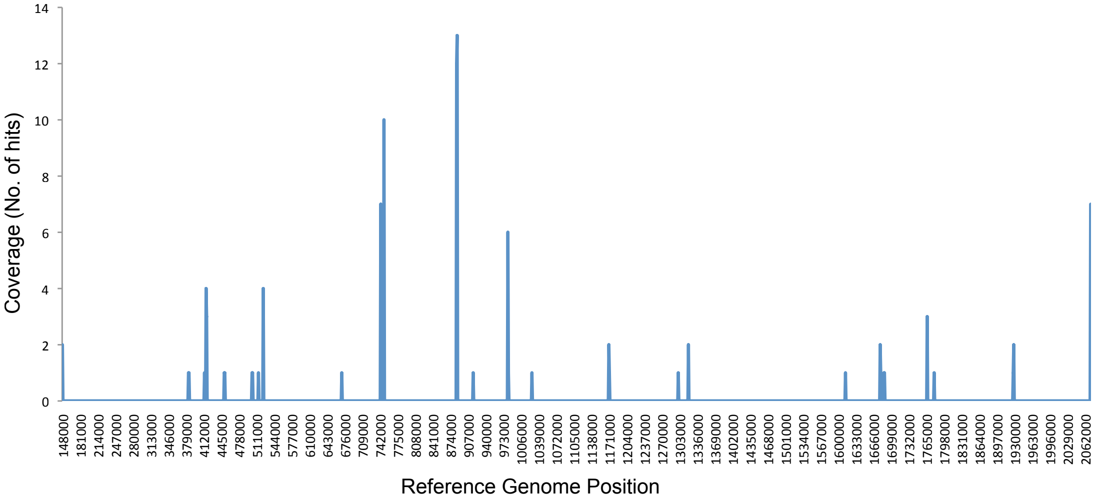
The patients were selected based on the eligibility criteria that include (i) a known clinical diagnosis of CF, (ii) a protocol defined exacerbation that requires intravenous antibiotics, and (iii) a drop in FEV_1 of at least 15% or more compared to the best FEV_1 in the past 12 months. Sputum samples were collected from six CF volunteers (Supplementary Table 5.2) at the Adult CF Clinic (San Diego, CA, United States) by expectoration into a sterile cup except sample CF4-A that was a tracheal aspirate. All collection was in accordance with the University of California Institutional Review Board (HRPP 081500) and San Diego State University Institutional Review Board (SDSU IRB#2121). Clinical status at the time of collection was designated as *exacerbation* (prior to systemic antibiotic treatment), *on treatment* (during systemic antibiotic treatment), *post treatment* (upon completion of systemic antibiotic treatment) or *stable* (when clinically stable and at their clinical and physiological baseline). The samples collected during exacerbation were designated as Day 0 sample.

(S2) Rothia mucilaginosa in community metatranscriptomes

Rothia was detected in 3/5 metatranscriptomes (Supplementary Figure 5.2). Even though *R. mucilaginosa* was present in high abundance in the CF1 samples, the number of metatranscriptomic sequences in these samples was too small for significant detection. Metatranscriptome was not generated for the CF1E sample. On the other hand, CF4C, the only time point from patient CF4 containing *R. mucilaginosa* (~20% in the microbiome), had 120 *R. mucilaginosa* hits to mRNAs (1.6% of microbial transcripts) and 15 hits to rRNAs (0.2% of microbial transcripts) in its metatranscriptome (Supplementary Figure 5.16). The transcripts were scattered randomly across the genome (Supplementary Figure 5.3) starting from position 148,000 (no coverage was detected before this position).



Supplementary Figure 5.2: The prevalence of *R. mucilaginosa* in the corresponding community transcriptomes.



Supplementary Figure 5.3: Coverage of 120 CF4C metatranscriptomic mRNA reads on the reference genome *R. mucilaginosa* DY-18 starting from position 148,000 bp.

(S3) Characteristics of CF1E

(i) Defining Gaps

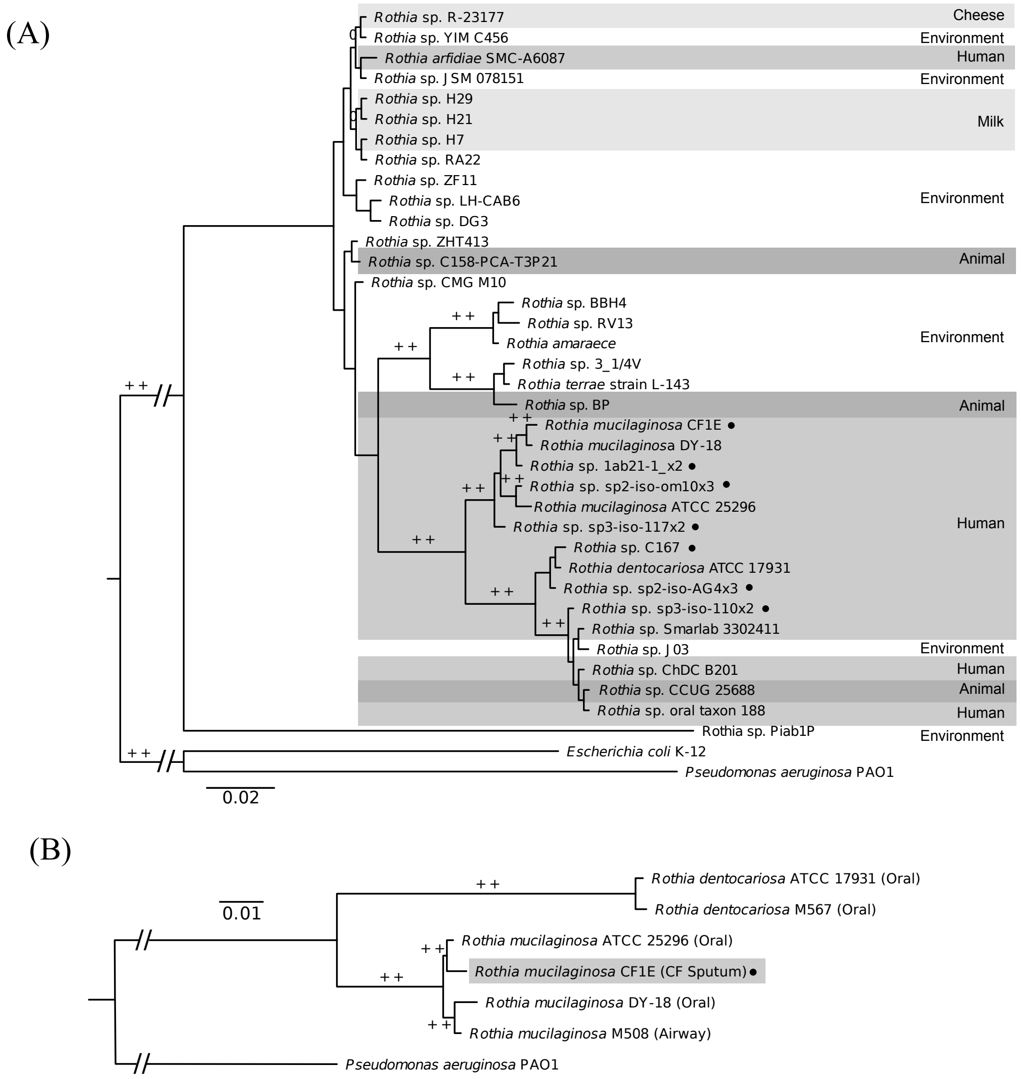
Gaps in the genome scaffold are defined as regions present in the reference DY-18 genome that were not covered by contigs in the final scaffold. The size of each gap was based on the difference in scaffolding coordinates at the end of the previous contig and the start of the next contig. For building the scaffold, gapped regions were filled with the ambiguous base N. There were only four gaps spanning more than 1 kbp, and in the reference genome these regions contained genes with unknown functions (i.e., hypothetical proteins). The genes present in the reference genome that correspond to the missing regions in CF1E are listed in Supplementary Table 5.15.

(ii) Phylogenetic analysis of *Rothia* spp.

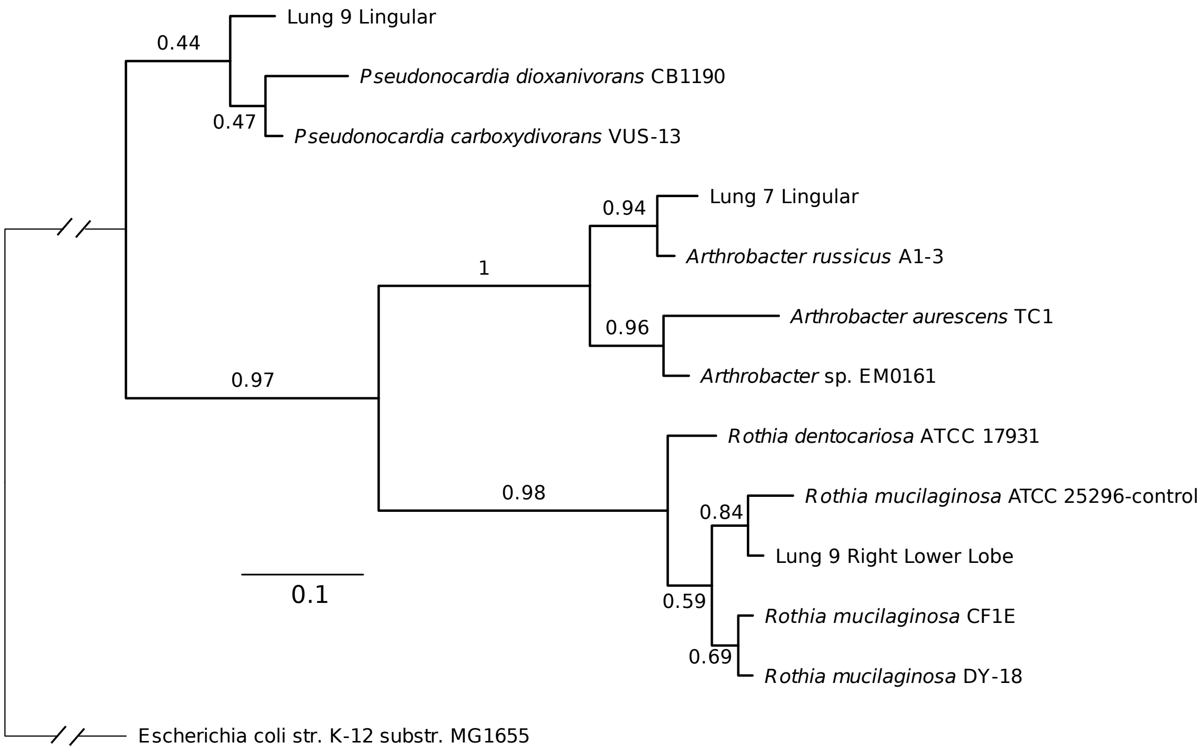
All 16S rRNA gene sequences from *Rothia* spp (n=36) were identified and retrieved from the NCBI nucleotide database (Supplementary Table 5.6) and aligned with a full-length 16S rRNA gene from the CF1E genome scaffold. The phylogenetic analysis of these 16S rRNA gene sequences revealed four distinct groups of human *Rothia*-isolates that were strongly supported in both Bayesian inferences (≥95% Bayesian posterior probability) and maximum likelihood-based resampling (bootstrap of ≥70%) (Supplementary Figure 5.4A). The closest neighbor of *R. mucilaginosa* CF1E was DY-18, followed by isolates of *Rothia* sp. from CF patients. In addition, the tree showed that other CF *Rothia* spp. isolates belong to either the *R. mucilaginosa* group or *R. dentocariosa* group. The two *Rothia* spp. that clustered with the human group were isolated from mouse nose (*Rothia* sp. CCUG 25688) and herbs in tumulus (*Rothia* sp. J03).

A phylogenetic analysis of *Rothia* species based on the RecA, RpoA, and Inf2 amino acid sequences placed CF1E closest to the ATCC strain 25296 instead of DY-18 (Supplementary Figure 5.4B). However, CF1E was placed in the *R. mucilaginosa* group with strong support according to both Bayesian posterior probability and maximum likelihood bootstrapping.

Methods used for Phylogenetic analysis: 16S rRNA sequences of the genus *Rothia* were downloaded from GenBank and compared to the 16S fragment from assembled CF1E contig 0173. RecA, RpoA, and Inf2 protein sequences of *R. mucilaginosa* and *R. dentocariosa* were downloaded from GenBank (accession numbers in Supplementary Table 5.13 and 5.14). Each gene was aligned separately using MUSCLE (version 3.8.31) (Edgar 2004) and the alignment was trimmed using trimAI (version v1.4.rev7) (Capella-Gutiérrez et al. 2009). The three separate gene alignments were then combined into a single alignment. For each dataset (16S rRNA and protein-coding genes), a phylogeny was estimated using MrBayes (version 3.1.2) (Ronquist and Huelsenbeck 2003) by four independent runs with the GTR+I+Γ model of evolution, sampling every 100 generations for 10 X 10^6 generations. Support for nodes was assessed with maximum likelihood bootstrapping as implemented in RAxML (version 7.2.6) (Stamatakis 2006) with the GTRGAMMAI model for 16S rRNA sequences and the PROTGAMMADAYHOFF model for amino acid sequences, both using default parameters.


(A)
Rothia sp. R-23177
Cheese
Rothia sp. YIM C456
Environment
Rothia arfidiae SMC-A6087
Human
Rothia sp. JSM 078151
Environment
Rothia sp. H29
Rothia sp. H21
Milk
Rothia sp. H7
Rothia sp. RA22
Rothia sp. ZF11
Rothia sp. LH-CAB6
Environment
Rothia sp. DG3
Rothia sp. ZHT413
Rothia sp. C158-PCA-T3P21
Animal
Rothia sp. CMG M10
Rothia sp. BBH4
Rothia sp. RV13
Environment
Rothia amaraece
Rothia sp. 3_1/4V
Rothia terrae strain L-143
Rothia sp. BP
Animal
Rothia mucilaginosa CF1E ●
Rothia mucilaginosa DY-18
Rothia sp. 1ab21-1_x2 ●
Rothia sp. sp2-iso-om10x3 ●
Rothia mucilaginosa ATCC 25296
Human
Rothia sp. sp3-iso-117x2 ●
Rothia sp. C167 ●
Rothia dentocariosa ATCC 17931
Rothia sp. sp2-iso-AG4x3 ●
Rothia sp. sp3-iso-110x2 ●
Rothia sp. Smarlab 3302411
Rothia sp. J03
Environment
Rothia sp. ChDC B201
Human
Rothia sp. CCUG 25688
Animal
Rothia sp. oral taxon 188
Human
Rothia sp. Piab1P
Environment
Escherichia coli K-12
Pseudomonas aeruginosa PAO1
++
0.02
(B)
0.01
Rothia dentocariosa ATCC 17931 (Oral)
Rothia dentocariosa M567 (Oral)
Rothia mucilaginosa ATCC 25296 (Oral)
Rothia mucilaginosa CF1E (CF Sputum) ●
Rothia mucilaginosa DY-18 (Oral)
Rothia mucilaginosa M508 (Airway)
Pseudomonas aeruginosa PAO1
++

Supplementary Figure 5.4: Phylogeny of *Rothia* species, a majority-rule consensus phylogram computed from the set of 36,000 credible trees. (A) Phylogeny based on the 16S rDNA. A "+" above a node indicates a branch supported by ≥95% Bayesian posterior probability and a "++" indicates additional maximum likelihood bootstrap of ≥70%. Branches with double diagonal lines have been shortened 8-fold to aid viewing. (B) Multilocus phylogeny based on the RecA, RpoA, and Inf2 protein sequences. A "++" above a node indicates a branch supported by ≥95% Bayesian posterior probability and maximum likelihood bootstrap of ≥85%. Branches with double diagonal lines have been shortened by 10.4 times to aid viewing. Detailed accession number, source of isolation and reference for each sequence are shown in Supplementary Table 13 and 14.


0.44
Lung 9 Lingular
Pseudonocardia dioxanivorans CB1190
0.47
Pseudonocardia carboxydivorans VUS-13
0.94
Lung 7 Lingular
1
Arthrobacter russicus A1-3
0.96
Arthrobacter aurescens TC1
0.97
Arthrobacter sp. EM0161
Rothia dentocariosa ATCC 17931
0.98
0.84
Rothia mucilaginosa ATCC 25296-control
Lung 9 Right Lower Lobe
0.1
0.59
Rothia mucilaginosa CF1E
0.69
Rothia mucilaginosa DY-18
Escherichia coli str. K-12 substr. MG1655

Supplementary Figure 5.5: Majority-rule consensus phylogram of sequenced partial Actinobacteria-specific 16S PCR fragments from lung sections. The branch support is indicated by Bayesian posterior probability. Lung 7 corresponds to patient CF7 who microbiomes were generated before the lung transplant, and the Lung 9 patient was not used in our metagenomic analysis. Branches with double diagonal lines have been shortened by ~3-fold to aid viewing.

Supplementary Table 5.1: Diseases associated with *R. mucilaginosa.*

Disease	**Reference(s)**
Oral squamous cell carcinoma	(Pushalkar et al. 2011)
Periodontitis	(Siqueira Jr et al. 2007)
Branchial pouch anomalies	(Pahlavan et al. 2010)
Upper respiratory tract infection in sickle cell disease	(Rogovik et al. 2010)
Pneumonia	(Ko et al. 2009; Fusconi et al. 2009)
Endocarditis	(Pinsky et al. 1989)
Pericoronitis	(Peltroche-Llacsahttanga et al. 2000)
Psoriasis	(Gao et al. 2008)
Peritonitis	(Hodzic and Snyder 2010)
Bacteremia	(Kaufhold et al. 1992; Vaccher et al. 2007; Ohno et al. 2010)
Bacterial meningitis	(Lee et al. 2008)
Arthritis	(Kaasch et al. 2010)
Central nervous system infections	(Pulzova et al. 2009; Blouin et al. 2010)
Granulomatous dermatitis	(Morgan et al. 2010)

Supplementary Table 5.2: Microbiomes used in this study. Clinical status was designated as *exacerbation* (prior to systemic antibiotic treatment), *on treatment* (during systemic antibiotic treatment), *post treatment* (upon completion of systemic antibiotic treatment) or *stable* (when clinically stable and at their clinical and physiological baseline). The samples collected during exacerbation were designated as Day 0 sample, and the times between samples are cumulatively calculated from Day 0.

Patient ID (Gender, Age)	**Time Point**	**Time Line (Day)**	**Health Status**
CF1 (Male, 38)	D	0	Exacerbation
	E	14	On Treatment
	F	33	Post Treatment
CF4 (Male, N/A)	A	0	Exacerbation
	B	11	Post Treatment
	C	58	Stable
CF5 (Female, N/A)	A	0	Exacerbation
	B	21	Post Treatment
CF6 (Female, 39)	A	0	Exacerbation
	B	12	On Treatment
	C	17	Post Treatment
	D	46	Stable
CF7 (Male, 36)	A	0	Exacerbation
	B	20	On Treatment
	C	27	On Treatment
	D	37	Post Treatment
CF8 (Male, 26)	A	0	Exacerbation
	B	17	Post Treatment

Supplementary Table 5.3A: Annotation of the gaps ≥5 kbp in CF1 metagenomic reference mapping against *R. mucilaginosa* DY-18. Refer to Supplementary Table 5.2 for detailed patient samples information.

CF1	CF1-D-Ex (Day 0)		CF1-F-Pt (Day 33)		Genes
	Start	Stop	Start	Stop	
*	116,888	127,413	118,226	127,208	Seryl-tRNA synthetase (EC 6.1.1.11) CDS; Conserved protein with diacylglycerol kinase catalytic domain CDS; FIG01029139: hypothetical protein CDS; FIG01029400: hypothetical protein CDS; conserved hypothetical protein, putative cell filamentation protein CDS; putative cell filamentation protein CDS; hypothetical protein (6)
	191,241	198,554	192,364	198,337	FIG01028594: hypothetical protein CDS; FIG01029378: hypothetical protein CDS; FIG01028871: hypothetical protein CDS (2); hypothetical protein 2)
**	317,814	325,245	319,381	325,118	4 hypothetical proteins + ATP-binding protein of ABC transporter
	1,044,230	1,052,204	1,045,478	1,052,421	GTP pyrophosphokinase (EC 2.7.6.5), (p)ppGpp synthetase I CDS; Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) CDS; Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) CDS; Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) CDS
**	1,214,127	1,222,178	1,214,321	1,221,292	FIG01293855: hypothetical protein CDS; hypothetical protein (2)
	1,369,419	1,382,713	1,369,198	1,377,838	FIG01028573: hypothetical protein CDS; FIG01028929: hypothetical protein CDS; FIG01029243: hypothetical protein CDS; FIG01029252: hypothetical protein CDS; CRISPR-associated protein Cas1 CDS
*	1,784,753	1,798,606	1,786,942	1,798,038	FIG01028641: hypothetical protein CDS; Acetyltransferase (EC 2.3.1.-) CDS; FIG01028952: hypothetical protein CDS; glycosyl transferase, family 2 CDS; FIG01028912: hypothetical protein CDS; protein tyrosine phosphatase CDS; Ribose-phosphate pyrophosphokinase (EC 2.7.6.1) CDS; N-acetylglucosamine-1-phosphate uridyltransferase (EC 2.7.7.23) / Glucosamine-1-phosphate N-acetyltransferase (EC 2.3.1.157) CDS; ABC transporter, ATP-binding protein CDS; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC 2.7.1.148) CDS
**	1,893,621	1,899,157	1,893,621	1,898,769	hypothetical protein (3); protein tyrosine phosphatase CDS
**	2,084,616	2,090,484	2,084,469	2,090,539	Amino acid ABC transporter, periplasmic amino acid-binding protein CDS; Cystathionine gamma-lyase (EC 4.4.1.1) CDS; O-acetylhomoserine sulfhydrylase (EC 2.5.1.49) CDS; hypothetical protein CDS

* Gaps present in CF1E

** Gaps present in CF1E with sizes slightly smaller than 5 kbp and therefore not shown in the line plots (Figure 5.2 and Supplementary Figure 5.1).

Supplementary Table 5.3B: Annotation of the gaps ≥5 kbp in CF6 metagenomic reference mapping against *R. mucilaginosa* DY-18. Refer to Supplementary Table 5.2 for detailed patient samples information.

CF6	CF6-B-Tr (Day 12)		CF6-C-Pt (Day 17)		Genes
	Start	Stop	Start	Stop	
	293,008	301,901	293,008	299,553	NAD-dependent protein deacetylase of SIR2 family CDS; FIG01029042: hypothetical protein CDS (2); FIG01029042: hypothetical protein CDS; FIG01028880: hypothetical protein CDS; FIG01028880: hypothetical protein CDS
	1,045,416	1,056,255	1,045,555	1,056,118	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) CDS; Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) CDS; Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) CDS; hypothetical protein (2)
	1,212,461	1,225,388	1,216,259	1,225,272	FIG01029335: hypothetical protein CDS; Triosephosphate isomerase (EC 5.3.1.1) CDS; FIG01293855: hypothetical protein CDS; hypothetical protein (2)
	1,236,551	1,241,615	1,236,259	1,242,067	FIG01293855: hypothetical protein CDS; FIG01293855: hypothetical protein CDS
	1,480,952	1,493,792	1,483,197	1,491,820	FIG01293855: hypothetical protein CDS; FIG01293855: hypothetical protein CDS; Glutamate-ammonia-ligase adenylyltransferase (EC 2.7.7.42) CDS; Glutamine synthetase type I (EC 6.3.1.2) CDS
	1,893,606	1,898,780	1,893,621	1,898,780	hypothetical protein (2)
	1,952,461	1,959,004	1,952,580	1,959,091	FIG01029071: hypothetical protein CDS; Cystathionine gamma-synthase (EC 2.5.1.48) CDS; Thymidylate kinase (EC 2.7.4.9) CDS; Thymidylate kinase (EC 2.7.4.9) CDS; Thymidylate kinase (EC 2.7.4.9) CDS; ACT domain protein CDS; Mannose-6-phosphate isomerase (EC 5.3.1.8) CDS
	2,160,027	2,187,871	2,160,027	2,187,948	Cell division protein FtsK CDS; Cell division protein FtsK CDS; FIG01029055: hypothetical protein CDS; hypothetical protein (14)

Supplementary Table 5.3C: Annotation of the gaps ≥5 kbp in CF7 metagenomic reference mapping against *R. mucilaginosa* DY-18. Refer to Supplementary Table 5.22 for detailed patient samples information.

CF7	CF7-A-Ex (Day 0)		CF7-B-Tr (Day 20)		Genes
	Start	Stop	Start	Stop	
	119,183	130,754	119,330	130,480	Mainly non-coding region + 1 hypothetical protein
	317,552	324,682	319,276	325,352	non-coding region, deleted from the cf1e scaffold too
	815,783	821,768	812,145	818,057	putative two-component system response regulator CDS; FIG01029073: hypothetical protein CDS; FIG01029073: hypothetical protein CDS; FIG01029073: hypothetical protein CDS
	1,046,072	1,058,265	1,044,326	1,059,604	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) CDS; Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) CDS; Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) CDS; hypothetical protein (4)
	1,213,573	1,227,654	1,216,352	1,221,526	non-coding region + half of hypothetical protein
	1,366,206	1,382,817	1,366,502	1,382,793	FIG01029408: hypothetical protein CDS; FIG01029391: hypothetical protein CDS; FIG01028573: hypothetical protein CDS; FIG01028929: hypothetical protein CDS; FIG01029243: hypothetical protein CDS; FIG01029252: hypothetical protein CDS; CRISPR-associated protein Cas1 CDS
	1,786,873	1,797,870	1,786,975	1,792,449	glycosyl transferase, family 2 CDS; FIG01028912: hypothetical protein CDS; protein tyrosine phosphatase CDS; Ribose-phosphate pyrophosphokinase (EC 2.7.6.1) CDS
			1,792,870	1,799,197	N-acetylglucosamine-1-phosphate uridyltransferase (EC 2.7.7.23) / Glucosamine-1-phosphate N-acetyltransferase (EC 2.3.1.157) CDS; ABC transporter, ATP-binding protein CDS; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC 2.7.1.148) CDS; Dimethyladenosine transferase (EC 2.1.1.-) CDS
	1,888,835	1,898,868	1,892,504	1,898,840	FIG01028568: hypothetical protein CDS; hypothetical protein (2)
	1,971,322	1,988,862	1,972,818	1,982,106	Probable ATP-dependent helicase lhr (EC 3.6.1.-) CDS; FIG01028989: hypothetical protein CDS; Ribosomal large subunit pseudouridine synthase A (EC 4.2.1.70) CDS; putative glycosyl hydrolase CDS; putative glycosyl hydrolase CDS
	2,160,027	2,165,406	2,160,373	2,166,081	hypothetical protein (7)

Supplementary Table 5.3D: Annotation of the gaps ≥5 kbp in CF8 metagenomic reference mapping against *R. mucilaginosa* DY-18. Refer to Supplementary Table 5.2 for detailed patient samples information.

CF8	CF8-A-Ex (Day 0)		CF8-B-Pt (Day 17)		Genes
	Start	Stop	Start	Stop	
	118,031	124,798	120,245	125,303	Non-coding region
	1,048,704	1,054,082			subunit R only that's missing
			1,045,578	1,054,259	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) CDS; Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) CDS; Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) CDS; hypothetical protein (4)
	1,339,023	1,346,126			predicted nucleic acid-binding protein CDS; predicted nucleic acid-binding protein CDS; FIG01029207: hypothetical protein CDS;
			1,341,734	1,348,705	FIG01029207: hypothetical protein CDS; gluconolactonase CDS; protein of unknown function DUF34 CDS; Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11) CDS
	1,365,871	1,382,363	1,366,786	1,383,290	FIG01029408: hypothetical protein CDS; FIG01029391: hypothetical protein CDS; FIG01028573: hypothetical protein CDS; FIG01028929: hypothetical protein CDS; FIG01029243: hypothetical protein CDS; FIG01029252: hypothetical protein CDS; CRISPR-associated protein Cas1 CDS
	1,701,452	1,706,504	1,701,544	1,706,768	ABC transporter related CDS; ABC transporter related CDS; ABC transporter related CDS; ABC transporter related CDS; Protease II (EC 3.4.21.83) CDS
	1,893,711	1,898,815	1,893,481	1,898,769	FIG01028568: hypothetical protein CDS; protein tyrosine phosphatase CDS; protein tyrosine phosphatase CDS; hypothetical protein (2)
	1,971,500	1,981,982	1,975,975	1,981,085	Probable ATP-dependent helicase lhr (EC 3.6.1.-) CDS; FIG01028989: hypothetical protein CDS; Ribosomal large subunit pseudouridine synthase A (EC 4.2.1.70) CDS; putative glycosyl hydrolase CDS

Supplementary Table 5.4: Statistics from BWA mapping of metagenomic reads against the reference genome *R. mucilaginosa* DY-18.

Metagenome	Number of Mapped Reads	Maximum Coverage	Mean Coverage	Number of Bases (% bases of reference genome)
CF1-D	4,772	12	0.9	1,239,444 (55%)
CF1-E	41,835	47	7.8	2,117,375 (93%)
CF1-F	12,512	17	2.4	1,892,620 (84%)
CF4-C	8,886	23	1.6	1,621,081 (72%)
CF6-A	355	6	0.06	131,287 (6%)
CF6-B	9,925	18	1.8	1,683,390 (74%)
CF6-C	10,790	20	1.9	1,728,332 (76%)
CF6-D	800	8	0.15	285,337 (13%)
CF7-A	5,088	16	0.9	1,220,642 (54%)
CF7-B	9,588	16	1.7	1,659,075 (73%)
CF7-C	133	7	0.02	42,330 (2%)
CF7-D	1,246	8	0.2	416,645 (18%)
CF8-A	10,225	50	1.8	1,6,45,293 (73%)
CF8-B	11,236	16	2.0	1,783,108 (79%)

Supplementary Table 5.5: General features of the CF1E *R. mucilaginosa* scaffold, DY-18 reference genome, and M508 draft genome.

Rothia mucilaginosa	**CF1E**	**DY-18 (reference)**	**M508***
Genome Size	2,278,618 bp	2,264,603 bp	2,313,271
GC content	59.6%	59.6%	59.6%
Predicted gene products**	1,739	1,739	1,790
Total subsystems** **(%)**	248 (42%)	254 (42%)	251 (41%)

* Draft genome in supercontigs format from the Genomes Online Database (GOLD).

** Numbers obtained from the RAST-annotation server

Supplementary Table 5.6: Subsystem feature counts of *R. mucilaginosa* CF1E, DY-18, and M508.

Subsystem features	**CF1E**	**DY-18**	**M508**
Amino Acids and Derivatives	153	162	148
Protein Metabolism	142	142	141
Cofactors, Vitamins, Prosthetic Groups, Pigments	136	130	128
Carbohydrates	101	101	107
RNA Metabolism	90	89	93
DNA Metabolism	63	81	79
Nucleosides and Nucleotides	59	60	60
Cell Wall and Capsule	58	60	57
Stress Response	50	50	50
Fatty Acids, Lipids, and Isoprenoids	29	29	29
Virulence, Disease and Defense	25	22	24
Respiration	22	21	22
Membrane Transport	18	18	17
Cell Division and Cell Cycle	13	14	14
Phosphorus Metabolism	13	13	13
Regulation and Cell signaling	10	10	10
Nitrogen Metabolism	9	9	9
Potassium metabolism	9	9	9
Iron acquisition and metabolism	7	7	4
Sulfur Metabolism	5	5	6
Dormancy and Sporulation	1	1	1
Metabolism of Aromatic Compounds	1	1	1
Miscellaneous	50	50	49

Subsystem features that do not have any count: photosynthesis; Phages, Prophages, Transposable elements, Plasmids; motility and chemotaxis; secondary metabolism.

Supplementary Table 5.7: Sequence identities of the genes encoding the Type I restriction modification system in CF1E and DY-18, determined using BLAST. The identity value is subjected to >97% query length coverage.

	Nucleotide level Identity against DY-18 (%)	**Protein level Identity against DY-18 (%)**
Subunit R	81	75
Subunit M	87	88
Subunit S	37	41

Supplementary Table 5.8: CRISPR positions in the CF1E genome scaffold.

start	**end**	**length**	**DR length**	**No. of Spacers**	**Blastx against NR**
1,389,037	1,389,864	827	36	11	hypothetical protein from *Caenorhabditis remanei*
1,392,263	1,393,579	1,316	36	17	no hit
1,394,037	1,395,297	1,260	36	17	hypothetical protein from *Propionibacterium freudenreichii* subsp. Shermanii CIRM-BIA1
1,395,433	1,395,686	253	36	3	no hit

Supplementary Table 5.9: Identification of the spacer sequences in CF1E CRISPR structure from human- and environmental-viral metagenomes at 100% length coverage and ≥90% identity (≤2 mismatches).

Virome (Habitat)	**Spacer sequence**	**No. of hits**
(Willner et al. 2011) **(Oral)** 5b11623defc42938f41589107350896f_186762_266 5b11623defc42938f41589107350896f_176283_256 5b11623defc42938f41589107350896f_176158_244 5b11623defc42938f41589107350896f_175980_260 5b11623defc42938f41589107350896f_175807_233 5b11623defc42938f41589107350896f_173039_268 5b11623defc42938f41589107350896f_159634_255 5b11623defc42938f41589107350896f_159104_258 5b11623defc42938f41589107350896f_158624_261 5b11623defc42938f41589107350896f_118402_230 5b11623defc42938f41589107350896f_84294_248 5b11623defc42938f41589107350896f_70709_261 5b11623defc42938f41589107350896f_60807_263 5b11623defc42938f41589107350896f_60398_267 5b11623defc42938f41589107350896f_58332_260 5b11623defc42938f41589107350896f_48083_250 5b11623defc42938f41589107350896f_37476_262 5b11623defc42938f41589107350896f_33576_249 5b11623defc42938f41589107350896f_18007_255	CAACGATTCCCACGCGGCGCGCCCAGTCTCCGTCTGA	**19**
(Willner et al. 2011) **(Oral)** f44d959b723905a049b0334f19668e5c_207520_157 f44d959b723905a049b0334f19668e5c_201888_260 f44d959b723905a049b0334f19668e5c_110173_251 f44d959b723905a049b0334f19668e5c_212940_122	CAACGATTCCCACGCGGCGCGCCCAGTCTCCGTCTGA	**4**

Supplementary Table 5.10: A comparison of putative adaptations and predicted metabolisms of *R. mucilaginosa* and *P. aeruginosa* that are hypothesized to enable persistence in the CF lung, based on literature and genomic data.

CF-lung adapted phenotype	***Pseudomonas aeruginosa ****	***Rothia mucilaginosa***
Respiration	Undergoes aerobic, microaerobic, & anaerobic respiration Increasing denitrification Fermentation	Aerobic Microaerobic Anaerobic (very slow growing) Reduces nitrate Fermentation
Food source	Free amino acids (Prefers L-alanine, L-arginine, L-glutamate) Lactate Pyruvate Arginine	Free amino acids Sucrose Fructose Lactate Glycerol & glycerol-3-phosphate Pyruvate
Motility	Uses flagella but this is lost when *P. aeruginosa* adapts towards persistence	Not known
Mucoidy / Biofilm	Conversion to mucoidy (overproduction of alginate) in persistent infection Biofilm production is alginate – dependent	Organism contains mucoid capsule Biofilm production is mannose-dependent
Signaling and communication systems	Loss of quorum sensing molecules e.g. AHLs due to ΔLasR *Rhs* may be responsible for the communication between self, competitors & host?	No known quorum sensing molecules *Rhs* may be responsible for the *communication* between self, competitors & host?
Type III secretion	Down-regulated or lost to reduce virulence	Not detected in the genome
Other virulence factors	Loss of virulence including secretion of elastase and exotoxin	-
Siderophores	Pyoverdine and pyochelin	Not detected in the genome
Iron acquisition	Multiple mechanisms; independent systems for heme, Fe3+ and Fe2+ uptake and use	Heme, hemin uptake and utilization systems in Gram Positives; Low-pH induced ferrous iron transporter
Antibiotic resistance mechanisms	β-lactamases Mex – multidrug efflux pumps Loss of OprD (Imipenem resistance) Mutation – induced	Modulator of Drug Activity (MdaB) Macrolide export ATP-binding/permease protein MacB EmrB/QacA subfamily drug resistance transporter Bcr/CflA family multidrug resistance transporter
Prophages	Detected in the genome	Not detected in the genome

* Mainly extracted from (Hogardt and Heesemann 2010)

Supplementary Table 5.11: Genes that are missing from the CF1E genome scaffold but present in the DY-18 reference. Genes are considered missing when the gap is within a contig.

Name	Length	Start	End	Region Size
predicted ARSR subfamily of helix-turn-helix bacterial transcription regulatory protein CDS	396	43502	44199	697
conserved hypothetical protein, putative cell filamentation protein CDS	693	119455	127207	7752
FIG01029139: hypothetical protein CDS	687	119455	127207	7752
FIG01029400: hypothetical protein CDS	756	119455	127207	7752
hypothetical protein CDS	210	119455	127207	7752
hypothetical protein CDS	156	119455	127207	7752
hypothetical protein CDS	210	119455	127207	7752
hypothetical protein CDS	189	119455	127207	7752
hypothetical protein CDS	561	119455	127207	7752
hypothetical protein CDS	285	119455	127207	7752
putative cell filamentation protein CDS	636	119455	127207	7752
exopolysaccharide biosynthesis protein related to N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase CDS	1317	138835	139305	470
predicted nucleic acid-binding protein CDS	489	138835	139305	470
UBA/THIF-type NAD/FAD binding fold CDS	1071	149644	150934	1290
hypothetical protein CDS	1179	260142	261857	1715
hypothetical protein CDS	117	260142	261857	1715
hypothetical protein CDS	1314	283582	285106	1524
ATP-binding protein of ABC transporter CDS	777	319311	324416	5105
FIG01028874: hypothetical protein CDS	1215	319311	324416	5105
FIG01029167: hypothetical protein CDS	792	319311	324416	5105
hypothetical protein CDS	333	319311	324416	5105
hypothetical protein CDS	903	319311	324416	5105
ABC-type amino acid transport system, permease component CDS	684	620953	624517	3564
COG1126: ABC-type polar amino acid transport system, ATPase component CDS	765	620953	624517	3564
Glutamate binding protein CDS	831	620953	624517	3564
putative glutamate transporter permease protein CDS	939	620953	624517	3564
FIG01029195: hypothetical protein CDS	1194	656126	657389	1263
hypothetical protein CDS	1791	1237261	1239756	2495
putative integral membrane protein CDS	687	1410305	1411220	915
Pyruvate oxidase [ubiquinone, cytochrome] (EC 1.2.2.2) CDS	1752	1817923	1820007	2084
cell wall surface anchor family protein CDS	735	1924698	1925692	994
Amino acid ABC transporter, periplasmic amino acid-binding protein CDS	846	2084618	2090544	5926
Cystathionine gamma-lyase (EC 4.4.1.1) CDS	1161	2084618	2090544	5926
hypothetical protein CDS	1029	2084618	2090544	5926
O-acetylhomoserine sulfhydrylase (EC 2.5.1.49) CDS	1152	2084618	2090544	5926
FIG01029286: hypothetical protein CDS	981	2105030	2107749	2719
FIG01029131: hypothetical protein CDS	1476	2209635	2211353	1718

Supplementary Table 5.12: Genes present in the CF1E genome scaffold but missing in the reference genome DY-18.

Name	Insertion site	region size
hypothetical protein CDS Phage lysin, N-acetylmuramoyl-L-alanine amidase CDS	53748	3174
Modulator of drug activity B CDS	65967	913
Putative DNA-binding protein CDS	70542	1534
hypothetical protein CDS	92872	1160
hypothetical protein CDS	111545	710
hypothetical protein CDS	119455	1338
predicted nucleic acid-binding protein CDS hypothetical protein CDS	156926	1615
putative hydrolase CDS	239949	1690
hypothetical protein CDS	374555	1065
acyltransferase 3 CDS	613830	2685
hypothetical protein CDS	625770	945
hypothetical protein CDS	686303	719
Mobile element protein CDS	808,821	318
hypothetical protein CDS DNA-cytosine methyltransferase (EC 2.1.1.37) CDS hypothetical protein CDS	992646	3146
2-oxoglutarate/malate translocator CDS	1013016	1866
putative helicase CDS putative helicase CDS hypothetical protein CDS	1084123	5306
FIG01029049: hypothetical protein CDS hypothetical protein CDS	1383788	1489
hypothetical protein CDS	1392429	754
PE-PGRS FAMILY PROTEIN CDS	1434185	4251
hypothetical protein CDS	1464431	1830
Mercuric ion reductase (EC 1.16.1.1) CDS	1671908	1271
Macrolide export ATP-binding/permease protein MacB (EC 3.6.3.-) CDS hypothetical protein CDS	1671564	3475
hypothetical protein CDS hypothetical protein CDS	1788171	2192
hypothetical protein CDS hypothetical protein CDS	1790071	2205
hypothetical protein CDS	1817923	1800
Mobile element protein CDS	1884458	1840
L-lactate dehydrogenase (EC 1.1.2.3) CDS	2044067	1547
Flagellar hook-length control protein fliK CDS	1924657	966
hypothetical protein CDS	2000796	1310

Supplementary Table 5.13: Isolation source and references of sequences extracted and used in the 16S phylogenetic analysis.

Category	Organism (Accession)	Isolation source	Underlying disease	Reference
Human	*Rothia arfidiae* SMC-A6087 (DQ673322)	Blood	Pneumonia	Ko KS et al. (2009)
	Rothia mucilaginosa CF1E	Sputum from CF	Cystic Fibrosis	This Paper (2012)
	Rothia mucilaginosa DY-18 (NC_013715)	Oral	Persistent apical periodontitis lesion	Yamane K. et al. (2010)
	Rothia sp. Lab 21-1_x2 (GQ900877)	Sputum from CF	Cystic Fibrosis	Guss AM. et al. (2011)
	Rothia sp. sp2-iso-om10x3 (GQ900840)	Sputum from CF	Cystic Fibrosis	Guss AM. et al. (2011)
	Rothia mucilaginosa ATCC 25296	Oral	-	HMP unpublished
	Rothia sp. sp3-iso-117x2 (GQ900845)	Sputum from CF	Cystic Fibrosis	Guss AM. et al. (2011)
	Rothia sp. ChDC B201 (AF543279)	Oral	-	unpublished
	Rothia sp. Smarlab 3302411 (AY538697)	Bronchic expectoration	-	unpublished
	Rothia sp. Sp3-iso-110x2 (GQ900846)	Sputum from CF		Guss AM. et al. (2011)
	Rothia sp. Oral taxon 188 (GU470892)	Oral	-	Dewhirst FE. (2010)
	Rothia sp. sp2-iso-AG4x3 (GQ900837)	Sputum from CF	-	Guss AM. et al. (2011)
	Rothia dentocariosa ATCC 17931 (CP002280)	Oral	-	HMP unpublished
Cheese	*Rothia sp.* R-23177 (AJ969174)	Smear-ripened cheese	-	unpublished
Animal	*Rothia sp.* C158-P CA-T3P21			
	Rothia sp. BP (EU725780)	Poultry - India	Otitis	unpublished
	Rothia sp. CCUG 25688 (AJ131122)	Mouse nose	-	Collins et al. (2000)
Outgroup	*Escherichia coli* str. K-12 substr. MG1655 (NC_000913)			
	Pseudomonas aeruginosa PAO1 (NC_002516)			

Supplementary Table 5.13: Isolation source and references of sequences extracted and used in the 16S phylogenetic analysis. (continued)

Category	Organism (Accession)	Isolation source	Underlying disease	Reference
Environmental	*Rothia sp.* CMG M10 (EU081515)	Pakistan coastline	-	Uzair B. (Ref unclear)
	Rothia sp.YIM C456 (EU135638)	Haloalkaline soil	-	Cui XL. et al. (Ref unclear)
	*Rothia sp.*ZHT413 (EU873349)	Venerupis philippinaram shell conglutination mud	-	unpublished
	Rothia sp. JSM 078151	Saline soil	-	unpublished
	Rothia sp. RA22 (FJ898305)	spring	-	Yang H and Lou K. (2011)
	Rothia sp. ZF11 (GQ891672)	water from high level natural radiation area	-	Zakei D. et al. (2010)
	Rothia sp. LH-CAB6 (HQ717389)	Air around 4200 m	-	unpublished
	Rothia sp. DG3 (JN208194)	Soil	-	unpublished
	Rothia sp. BBH4 (AM183255)	Deep sea sediment	-	unpublished
	Rothia amarae (AY043359)	Foul water sewer sludge	-	Fan Y. et al. (2002)
	Rothia sp. RV13 (GU318366)	Dysidea tupha (marine sponge)	-	unpublished
	Rothia sp. 3_1/4V	Semi-coke	-	unpublished
	Rothia terrae strain L-143 (NR_043968)	soil	-	Chou YJ. et al. (2008)
	Rothia sp. J03 (DQ409139)	Herbs in tumulus	-	unpublished
	Rothia sp. Piab1P (DQ457602)	nodules of Hedysarum glomeratum	-	Muresu r. et al. (2008)
Milk	*Rothia sp.* H29 (EF204383)	Raw milk	-	Hantsis-Zacharov E. et al. (2007)
	Rothia sp. H7 (EF204384)	Raw milk	-	Hantsis-Zacharov E. et al. (2007)
	Rothia sp. H21 (EF204385)	Raw milk	-	Hantsis-Zacharov E. et al. (2007)

Supplementary Table 5.14: Protein-coding genes used for multilocus phylogenetic inference.

Organism	Isolation Source	Genes (Accession)
Rothia dentocariosa ATCC 17931	Human oral	RecA (ADP41036) RpoA (ADP39732) Inf2 (ADP41005)
Rothia dentocariosa M567	Human oral cavity	RecA (EFJ76920) RpoA (EFJ77777) Inf2 (ZP_07072987)
Rothia mucilaginosa ATCC 25296	Human oral	RecA (EET75911) RpoA (EET75429) Inf2 (ZP_05367468)
Rothia mucilaginosa DY-18	Human oral	RecA (BAI65285) RpoA (BAI64339) Inf2 (BAI65313)
Rothia mucilaginosa M508	Human airways sample	RecA (EHB87709) RpoA (EHB88703) Inf2 (EHB87687)
Rothia mucilaginosa CF1E	Cystic Fibrosis sputum	RecA RpoA (This study) Inf2

Supplementary Table 5.15: Genes missing from the CF1E genome scaffold, based on contig mapping to the reference genome DY18.

Gene name	Gene length
1-acyl-sn-glycerol-3-phosphate acyltransferase (EC 2.3.1.51) CDS	681
Cell division protein FtsI [Peptidoglycan synthetase] (EC 2.4.1.129) CDS	1,866
COG0834: ABC-type amino acid transport/signal transduction systems, periplasmic component/domain CDS	840
COG4420: Predicted membrane protein CDS	933
Dihydrolipoamide dehydrogenase (EC 1.8.1.4) CDS	1,407
FIG01028708: hypothetical protein CDS	1,464
FIG01029080: hypothetical protein CDS	153
FIG01029402: hypothetical protein CDS	1,749
FIG01114456: hypothetical protein CDS	1,146
PTS system, mannose-specific IIC component (EC 2.7.1.69) CDS	867
putative ABC transporter ATP-binding protein CDS	114
Ribonuclease D (EC 3.1.26.3) CDS	1,266
Transcriptional regulator, FUR family CDS	591

Supplementary Table 5.16: Metatranscriptomics sequence data summary.

Sample	Total number of sequences	Total number of non-rRNA reads	Total number of microbial transcripts	Total *R. mucilaginosa* hits (rRNA)	Total *R. mucilaginosa* hits (non-rRNA)
CF1D	1,991	1,900	283	4	0
CF1F	25,238	7,971	312	3	2
CF4A	68,414	59,394	1,030	0	0
CF4B	32,737	32,446	471	0	0
CF4C	36,172	34,411	7,442	15	120

* No metatranscriptome was generated from CF1E

References

Blouin P, Yvert M, Arbion F, Pagnier A, Emile JF, Eitenschenck L, Machet M, Plantaz D, Colombat P, Alla CA, Donadieu J (2010) Juvenile xanthogranuloma with hematological dysfunction treated with 2CDA-AraC. Pediatric Blood & Cancer 55:757–760. doi: 10.1002/pbc.22629

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. doi: 10.1093/bioinformatics/btp348

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32:1792–1797. doi: 10.1093/nar/gkh340

Fusconi M, Conti C, De Virgilio A, de Vincentiis M (2009) [Paucisymptomatic pneumonia due to Rothia mucilaginosa: case report and literature review]. Infez Med 17:100–104.

Gao Z, Tseng C, Strober BE, Pei Z, Blaser MJ (2008) Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS ONE 3:e2719. doi: 10.1371/journal.pone.0002719

Hodzic E, Snyder S (2010) A case of peritonitis due to Rothia mucilaginosa. Perit Dial Int 30:379–380. doi: 10.3747/pdi.2009.00146

Hogardt M, Heesemann J (2010) Adaptation of Pseudomonas aeruginosa during persistence in the cystic fibrosis lung. International Journal of Medical Microbiology 300:557–562. doi: 10.1016/j.ijmm.2010.08.008

Kaasch AJ, Saxler G, Seifert H (2010) Septic arthritis due to Rothia mucilaginosa. Infection 39:81–82. doi: 10.1007/s15010-010-0065-5

Kaufhold A, Reinert RR, Kern W (1992) Bacteremia caused byStomatococcus mucilaginosus: Report of seven cases and review of the literature. Infection 20:213–220. doi: 10.1007/BF02033062

Ko KS, Lee MY, Park YK, Peck KR, Song J-H (2009) Molecular identification of clinical Rothia isolates from human patients: Proposal of a novel Rothia species, Rothia arfidiae sp. nov. Journal of Bacteriology and Virology 39:159. doi: 10.4167/jbv.2009.39.3.159

Lee AB, Harker-Murray P, Ferrieri P, Schleiss MR, Tolar J (2008) Bacterial meningitis from Rothia mucilaginosa in patients with malignancy or undergoing hematopoietic stem cell transplantation. Pediatr Blood Cancer 50:673–676. doi: 10.1002/pbc.21286

Morgan EA, Henrich TJ, Jarell AD, Shieh W-J, Zaki SR, Marty FM, Thorner AR, Milner DA, Velazquez EF (2010) Infectious granulomatous dermatitis associated with Rothia mucilaginosa bacteremia: A case report. The American Journal of Dermatopathology 32:175–179. doi: 10.1097/DAD.0b013e3181b1c5ad

Ohno H, Mizumoto C, Otsuki Y, Oguma S, Yoshida Y (2010) The duration of functioning of a subcutaneous implantable port for the treatment of hematological tumors: a single institution-based study. International Journal of Clinical Oncology 15:172–178. doi: 10.1007/s10147-010-0039-8

Pahlavan S, Haque W, Pereira K, Larrier D, Valdez TA (2010) Microbiology of third and fourth branchial pouch cysts. The Laryngoscope 120:458–462. doi: 10.1002/lary.20724

Peltroche-Llacsahttanga H, Reichhart E, Schmitt W, Lüitticken R, Haase G (2000) Investigation of infectious organisms causing pericoronitis of the mandibular third molar. Journal of Oral and Maxillofacial Surgery 58:611–616. doi: 10.1016/S0278-2391(00)90151-4

Pinsky RL, Piscitelli V, Patterson JE (1989) Endocarditis caused by relatively penicillin-resistant Stomatococcus mucilaginosus. J Clin Microbiol 27:215–216.

Pulzova L, Bhide MR, Andrej K (2009) Pathogen translocation across the blood–brain barrier. FEMS Immunology & Medical Microbiology 57:203–213. doi: 10.1111/j.1574-695X.2009.00594.x

Pushalkar S, Mane SP, Ji X, Li Y, Evans C, Crasta OR, Morse D, Meagher R, Singh A, Saxena D (2011) Microbial diversity in saliva of oral squamous cell carcinoma. FEMS Immunology & Medical Microbiology 61:269–277. doi: 10.1111/j.1574-695X.2010.00773.x

Rogovik AL, Friedman JN, Persaud J, Goldman RD (2010) Bacterial blood cultures in children with sickle cell disease. The American Journal of Emergency Medicine 28:511–514. doi: 10.1016/j.ajem.2009.04.002

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. doi: 10.1093/bioinformatics/btg180

Siqueira Jr JF, Rôças IN, Paiva SSM, Magalhães KM, Guimarães-Pinto T (2007) Cultivable bacteria in infected root canals as identified by 16S rRNA gene sequencing. Oral Microbiology and Immunology 22:266–271. doi: 10.1111/j.1399-302X.2007.00355.x

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690. doi: 10.1093/bioinformatics/btl446

Vaccher S, Cordiali R, Osimani P, Manso E, Benedictis FM (2007) Bacteremia caused by Rothia mucilaginosa in a patient with Shwachman-Diamond syndrome. Infection 35:209–210. doi: 10.1007/s15010-007-6284-8

Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, Angly FE, McDole T, Mariella RP, Rohwer F, Haynes M (2011) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci U S A 108:4547–4553. doi: 10.1073/pnas.1000089107

Chapter 6

Metabolic activities of CF anaerobic communities and their responses to perturbations caused by oxygen and pressure

Abstract

The pulmonary disease of Cystic Fibrosis (CF) patients is characterized by mucus plugging of the airways, polymicrobial infections, prolonged ineffective immune responses, and ultimately irreversible airway tissue remodeling. Despite aggressive antimicrobial and anti-inflammatory treatments, CF patients undergo periodic pulmonary exacerbation (CFPE). The cause of CFPE is unknown and has been attributed to multifactorial elements. The diverse community assemblages in CF airways include aerobic bacteria, as well as facultative and obligate anaerobes. Metagenomic analysis showed high diversity of metabolic potentials and many of these anaerobes are capable of fermentation processes. Fermentation products have been found in CF sputum and breath gas. A preliminary longitudinal study also showed that an increase in fermentation products coincides with CFPE. The fermentation products can be toxic to both host cells and other microbial cells, altering the immune response and the physiology of opportunistic pathogens in the CF airways. This study investigates if microbial fermentative processes are the cause of CFPE. In addition, perturbations through oxygen and pressure were used in an attempt to inhibit fermentation processes. Here we collected forty-one CF sputum samples from 15 CF patients over a 6 months period for community culture and experimental perturbations *in vitro*. The fermentative signature was significantly higher during stable and exacerbation states compared to during and after

antimicrobial treatments. Hyperbaric treatments with and without oxygen significantly reduced the fermentative signature. Metatranscriptomic analysis showed that the *in vitro* culture model was highly enriched for anaerobes when compared to the initial communities characterized from the sputum samples. Environmental stress-related genes including superoxide reductase, rubrerythrin and neutrophil activating proteins are highly expressed by the anaerobes during their growth in the WinCF system. Genes involved in redox homeostasis and pyruvate fermentation are significantly correlated with the amount of fermentative signatures. Overall taxonomic and gene expression profiles revealed strong patient- and sample-specific signals despite the different experimental manipulations. High expression levels of ferroxidase and acyl carrier protein genes provide alternative therapeutic targets to alleviate the effect of anaerobes in CF lung pathogenesis.

Introduction

Pulmonary failure is the main cause of morbidity and mortality in Cystic Fibrosis (CF) patients (O'Sullivan and Freedman 2009). CF lung disease is characterized by polymicrobial infections (Sibley et al. 2008) and mucus plugging (Fuchs et al. 1994) of the airway, prolonged ineffective immune responses (Cohen and Prince 2012), and ultimately irreversible airway tissue remodeling. Despite aggressive antimicrobial and anti-inflammatory treatments, CF patients undergo periodic pulmonary exacerbation (CFPE). CFPE is characterized by increased cough and sputum production, as well as declined lung function, eventual loss of respiratory function, and death (Fuchs et al. 1994).

Based on the data generated with standard microbial culture techniques, *Staphylococcus aureus* and *Haemophilus influenza* are commonly found in younger CF patients (LiPuma 2010). As the disease progresses the communities often become dominated by *Pseudomonas aeruginosa*, *Burkholderia cepacia* complex, *Stenotrophomonas maltophilia*, *Achromobacter* spp. and *Rohia* spp.; many of which are highly resistant against multiple antibiotics (LiPuma 2010; Lim et al. 2013). Using 16S rRNA gene analysis by microarray and high-throughput sequencing, more than 1,000 different taxa across ~90 genus have been associated with CF airways infection (Cox et al. 2010; Guss et al. 2011). Many of the species identified are environmental and oral-associated opportunistic pathogens, consisting of both aerobes and anaerobes. Every adult CF patient however presents a unique microbial community (Lim et al. 2012; Lim et al. 2014a). Although CF associated community assemblages are taxonomically diverse, they carry very similar metabolic properties (Lim et al. 2012; Lim et al. 2014a). This suggests that the CF airway microenvironment is highly similar across CF patients and drives the community towards acquiring similar metabolic properties as the disease progresses.

Microbial metabolism in CF lungs is associated with disease state (Twomey et al. 2013) and potentially influence the progression of CF pulmonary disease. The lungs present a spatially structured microenvironment that ranges from highly oxygenated in the upper airways to anoxic within the mucus plugs (Worlitzsch et al. 2002; Cowley et al. 2015). Opportunistic pathogens including viruses, bacteria, and fungi evolve and stratify in this structured microenvironment to persist in the CF lungs. As the disease progresses, mucus and biofilm build up in the airway and create an environment with steep oxygen gradient (Worlitzsch et al. 2002). In addition, atypical pH values and nutrient availability within the mucus plugs are capable of driving microbes to use alternative heterotrophic metabolisms. When readily available electron acceptors (i.e. O_2, NO_3^-, and SO_4^{2-}) are depleted, facultative anaerobes tend to switch to fermentative metabolism.

The diverse community assemblages in CF airways include aerobes as well as facultative and obligate anaerobes. Metagenomics and metatranscriptomics studies revealed a range of metabolic functions that led to a theoretical electron tower model consisting of possible aerobic, anaerobic, and fermentation pathways represented by the conventional and non-conventional CF-associated bacteria (Quinn et al. 2015). However, the metabolism of the CF microbes as a community in CF airways remains uncharacterized. *P. aeruginosa* PAO1 preferentially penetrates into hypoxic airway surface liquid on human airway epithelial cultures (Worlitzsch et al. 2002; Alvarez-Ortega and Harwood 2007). This behavior is supported by metabolic adaptation of *P. aeruginosa* through increased expression of genes involved in anaerobic and microaerobic respiration (Alvarez-Ortega and Harwood 2007; Hoboth et al. 2009). A strong relationship between the presence of anaerobes and the level of metabolites such as lactate and putrescine has also been reported during CF patient exacerbation (Twomey et al. 2013).

In the lungs of CF patients, the perturbation from drugs, physical therapies, and airway remodeling following exacerbation events destabilizes the established community (Zarei et al. 2012). This creates new spatial structures, which select for populations capable of living under anoxic conditions and more difficult to eradicate through treatment. Therapeutic application of hyperbaric oxygen (HBO) is an established treatment for various conditions, especially in the management of chronic infections (Kranke et al. 2012). Many of these clinical successes provide an impetus for studies of HBO effects on microbial persistence and resistance against antimicrobial compounds in CF airways. For instance antibacterial effects of antibiotic exposure increases under increased oxygen partial pressure compared to normoxic conditions (Brown et al. 1968; Muhvich et al. 1989). According to Fick's first law of diffusion, oxygen penetration is dependent on the diffusion coefficient, which is a function of the viscosity of the solution. Therefore, the diffusion of O_2 through mucus is slow compared to water, and even slower across the thick dehydrated (and thus more viscous) mucus in the CF airways. Fick's law of diffusion across a fluid membrane (Mathieu 2006) further states that the amount of gas transferred per unit time ($\Delta N/\Delta t$) across thickness Δx is proportional to the area (A) available for exchange and the partial pressure difference (ΔP) of the gas across the membrane:

$$\Delta N/\Delta t = (K)(A)(\Delta P)/\Delta x$$

K = Krogh's Diffusion Coefficient

$$K = \alpha D$$

α = solubility of the gas

D = Diffusion rate

As the area and thickness are assumed to be the same in all conditions, the amount of gas transferred across the mucus plug per unit time should thus increase with increasing pressure (ΔP). In the context of the CF lung, increases in pO_2 within the mucus plug will disrupt anaerobic respiration and fermentation processes that may facilitate CF disease pathogenesis.

The present study therefore addresses two pivotal questions. First the roles of anaerobes in CF-associated microbial communities were investigated, comparing their fermentative signatures to pulmonary exacerbation through a longitudinal study design. Metagenomics, metatranscriptomics, and metabolomics approaches were concurrently collected with the WinCF community culture model (Quinn et al. 2015) to acquire an insight of the community structure and metabolic activities of CF anaerobes. Secondly we investigate the effect of perturbations through increased oxygen concentrations and pressure on microbial community structure and metabolism.

Materials and Methods

Sampling procedure: Fifteen CF patients from the University of California, San Diego adult CF clinic were recruited for this study (Table 6.1). The patients were selected based on eligibility criteria that included: (i) an established diagnosis of CF based on consistent clinical symptoms and confirmation by genetic screening of the CFTR gene, (ii) ability to perform induced sputum procedures and willingness to provide fresh sputum samples during every clinic visit, and (iii) the patients need to be classified as "Frequent Exacerbating Patient", i.e. 2 or more exacerbations per year according to attending physician. Informed consent was collected from each patient prior to the study.

Induced sputum samples and patients' information were collected in accordance with the University of California Institutional Review Board (HRPP #081500) and the San Diego State University Institutional Review Board (SDSU IRB #2121). Sterile saline was used as mouth rinse prior to the inhalation of 7% hypertonic saline via a nebulizer. Sputum was expectorated into a sterile cup over a thirty-minute period.

'Stable samples' were collected when the patients were clinically stable without respiratory symptoms (i.e., increase cough, increased sputum production, and suffering shortness of breath) and not treated acutely with antimicrobial compounds. 'Exacerbation samples' were collected when the patients were presented with increase respiratory symptoms and a drop in the FEV_1 for ≥15% compared to the patient's best FEV_1 in the past 12 months. 'Treatment samples' indicate samples that were collected during either oral or intravenous antibiotic treatments for exacerbation episodes. 'Post-treatment samples' were collected within 24 hours after either oral or intravenous antibiotic treatments for the patient's exacerbation episode.

Sample processing: Sputum samples were processed immediately after each collection as described in (Lim et al. 2012; Lim et al. 2014b). In short, each sputum sample was homogenized with a 3 ml syringe until no visible clumps remained. The homogenized sample was aliquoted into: (i) one tube containing Trizol-LS (Life Technologies) and 0.1 mm Zirconia silica beads (BioSpec) for the metatranscriptomics study, (ii) three sterile cryovials for metabolomics analysis, and (iii) one sterile cryovial for the community culture model. Extra sputum samples (if any) were aliquoted into 15 ml falcon tubes for metagenomics analysis, and into separate cryovials for storage. Metatranscriptomic tubes were homogenized immediately at medium speed for 10 minutes. Samples for community culture model and metagenomics were transported in ice to the lab prior to culture study (see below) while others were transported on dry ice and stored at -80°C until further processes.

Community culture model: Artificial sputum media (ASM) that mimics the constituents of sputum from CF patients was made based on the modified recipes from (Sriramulu et al. 2005; Palmer et al. 2007; Fung et al. 2010). The modifications include (i) increased mucin concentration to 20 mg/ml to better reflect the viscosity in CF sputum, (ii) additional 3 μg/ml ferritin was added as an alternative iron source and as substitute of diethylene triamine pentaacetic acid (DTPA) (Reid et al. 2007; Hare et al. 2012), and (iii) omitting the addition of antibiotics onto the media (Quinn et al. 2015). The media was kept sterile by using sterile 1X PBS, autoclaving the 5% mucin for 10 mins, and ensuring sterile microbiology techniques during handling. All other reagents were sterilized through 0.22 µm – filtration prior to the mixing.

Within 24 hours post-collection, homogenized sputum was diluted 10-fold using sterile ASM. Physiochemical variables including pH changes, respiratory activity, and macrocolonies formation were monitored through the addition of 1 mg ml^{-1} phenol red/bromocresol purple (Sibley et al. 2008), tetrazolium dye (Bhupathiraju et al. 1999), and 3% coomassie blue, respectively, at 10^{-1} volume of the sputum-ASM mixture (Supplementary Figure 6.1). The mixture was then inoculated into the capillary tubes. The tubes were plugged at one end and laid horizontally in the MH-holder (Supplementary Figure 6.2). Each holder holds seventeen tubes (the sample arrangement within each holder is shown in Supplementary Figure 6.2). Each holder contained tubes for either (i) Hyperbaric oxygen (HBO; 2.4 atm at 100% O_2), (ii) Normobaric hyperoxic (HO; 1.0 atm at 100% O_2), (iii) Hyperbaric normoxic (HBA; 2.4 atm at 21% O_2), (iv) Hyperbaric hypoxia (HBN; 2.4 atm at <1% O_2 using nitrogen), or (v) Control (CTRL; 1.0 atm at 21% O_2) treatment, per sputum sample. The pressure at sea level is represented by 1 atmosphere (atm) and the treatment protocols are based on UCSD hyperbaric oxygen chamber treatment protocol guidelines for monochamber. All treatments were conducted for 60 minutes prior to incubation in 5% CO_2 environment at 37 °C (Supplementary Figure 6.3). The tubes were then monitored and imaged every 24 hours for 72-96 hours.

Capillary tube imaging and image analysis: All images were taken in the dark on white backlight (Logan 5.5 X 9 inches^2 light box) using a Canon EOS Rebel T3 camera (Canon U.S.S Inc., Melville, NY, USA). All images were taken under identical settings (manual focus, ISO 3,200, Aperture F4.5) and saved in raw and JPEG format. Images were cropped to retain only the tubes (Supplementary Figure 6.1) and used for further analysis. Tubes 14-19 (Supplementary Figure 6.1) within each image were used for calculating the volume of tubes occupied with gas bubbles and the final percentage represent the average of all six tubes. Statistical analysis was carried out using the PRISM software (Graphpad).

Processing capillary tubes contents for metatranscriptomics and metabolomics: At the end of 72 – 96 hour incubation, the content of tubes 14-19 (tubes with ASM + sputum only) was collected for metatranscriptomic and metabolomic analysis. ASM and sputum from three tubes were added into a cryovial containing 1 mm zirconia beads and 1 ml Trizol LS®, mixed, and kept at -80°C until further processing. ASM and sputum from the other three tubes were separately added into three sterile cryovials for metabolomic analysis. Three sterile ASM without sputum (Tubes 20-22) were added into one sterile cryovial to serve as negative control for the metabolomcis study. Metabolomic samples were all kept at -80°C and transferred to the University of California, Davis West Coast Metabolomics Center. Metabolites were extracted using the standard plasma extraction protocol (20 µl of sputum in 1 ml of 3:3:2 acetonitrile:isopropanol:water) and analyzed using gas chromatography Time-Of-Flight (GC-TOF) mass spectrometry.

Microbial metagenomes: Sputum collected for microbial metagenomes were treated as described in (Lim et al. 2014b). Briefly, samples were treated with β-mercaptoethanol at 0.2-sample volume and rocked at room temperature for one hour.

Cells were spun down and repeatedly washed three times with sterilized deionized water to lyse human cells. The washed cells were then resuspended in 1X DNase buffer (50 mM NaAc; 10 mM $MgCl_2$; 2 mM $CaCl_2$) and treated with 15 units ml^{-1} DNase I (Calbiochem) to remove extracellular DNA. Total DNA was extracted using the Nucleospin® Tissue kit (Macherey-Nagel) following the gram-positive variation protocol. Library preparation was done using the Nextera XT kit (Illumina) and all samples were sequenced in-house on Illumina MiSeq platform using the 600-flow sequencing chemistry.

Metatranscriptome RNA isolation and RNA-seq library preparation: Sputum and capillary tubes contents in Trizol LS® were thawed on ice. RNA was extracted using a combination of chloroform phase separation and Zymo RNA Clean & Concentrator column (Zymo). Interphase and phenol-chloroform layers were saved for DNA extraction. RNA-seq libries were constructed using a combination of the Ribo-Zero Gold Epidemiology rRNA removal kit (Epicentre) and TruSeq Stranded mRNA kit (Illumina). Libraries were sequenced at the University of California, Davis Genome Center on Illumina HiSeq platform using the rapid-mode paired-end 2 X 250 bp chemistry.

Metagenomics Data Preprocessing: Metagenomic reads were preprocessed using PrinSeq (Schmieder and Edwards 2011) to remove low quality reads (minimum quality score of 25). The resulting reads were then compared to a human genome database (hg19 genome) using SNAP aligner (version 1.0beta; Zaharia et al. 2011) and non-human reads were extracted using bamtoFastq (http://gsl.hudsonalpha.org/information/software/bam2fastq).

Metatranscriptomics Data Preprocessing: FASTQ files were manually uploaded to Illumina BaseSpace for quality check and data preprocessing. Read quality and adapter trimming were performed using FASTQToolKit (in BaseSpace Labs) using the parameter:

(--adapters-5
ACACTCTTTCCCTACACGACGCTCTTCCGATCT,GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
--adapters-3
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT,GATCGGAAGAGCACACGTCTGAACTCCAG
--adapter-trimming-stringency 0.9 --min-length 70 --min-seq-complexity 70 --trim-tail-5 5 --trim-tail-3 5 --trim-qual-score 25 --trim-ns). All reads shorter than 70 bp were discarded. SortMeRNA v.2.0 was used to extract non-rRNA sequences.

Data Analysis: Bacterial taxonomy of the metagenomic sequences were assigned based on BLASTn search against the NCBI nucleotide (NT) database (Figure 6.1). Sequences assigned to the phylum Chordata, or any synthetic/vector sequences were further removed prior to the comparison of metagenomic datasets against the KEGG database using BLASTx. The resultant KEGG-BLASTx output was analyzed using the

HUMAnN pipeline (Abubucker et al. 2012). The normalized relative abundance values (file *-04b.txt) were used for subsequent analysis.

All metagenomic sequences were also concatenated into a single file and assembled with SPADES assembler (Bankevich et al. 2012) using the parameter –k 127 and --careful. Large contigs (>1,000 bp) were extracted and used for open reading frames (ORFs) prediction via MetaProdigal (Hyatt et al. 2012). Final contigs and resulting ORFs were then used as the backbone for metatranscriptomics sequence recruitment.

MetaphlAn2 (Segata et al. 2012) was used for taxonomic assignment of the metatranscriptomics reads. Kyoto Encyclopedia Genes and Genomes (KEGG) database was used for functional assignment based on BLASTx. The KEGG-BLASTx output of was analyzed using the HUMANn software (Abubucker et al. 2012) and further analyzed using R statistical software packages. Using the taxonomical and gene annotations, only read 1 (R1) was used for the calculation of relative and normalized abundance.

An in-house WinCF-specific transcriptomics database (CFaDB) was constructed using all annotated gene sequences from the genomes of the six genera (*Veillonella* spp., *Prevotella* spp., *Streptococcus* spp., *Haemophillus* spp., *Fusobacterium*, and *Granulicatella* spp.) identified from the capillary cultures (Figure 6.5). The gene sequences were downloaded from the PATRIC database (www.patricbrc.org). Non-rRNA sequences were compared against the CFaDB using BLASTx. Hit counts were normalized based on the formula:

$$f(i) = \frac{\left(\frac{p(i)}{L(i)}\right)}{\sum_{j=1}^{K}\left(\frac{p(j)}{L(j)}\right)}$$

where $f(i)$ = Normalized abundance of gene i

$p(i)$ = the fraction of the identifiable reads that map to gene i

i.e. $\left(\frac{n(i)}{N}\right)$ where $(n(i))$ is the number of hits to gene i and N is the total number of hits to all genes

$L(i)$ = The effective length of gene i in bases

K = the number of different genes within that sample

Normalized abundance where used for all downstream analysis. Random Forests analyses (Breiman 2001) were done using the R package *randomForest* (Liaw and Wiener 2002).

Table 6.1: Total number of samples collected at different health status. Fifteen patients were recruited for this study and a total of 35 samples were collected.

Sample ID	Age	Genotype	Exacerbation	Post - Treatment	Stable	During Treatment	Total
CF1	36	dF508	1	0	0	0	1
CF2	23	dF508/3849	1	0	0	0	1
CF3	32	dF508/c.1340delA	0	1	1	0	2
CF4	22	dF508/394delTT	1	0	0	1	2
CF5	22	dF508/1249	1	1*	1	0	3
CF6	32	dF508/unknown	0	0	1	0	1
CF7	24	dF508/p.E585	2	2	1	0	5
CF8	60	dF508	3	0	1	0	4
CF9	40	dF508	0	0	1	5	6
CF10	29	dF508	1	0	1	0	2
CF11	29	dF508	3	0	2	0	5
CF12	23	dF508/3849	0	0	0	1	1
CF13	26	dF508	0	0	0	1	1
CF14	24	W1282X/L206W	1/1*	0	1	0	3
CF15	29	dF508/3737C>T	0	0	1	1/1*	3
CF16**	24	dF508	1	0	0	0	1
						Total	**41**

* Capillary culture experiment omitted.

** Patient dropped out of the study. No metagenome was constructed.

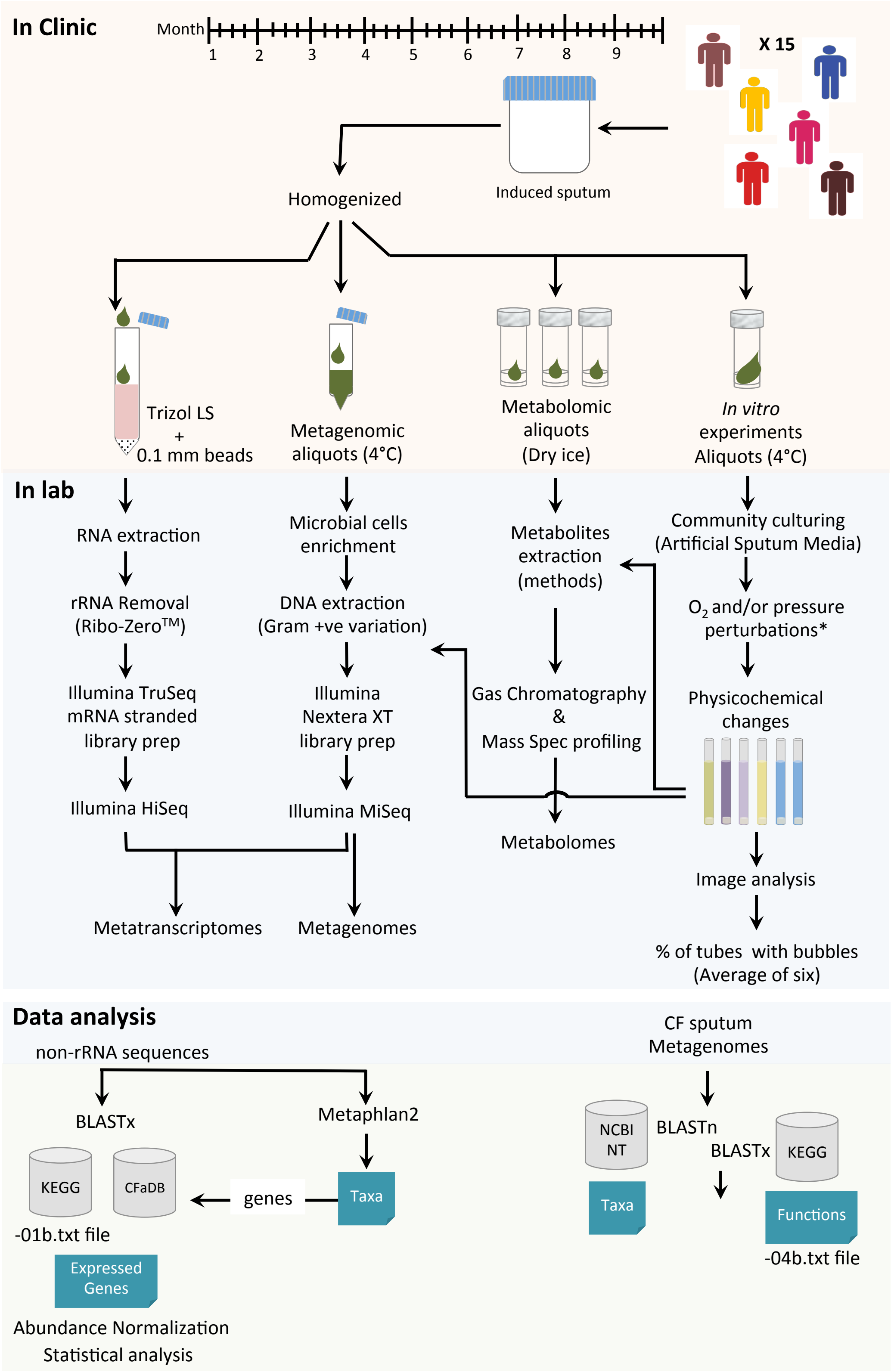


Figure 6.1: Overview of the experimental and data analysis processes.

Results

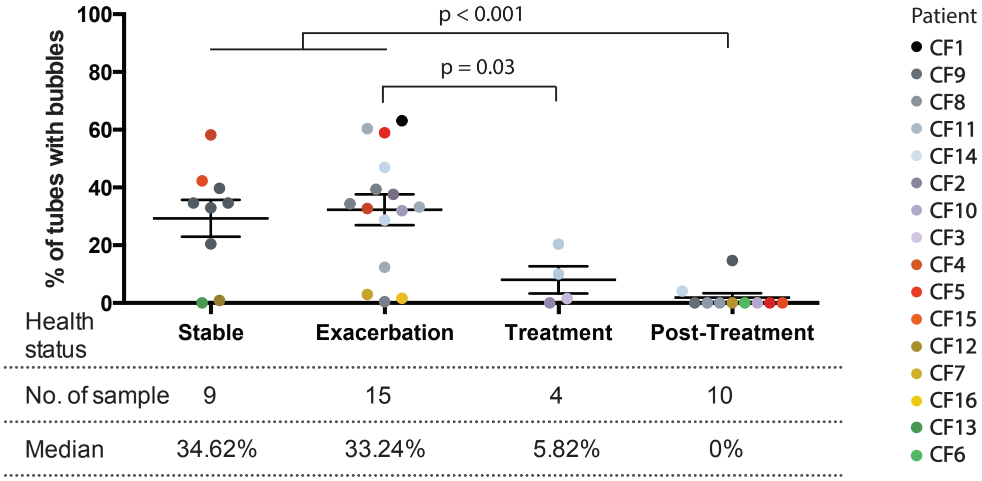
Induced sputum samples (n=41) were collected from fifteen CF subjects ages 22 – 60 (median = 27.5) across a 6-month period (Table 6.1). The patients' clinical stability and exacerbation events during sampling were defined based on pulmonary function tests and patient's reported outcomes such as increased sputum production and cough (See Methods) (Fuchs et al. 1994).

The main bacterial species (relative abundances > 5%) detected by metagenomic sequencing (Supplementary Table 6.1) across all CF patients include *Pseudomonas aeruginosa*, *Stenotrophomonas maltophilia*, *Staphylococcus aureus*, *Rothia mucilaginosa*, *Streptococcus* spp., *Prevotella melaninogenica*, *Veillonella parvula*, *Achromobacter xylosoxidans*, and *Marinomonas* sp. MQYL1 (Table 6.2). Eight out of the fifteen patients (53%) had *P. aeruginosa* as their major colonizer (relative abundances range from 40% - 97%), while in the other patients the dominant species differed between *R. mucilaginosa*, *S. aureus,* or *S. maltophilia* (Table 6.2; Supplementary Figure 6.5). Despite the differences in the community profiles, the functional compositions based on KEGG annotation were highly similar across all patients (Supplementary Figure 6.6).

Table 6.2: Microbial communities associated with each CF patients. Taxanomy was assigned based on sequence comparison using BLASTn against the NCBI non-redundant nucleotide (nt) database. No metagenome was generated for patient CF15 & CF16.

Patient ID	Major bacteria	Patient ID	Major bacteria
CF1	*Pseudomonas aeruginosa* (92%) Other bacteria entails < 1% of the total reads	CF8 (August; November)	*Pseudomonas aeruginosa* (97%; 91%) Other bacteria entails < 1% of the total reads
CF2	*Streptococcus* spp. (39%) *Rothia mucilaginosa* (20%) *Staphylococcus aureus* (8%) *Prevotella melaninogenica* (7%)	CF9 (July; November)	*Pseudomonas aeruginosa* (62%; 41%) *Prevotella melaninogenica* (2%; 15%) *Streptococcus salivarius* (2%; 14%) *Veillonella parvula* (1%; 8%) *Rothia mucilaginosa* (8%; 3%) *Streptococcus parasanguinis* (6%; 5%)
CF3	*Rothia mucilaginosa* (65%) *Streptococcus pneumonia* (8%) *Streptococcus mitis* (6%) Other *Streptococcus* spp. (32%)	CF10	*Stenotrophomonas maltophilia* (43%) *Pseudomonas aeruginosa* (24%) *Veillonella parvula* (16%) *Streptococcus parasanguinis* (8%)
CF4	*Staphylococcus aureus* (53%) *Veillonella parvula* (5%) *Streptococcus pneumoniae* (5%) *Pseudomonas aeruginosa* (4%) *Rothia mucilaginosa* (3%) Other *Streptococcus* spp. (18%)	CF11 (August; October)	*Pseudomonas aeruginosa* (97%; 40%) *Streptococcus intermedius* (< 1%; 29%) *Achromobacter xylosoxidans* (< 1%; 10%) *Staphylococcus aureus* (10%)
CF5	*Rothia mucilaginosa* (47%) *Marinomonas* sp. MWYL1 (14%) *Staphylococcus aureus* (4%)	CF12	*Pseudomonas aeruginosa* (77%) *Streptococcus parasanguinis* (11%) Other bacteria entails < 1% of the total reads
CF6	*Pseudomonas aeruginosa* (92%) Other bacteria entails < 1% of the total reads	CF13	*Pseudomonas aeruginosa* (66%) *Streptococcus parasanguinis* (11%) *Rothia dentocariosa* (6%) *Streptococcus salivarius* (3%) *Rothia mucilaginosa* (3%) *Veillonella parvula* (1%)
CF7	*Pseudomonas aeruginosa* (47%) *Streptococcus salivarius* (12%) *Streptococcus parasanguinis* (9%) *Veillonella parvula* (6%)	CF14	*Staphylococcus aureus* (46%) *Rothia mucilaginosa* (12%) *Streptococcus parasanguinis* (12%) *Prevotella melaninogenica* (8%)

Fermentation signatures were found in "stable" and "exacerbation" samples: In the fresh sputum samples that were added to sterile ASM (Figure 6.1) microbial growth and chemical changes due to respiration, pH changes, gas production were clearly visible through the capillary tubes (Supplementary Figure 6.1). The volume of capillary tubes occupied by gas (Figure 6.2) was calculated as mean percentages from six capillary tubes containing ASM and sputum without additional chemical (Supplementary Figure 6.1; Tube 14-19). After 72 – 96 hours of incubation at 5% CO_2, the tubes containing samples collected during exacerbation and stable states showed higher amounts of gas bubbles compared to those with treatment and post-treatment samples (Figure 6.2). Stable samples produced 0-58% gas (median = 35%), while exacerbation samples produced 0-63% gas (median = 34%) (Figure 6.2). The amount of gas production by the microbial community from exacerbation samples was significantly higher than treatment and post-treatment samples (Figure 6.2; Mann-Whitney test; $p = 0.03$ with $df = 17$ and $p < 0.0001$ with $df = 23$, respectively). However, the amount of gas production from stable samples was only significantly higher compared to post-treatment samples (Mann-Whitney test; $p < 0.001$; $df = 17$) but not to treatment samples (Mann-Whitney test; $p=0.0853$; $df = 11$). Interestingly, overall gas production can be prevented when tetrazolium dye was added to the same ASM-sputum mixture (Supplementary Figure 6.1). At least one stable or exacerbation sample was collected from fourteen of the total fifteen CF patients recruited and 93% (13/14) of these patients carry microbial communities that produce gas in the WinCF system (Supplementary Table 6.2).



Health status	Stable	Exacerbation	Treatment	Post-Treatment
No. of sample	9	15	4	10
Median	34.62%	33.24%	5.82%	0%

Figure 6.2: The amount of gas production through CF microbial community metabolisms. Fresh sputum samples were inoculated into capillary tubes and the volume of tubes occupied by gas bubbles was calculated after 72 – 96 hours incubation at 37°C and 5% CO_2. Samples collected during stable and exacerbation states demonstrated significantly higher amount of fermentative characteristic compared to treatment and post-treatment samples. (Mann-Whitney test; Stable vs Treatment, $p = 0.0853$, df = 11; Stable vs Post-treatment, p = < 0.001, df = 17; Exacerbation vs Treatment, $p = 0.03$, df = 17; Exacerbation vs Post-treatment, p = < 0.0001, df = 23)

Pressure with and without oxygen decreased fermentation signatures: The hypothesis implicit to this work is that increased oxygen concentration within the mucus plug will prevent gas production. Twenty-eight sputum samples collected were subjected to hyperbaric oxygen (HBO; 100% O_2 and 2.4 atm), pressure (HBA; 21% O_2 and 2.4 atm), hyperoxic (HO_2; 100% O_2 and 1 atm), and hyperbaric nitrogen (HBN; <1% O_2 and 2.4 atm) treatments prior to the 37°C incubation under 5% CO_2 for 72 – 96 hours (Supplementary Figure 6.3).

HBO_2, HBA, and HO_2 treatments all increased the penetration of O_2 into the capillary tube mucus plugs (Supplementary Figure 6.4). In the control tubes oxygen was only detected in the first 2-3 mm of the mucus plug within the capillary tube. Upon exposure at a normobaric condition to 100% O_2 for 60 minutes elevated O_2 concentrations were detected in the first 9-10 mm of the mucus plug surface. An increase in pressure to 1.5 atm and 2.4 atm at 100% O_2 did not increase penetration of O_2 (Supplementary Figure 6.4). Absolute concentrations of the O_2 within the tubes were not assessed.

The effect of changes in pressure and oxygen concentrations on microbial metabolism under these conditions is markedly different. Compared to the controls (CTRL: non-treated), the effect of pressure with or without oxygen significantly reduced the amount of gas produced by the sputum microbial communities (Figure 6.3). The addition of antibiotics also reduced the amount of gas produced (Supplementary Table 6.3) in addition of the effect of pressure. The amount of recovered total RNA per volume of liquid showed that the pressure and oxygen treatments did not affect the overall growth of bacteria (Supplementary Figure 6.7).

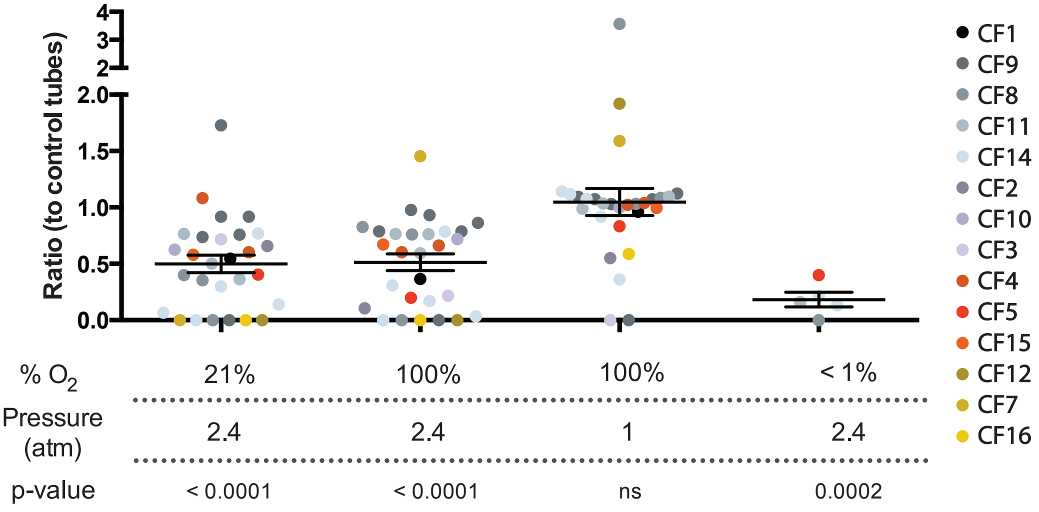


Figure 6.3: The effect of oxygen and pressure on microbial metabolism in the WinCF system. Sputum samples were subjected to hyperbaric oxygen (HBO; 100% O_2 and 2.4 atm), pressure (21% O_2 and 2.4 atm), hyperoxic (100% O_2 and 1 atm), and hyperbaric nitrogen (HBN; <1% O_2 and 2.4 atm) treatments prior to 37°C incubation under 5% CO_2 for 72 – 96 hours. Elevated pressure significantly reduced the amount of gas produced by the sputum microbial communities.

Thirty-one individual samples were selected from 7 experiments across 3 patients (CF9, CF11, and CF14) to characterize the molecular signatures within the tubes through metatranscriptomics and metabolomics (Supplementary Table 6.4). These samples represent a range of gas production, form 0% to 66% (Supplementary Table 6.4). Total RNA was extracted from 3 sputum samples and 31 WinCF contents stored in Trizol LS. Total rRNA was depleted using the Ribo-Zero kit (Epicentre) prior to high-throughput sequencing (See Methods). More than 19 million paired reads were retained after filtering low quality sequences and removing rRNA-like sequences. The same samples set with additional 45 sputum samples were also subjected to gas chromatography mass spectrometry (GC-MS) analysis. Across all sputum and WinCF samples, 255 metabolites were detected. Across all metatranscriptomes (from sputum and WinCF culture) 13,610 unique genes were annotated from the KEGG database. Based on a supervised random forests classification model, the metabolites and transcripts from the WinCF communities are distinctly different from the sputum communities (Figure 6.4).

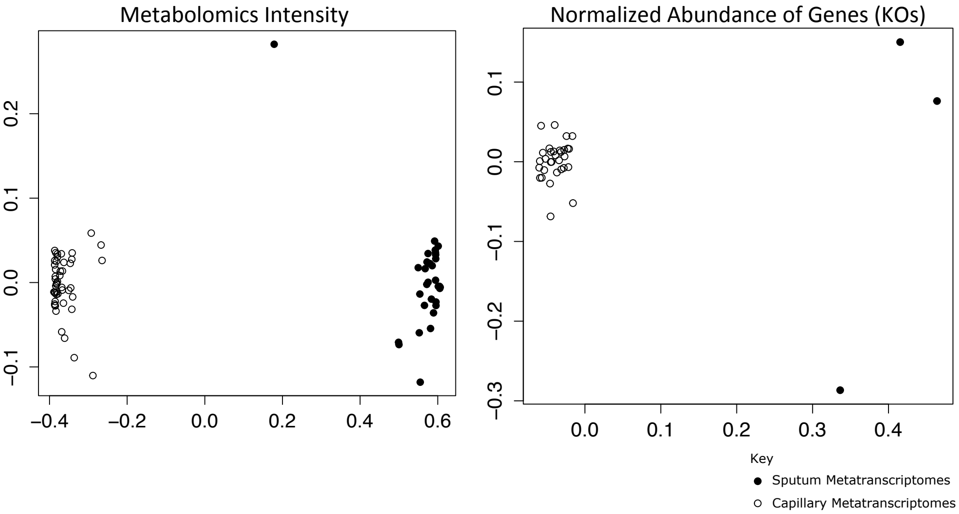


Figure 6.4: Comparison of metabolites (A) and genes (KOs) identified in sputum samples (filled circles) and capillary samples (unfilled circles) as shown by multidimentional scaling (MDS) plot based on the output of supervised random forests analysis. The classification from random forests distinctly differentiates sputum and capillary samples, with out-of-bag (OOB) error at 0% based on metabolites and 2.94% based on genes.

Mainly anaerobes were growing within the WinCF system: Metaphlan2 (Segata et al. 2012) was used to assign the taxonomic group to each sample using the R1 read of the non-rRNA sequences (Figure 6.5). The main bacteria in the WinCF are anaerobes that include *Prevotella* spp. (*P. melaninogenica* and *P. naceiensis*), *Streptococcus* spp., *Veillonella* spp. (Unclassified sp., *V. atypica*, and *V. parvula*), *Fusobacterium nucleatum*, *Haemophilus parainfleunza*, and *Granulicatella* spp. (Figure 6.5). The taxonomical profiles showed that patients are more similar within themselves than across the perturbation experiments (Random Forests Classification OOB error of 0% and 100%, respectively; Figure 6.5). Therefore, we hypothesized that the CF microbial communities switch their metabolism in response to the biochemical perturbations within the mucus plug caused by pressure and oxygen without significantly changing the composition of the community members.

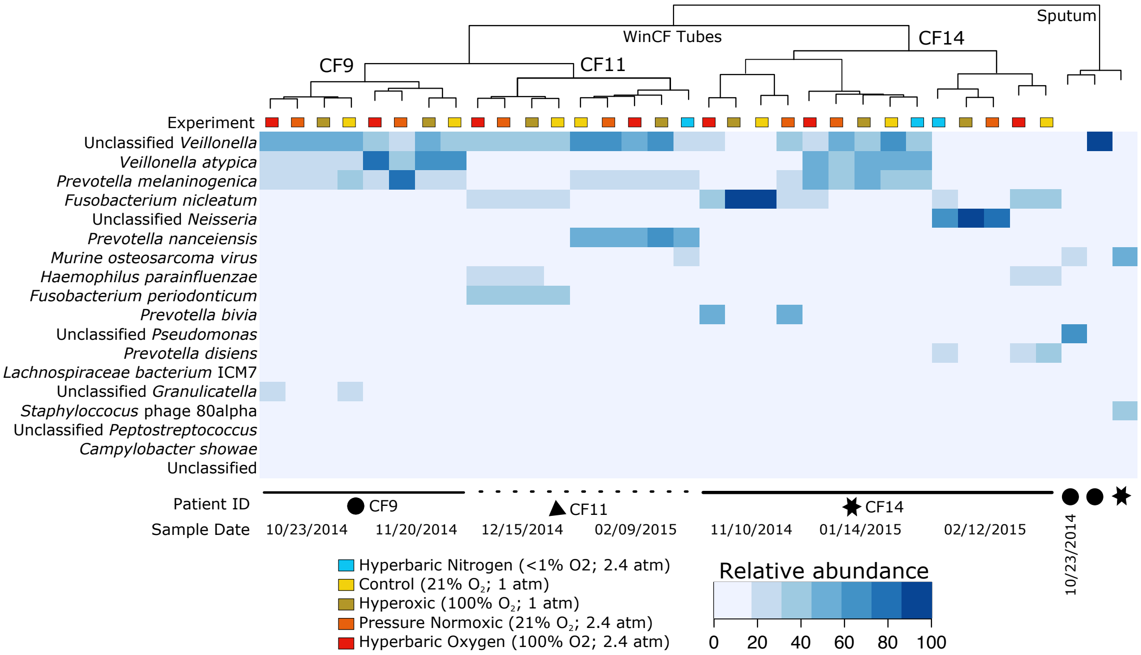


Figure 6.5: Taxonomic group assigned to non-rRNA sequences using Metaphlan2. Dendogram shows the hierarchical clustering of samples based on Euclidean distance calculated from the overall relative abundance of species assigned via Metaphlan2. Black symbols represent the three patients and the colored squares describe different experimental manipulations perform during WinCF culture (See Methods). Patients are more similar within themselves than the different experimental manipulations (Random Forests Classification OOB error 0% and 100%, respectively).

Gene expression profiles present highly patient- and sample-specific signals: The most abundant KEGG assigned metabolic pathways identified from the WinCF communities include ribosome, phosphotransferase system (PTS), homologous recombination, aminobenzoate degradation, glycolysis and gluconeogenesis, protein export, as well as arginine, proline, thiamine, beta-alanine, and selenocompound metabolism (Figure 6.6A). The most abundant modules (defined as functional units in the KEGG metabolic pathway maps) include the PTS system Mannose- and N-acetylgalactosamine specific II component, NAD- and NADP-malic enzyme type in C4-dicarboxylic acid cycle, F-type ATPase bacteria, reductive citric acid cycle (Arnon Buchanan cycle), adenine nucleotide biosynthesis, ATP synthase, methane oxidation, and gluconeogenesis (Figure 6.6B). We hypothesized that individual genes and not the organism is important in the observed microbial responses to biochemical perturbations. Therefore, protein sequences from the genomes of the six most abundant genus observed in the WinCF transcriptomes (*Prevotella*, *Veillonella*, *Fusobacterium*, *Streptococcus*, *Haemophilus*, and *Granulicatella*) were downloaded from the PATRIC database (www.patricbrc.org) to generate a WinCF-specific database (CFaDB).

The WinCF transcriptomes were compared against the CFaDB using BLASTx. On average, 66% of the transcripts mapped against the CFaDB compared to 46% of transcripts that mapped against the KEGG database (Supplementary Table 6.4). A total of 7,993 unique genes were identified from the CFaDB across all WinCF transcriptomes. The most abundant protein expressed was feroxidase (1.3% of total transcripts across all samples) while other interesting proteins are environmental stress-related genes that include superoxide reductase, acyl carrier protein, neutrophil activating protein A, rubrerythrin, and Thioredoxin (Figure 6.6C) were present in high abundance. The overall WinCF gene expression profiles showed strong patient- and sample-specific signals despite the different experimental conditions (Supplementary Figure 6.8). A supervised random forests analysis classified the gene expression profiles at 0% out-of-bag error based on individual patient, and 100% out-of-bag error based on experimental treatments.

Correlation of gene abundances against the percent gas production: A supervised random forests model based on the normalized RNA gene abundance and the percent bubbles production were able to predict 78.6% of the variance. The most important genes identified from the model were used to perform a linear regression of normalized gene abundance against the percent gas production (Supplementary Figure 6.9). Some of the genes that are significantly represented are presented in Figure 6.7. As the percent of gas production increased, the expression level of carbon starvation protein A, 3-hydroxybutyryl-CoA / 3-hydroxyacyl-CoA dehydrogenase, thiol:disulfide oxidoreductase related to ResA genes, and all components of acetoin dehydrogenase (E1-E3) increased significantly (Figure 6.7).

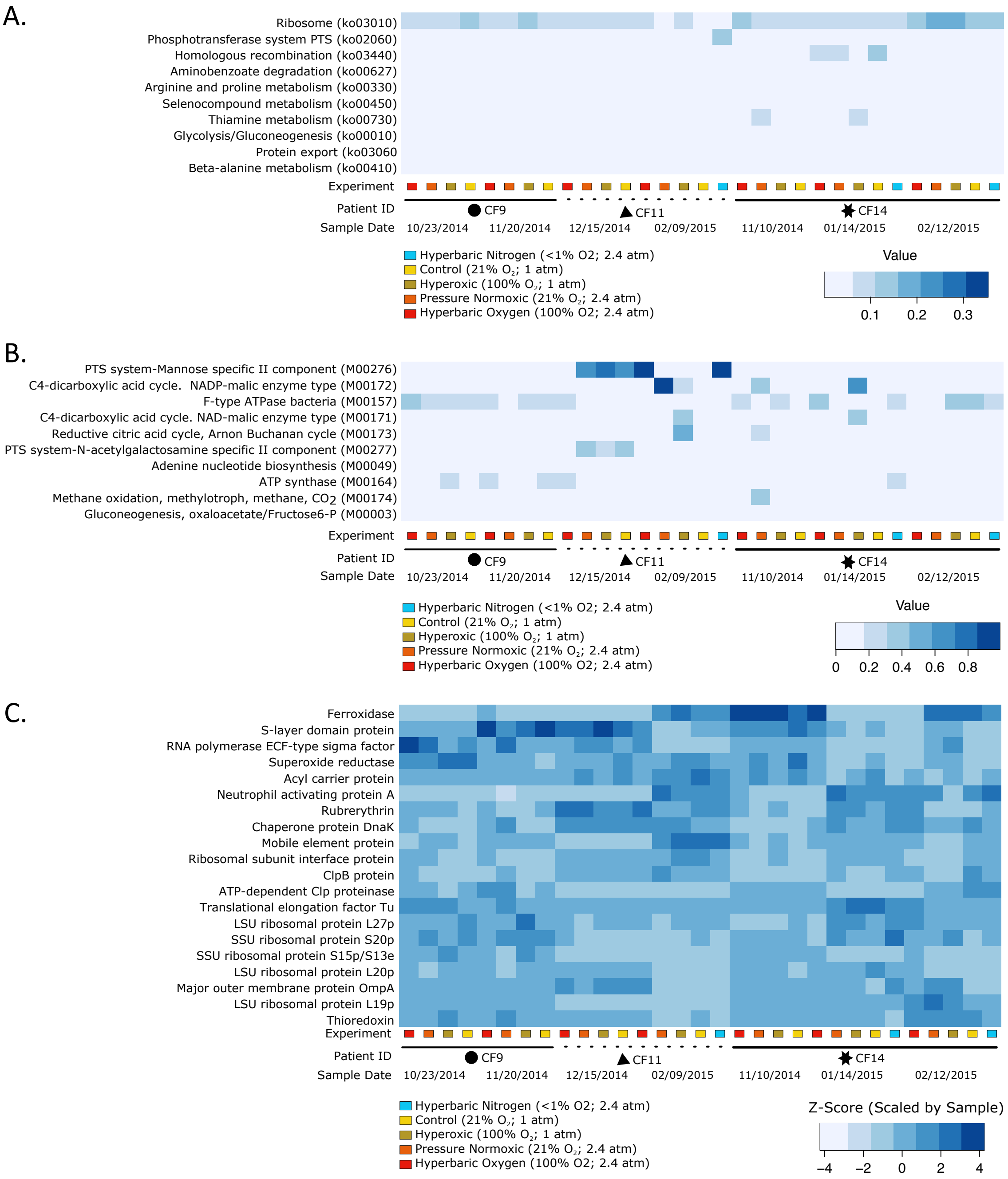


Figure 6.6: Functional characterization of the capillary communities based on the normalized abundance of (A) pathways and (B) modules annotated based on the KEGG database, and (C) genes annotated based on the CFaDB.

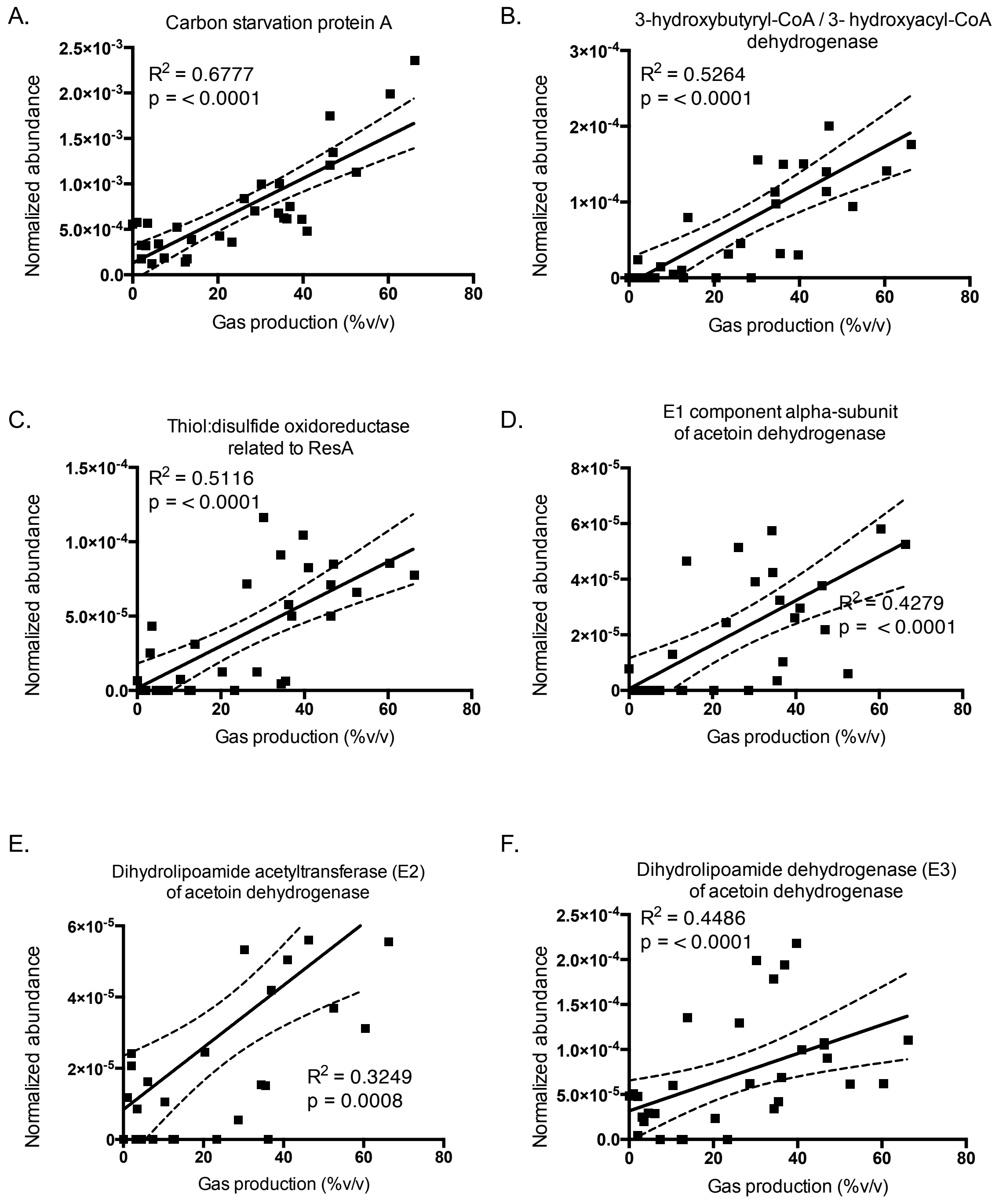


Figure 6.7: Linear regression of the normalized transcripts abundance against the amount of gas production (%v/v). The selected transcripts (A-C, and F) were identified from a supervised random forests model (Supplementary Figure 6.9), while D and E (the E1 and E2 components of acetoin dehydrogenase complex) were extracted to compliment the E3 component (F) of acetoin dehydrogenase complex. The expression level of the E1 component beta-subunit is presented in Supplementary Figure 6.10.

Metabolite profiles were sample-specific: GC-MS was used to characterize the metabolites produced by the WinCF communities. Three control tubes with sterile ASM were included during the experiment and were submitted simultaneously for GC-MS analysis. The reported average intensity from the three control sterile samples was used to subtract the intensity from the samples with sputum and negative values were treated as zeros. Subsequent analysis used the corrected intensity for each sample and therefore, did not take into account the substrates that were consumed by the WinCF communities. Similar to the transcriptomes analysis, the metabolomics profiles were more likely to cluster by patients than the experimental perturbations (Figure 6.8A). However, some treated samples were randomly distributed across the dendogram (Figure 6.8A). The supervised classification based on random forests analysis showed 0% OOB error when the profiles were classified based on patients, and 100% OOB error when the profiles were classified based on experimental perturbations, showing strong patient-specific signals.

Relative abundances of the 255 GC-TOF metabolite intensities can be compared between samples for the same metabolite, but each metabolite intensity is not a direct reflection of abundance. The metabolites with the highest intensities include maleimide, leucine, ethanolomine, indole-3-lactate, tocopherol alpha, heptadecanoic acid, erythritol, inosine, arachidic acid, trimetryllysin, threonic acid, and 4 unknown compounds (Figure 6.8B). A combination of unsupervised random forests analysis and Partitioning Around Medoids (PAM) clustering classified the samples into two distinct groups with an OOB error of 6.45%. The top 15 most important variables separating the groups were identified using the random forests analysis and used for principal component analysis. The PCA plot revealed two distinct groups of samples with small clusters of samples by sampling date, especially for samples from patient CF14 (Figure 6.9). The negative loadings of PC1, which accounts for around 58% of the variability between these samples, are mainly from CF14 February and January collection with high levels of inosine, tocopherol alpha, trimethyllysin, arachidic acid, threonic acid, maleimide, heptadecanoic acid, indol-3-lactate, erythritol, and the three unknown compounds (Figure 6.9). On the other hand, the positive loadings of PC1 separate the rest of the samples that have high level of ethanolamine and leucine (Figure 6.9).

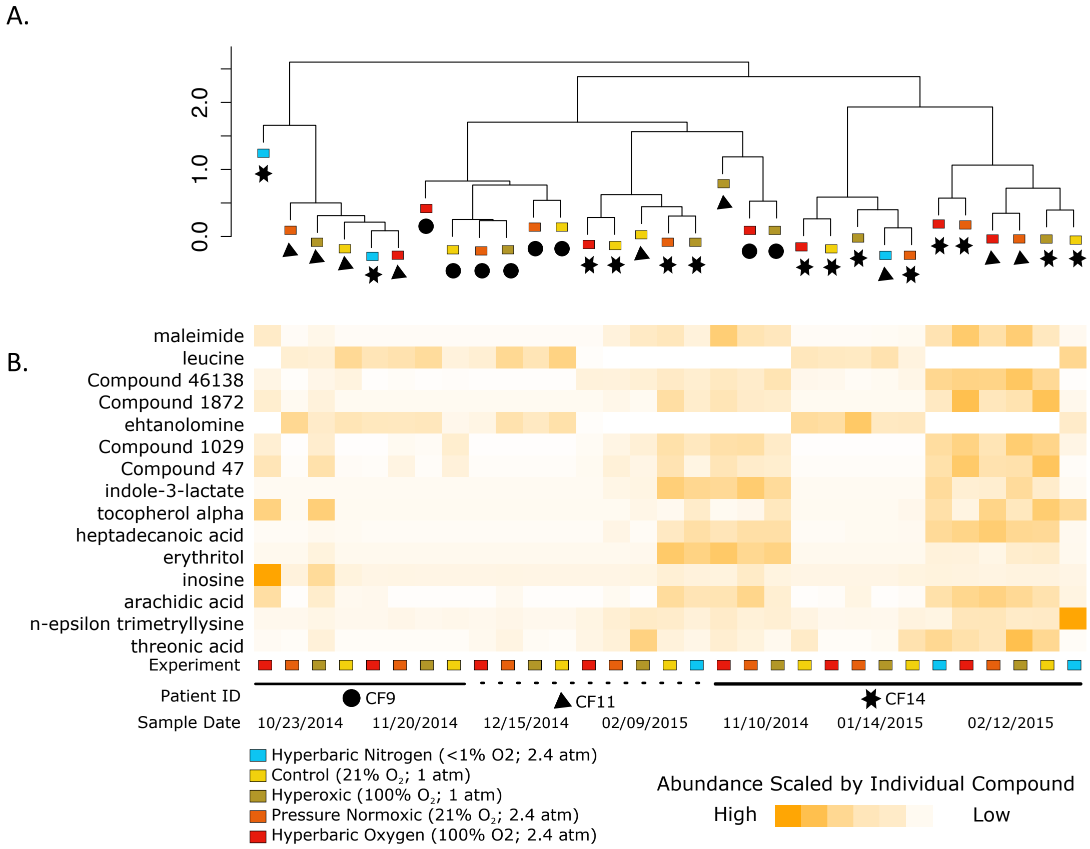


Figure 6.8: Metabolites characterization of the WinCF community. (A) The overall corrected compound intensity was used to group the samples based on Bray-Curtis similarity cluster analysis. (B) The corrected abundance based on of the top 16 metabolites. The samples were organized to observe changes due to experimental perturbations.

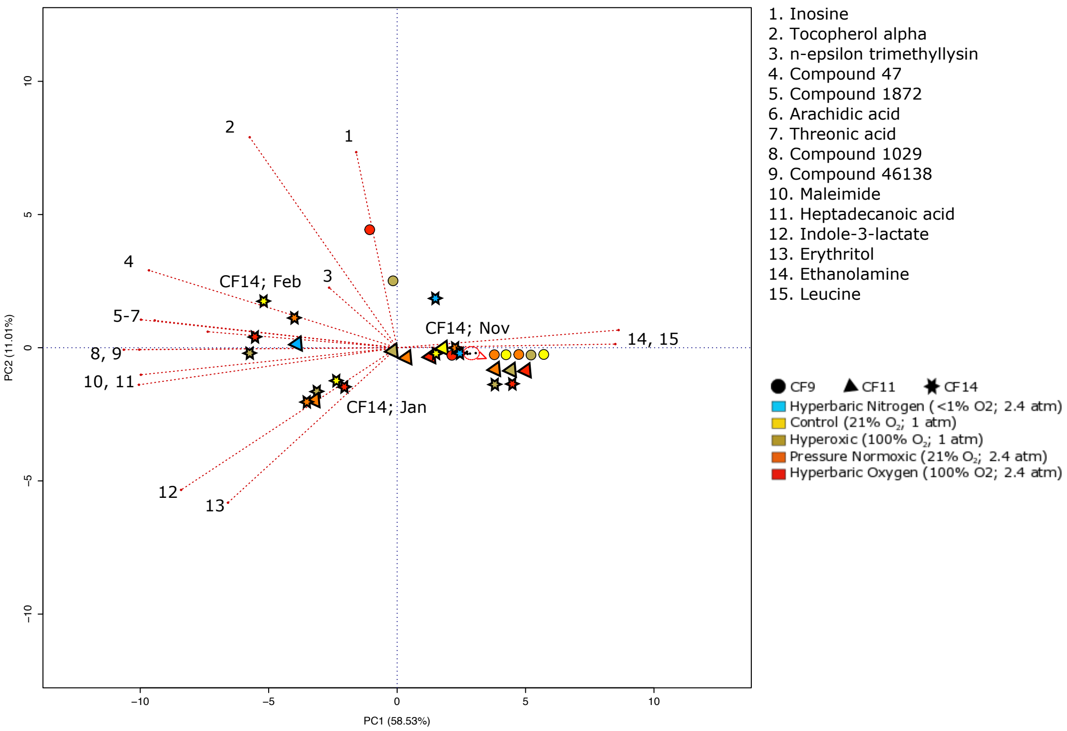


Figure 6.9: Principal component analysis of the top 15 most important metabolites identified from random forests analysis using the corrected intensity from the WinCF metabolomes. The red lines indicate vectors from PC1 and PC2.

Discussion

In addition to the commonly known CF-associated bacteria, a range of anaerobes including the genera of *Prevotella*, *Veillonella*, *Porphyromonas* and *Actinomyces* have been detected in the CF lungs through culture-dependent and culture-independent studies (Tunney et al. 2008; van der Gast et al. 2011). However, the role of these anaerobes in CF lungs pathogenicity is unclear. Several recent studies have shown that microbial fermentation processes in the CF lungs are important and potentially contribute to the decline of a patient's health status (Twomey et al. 2013; Whiteson et al. 2014; Quinn et al. 2015). The amount of fermentation precursors or metabolic products, specifically pyruvate, lactate, and putrescine varies with health status of the CF patients (Twomey et al. 2013). Utilizing the WinCF system (Quinn et al. 2015) this study explores the signatures of fermentation in 15 CF patients through visual characterization of microbial community metabolism. Metatranscriptomes and metabolomes were used to identify the molecular signatures generated by the communities.

Microbes inhabiting the CF lungs experience an oxygen gradient and the WinCF model was shown to create a steep oxygen gradient (Supplementary Figure 6.4 and (Quinn et al. 2015). On average, 60 – 65 µl of sputum-ASM mixture was inoculated into the capillary tube with an inner diameter of 1.15 mm. As one end was sealed to mimic the plugged airways (Supplementary Figure 6.2), the net amount of oxygen diffused into the artificial sputum media within the long capillary tube was limited. Due to the steep oxygen gradient within the capillary tubes, the WinCF model presented here was highly enriched with anaerobes such as *Veillonella* spp., *Prevotella* spp., *Streptococcus* spp., *Haemophilus* spp., *Fusobacterium* spp. and *Neisseria* spp. (Figure 6.5).

While *Veillonella* spp., *Prevotella* spp. and *Fusobacterium* are well known anaerobes, some clinical isolates have shown to be highly aerotolerant (Tally et al. 1975; Silva et al. 2003). Homogenization during sample processing introduced molecular oxygen into the sputum. Nevertheless these anaerobes represent the majority of the community in the WinCF model (Figure 6.5). The observation can be explained by high expression level of the genes superoxide reductase and rubrerythrin (Figure 6.6C). Superoxide reductase is an enzyme that reduce toxic superoxide radicals (O_2^-) to hydrogen peroxide (H_2O_2) while rubrerythrin is able to reduce H_2O_2 into non-reactive oxygen species (Sztukowska et al. 2002; Kovacs and Brines 2007). CF lungs are known to contain high level of oxidative stress from the host and microbial activities (Galli et al. 2012). Aerotolerance and having antioxidant abilities are one of the major adapted advantages of CF anaerobes to survive in CF lungs. Expression of these genes protects the bacterial cells from reactive oxygen species including molecular oxygen, hydrogen peroxides, and free radicals. In addition, the expression of carbon starvation protein A is highly correlated with the amount of gas production (Figure 6.7). The induction of this gene has been shown to induce resistance to other stresses including low pH and oxidative stresses (Kolter et al. 1993; Hengge-Aronis 2002), suggesting that the gas production is a response to or an effect of the low pH and high oxidative stress environment.

Within the WinCF system, the production of gas bubbles (Figure 6.2) is a result of catabolic processes such as anaerobic respiration and fermentation. The amount of gas production within the WinCF model is highly dependent on the redox states of the environment. Multiple oxidoreductases such as 3-hydroxybutyryl-CoA and 3-

hydroxyacyl-CoA dehydrogenase, dihydrolipoamide dehydrogenase of acetoin dehydrogenase, and thiol-disulfide oxidoreductase related to ResA were specifically involved (Figure 6.7). These oxidoreductases contain redox-active properties that transfer electrons from a donor to acceptor, using NAD(P) as cofactors. Specifically, thioredoxin, a disulfide oxidoreductase that is important in cell redox homeostasis is present at high level (Figure 6.6C). As thioredoxins have low redox potential (-270 to -330 mV measured in *Escherichia coli*), they are known to be efficient thiol-disulfide reductants (Krause et al. 1991; Aslund et al. 1997). The expression level of thiol-disulfide oxidoreductase is linearly related to the amount of gas production (Figure 6.7), suggesting that the thiol-disulfide reactions are crucial in regulating the redox potential that affect major proteins responsible for the gas production.

Interestingly, the production of gas bubbles within the ASM-sputum mixture can be prevented by the addition of tetrazolium dye (Supplementary Figure 6.1). The reduction of tetrazolium dye requires an electron donor and most often NADH is used. The new generations of commercially available tetrazolium dye contain a mixture of intermediate electron acceptors to facilitate the reduction of the dye, which makes it more efficient in accepting electrons from NADH (Berridge et al. 2005). The addition of tetrazolium dye in the WinCF experiment provides excessive amounts of secondary electron acceptor and perturbs the “normal” redox balance within the CF microbial communities environment. CF microbes need diverse terminal electron acceptors to complete oxidative phosphorylation. However, if an electron acceptor is unavailable, another alternative strategy is fermentation. The observation of diminished gas bubbles production in the presence of tetrazolium dye suggests that this process is a result of microbial fermentation processes. Alternatively, the added electron acceptors potentially interfere with redox active proteins such as thioredoxins that affect downstream crucial major pathways responsible for the gas production.

Acetoin dehydrogenase is composed of three components including acetoin-dependent dichlorophenolindophenol oxidoreductase (E1), dihydrolipoamide acetyltransferase (E2) and the dihydrolipoamide dehydrogenase (E3) (Peng et al. 2007). Additional evidence of fermentation is the linear relationship between all three components of acetoin dehydrogenase and the amount of gas production (Figure 6.7; Supplementary Figure 6.10). The expression of acetoin dehydrogenase is induced by the high level of acetoin (Ould Ali et al. 2001). Acetoin is a major fermentation product of pyruvate in many bacteria via the Embden-Meyerhof pathway. The fermentation of pyruvate generates high level of CO_2, which is likely to be the gas bubbles produced within the tubes. The production of acetoin is significant in low pH and the fluctuating redox CF-mucus environment in order to avoid acidification while regulating the NAD/NADH ratio. In addition, acetoin can also be used as a carbon storage metabolite for competitive advantage in a polymicrobial environment. Acetoin catabolism by acetoin dehydrogenase produces acetaldehyde (López et al. 1975; Oppermann et al. 1991). Acetaldehyde was detected at high level in a previous study using the WinCF system (Quinn et al. 2015). However, the source of acetaldehyde was unknown. This data suggests that acetaldehyde is the product of acetoin metabolism and pyruvate fermentation.

The clinically relevant conditions of the WinCF system showed that both stable and exacerbation samples produced significantly higher amounts of gas bubbles compared to samples collected during or after antibiotic treatment (Figure 6.2). This finding suggests that the presence of fermenters and/or fermentative processes in this community can be prevented through antibiotics. The observation leads to a bold hypothesis that fermenters start their activities immediately after the antibiotic stress is lifted (Figure 6.10). The subsequent accumulation of fermentative products then initiates a cascade of reactions that include (i) the use of fermentative products by other microbes (Létoffé et al. 2014; Whiteson et al. 2014; Filkins et al. 2015), (ii) the production of toxic products, where both mechanisms subsequently cause exacerbation directly or indirectly (Murray et al. 2014) (Figure 6.10).

We hypothesize that increases in oxygen concentration within the mucus plug through the application of oxygen and pressure will disrupt these anaerobic respiration and fermentation processes which contribute to CF disease pathogenesis. Our results showed that pressure with and without additional oxygen reduced the amount of fermentation (signatures of gas production) within the WinCF model (Figure 6.3). The comparable amounts of total RNA recovered per volume of liquid within the capillary tubes (Supplementary Figure 6.7) suggests that pressure did not affect the overall biomass of the microbial communities. From the metatranscriptomics analysis, it is likely that the anaerobes were responding to pressure through redox homeostasis (see discussion above). Pressure affects redox balance through (i) changing the membrane fluidity and hence changes in the ion permeability and microbial bioenergetics processes, (ii) affecting the transmembrane protein that changes ions and metabolites transport across the membrane, altering microbial metabolism, or (iii) affecting bacterial motility, preventing the stratification of the communities along the oxygen and chemical gradients within the capillary tube (Zobell and Cobet 1962; Meganathan and Marquis 1973; Welch et al. 1993). Many of these mechanisms have been well studied in the model mesophilic bacterium *Escherichia coli* by subjecting the bacteria to pressure up to 100MPa (Bartlett 2002). The pressure exerted on the microbial community in our study (2.4 atm; ~ 48 feet water; 17 PSI; 142 kPa) is significantly less than that used in the *E. coli* experiments and therefore, more likely to induce microbial stress response without impeding essential metabolic processes. The correlation of the general stress response element indicated by the expression of carbon starvation protein A with the amount of gas production provided a strong evidence of the stress response induction (Figure 6.7A).

In all conditions, ferroxidase, an enzyme that catalyzes the oxidation of of Fe(II) to Fe(III), was expressed at high level within the capillary culture model (Figure 6.6C). This characteristic may be an adapted trait of the anaerobes as the capillary tube is highly anoxic and anoxic microenvironment has been shown to favor the maintenance of bioavailable Fe(II) (Worlitzsch et al. 2002). In addition, Fe(II) has been found in high level and dominating the iron pool in CF lungs while correlated negatively with the lung function of CF patients (Hunter et al. 2013). Iron availability is critically important in disease pathogenesis. The availability of Fe(III) is often limited due to the chelation by host immune proteins lactoferrin and the competition with other bacteria. However, the ability to utilize Fe(II) provides these anaerobes an advantage to survive and persist in the CF lungs. Collectively, the addition of Fe(II) chelating therapy may be a potential CF therapeutic in controlling the effect of anaerobes in CF lung disease and this has been proposed to control biofilm development in *P. aeruginosa* infection (Hunter et al. 2013). A potential chelating agent includes the Chinese herbal medicine *Lingusticum wallichi*

Franchat, which contain two active ingredients, tetramethylpyrazine and ferulic acid that have been shown to chelate iron and decrease the level of hydroxyl radicals induced damage (Zhang et al. 2003).

Additionally, the high expression level of acyl carrier protein (ACP) is significant. ACP is important in the biosynthesis of fatty acids, phospholipids, endotoxins, glycolipids, and signaling molecules for growth and pathogenesis of many bacteria (Byers and Gong 2007). Pantothenamide antimicrobial compounds have been shown to inhibit fatty acid biosynthesis through the formation and accumulation of inactive ACP (Zhang et al. 2004). The high expression level of ACP by these communities (Figure 6.6C) provides a potential target for the pantothenamide class of antimicrobial compounds (Zhang et al. 2004).

It is worth noting that the expression profiles that were examined here were taken 72 – 96 hours into the WinCF culture. The gene expression profiles presented in this study is depicting microbial responses to an established CF lung-like environment within the capillary tube through the first two days of experiment, including low pH and high oxidative stress (Figure 6.10). The anaerobes also responded to these conditions through regulating the redox potential to achieve fermentation while alleviating the effect of pH and oxidative stress through multiple routes. The WinCF culture system can be improved to closer represent the plugged CF airways by reducing the length of the capillary tube or inoculants to a more appropriate inner diameter to length ratio. An experimental and mathematical model has shown that the inner diameter to length ratio according to sheep airways is predicted at 1:30 (Lipsett 2002). In this experiment, the inoculants reached

1:60 (1.15 mm inner diameter to 65 mm length of the tube) and therefore created an environment that is highly enriched for anaerobic communities.

Conclusion

The metabolic activities of microbes, specifically the anaerobes, can be easily examined and amplified through simple inoculation of sputum into commonly used blood capillary tubes. Anaerobes played a significant role in the health of CF patients even though present in low abundance within the sputum or mucus plug. This study showed that fermentative signatures from the anaerobes are significantly represented during patient's stability and exacerbation events. The fermentation-associated pathway(s) is highly sensitive to antibiotics clinically, as well as additional electron acceptors and pressure experimentally. A summary of the major findings and the proposed role of anaerobes in CF lungs are presented in Figure 6.10. The enriched anaerobes within the WinCF culture model are highly adapted to the highly reduced, low pH, and high ROS environment. The ability to regulate cellular and environmental redox potentials through mixed-acid fermentation processes provides additional advantage. Subsequent pyruvate fermentation generated high level of CO_2, which is visible through the WinCF culture model. Downstream production of acetoin and acetaldehyde as byproducts allowed the anaerobes to neutralize acidification and further regulate the ratio of NAD:NADH. Most importantly, the preference of anaerobes for Fe(II) and consistent high expression of ACP provided two potential drug targets to alleviate the effect of anaerobes in CF lungs.

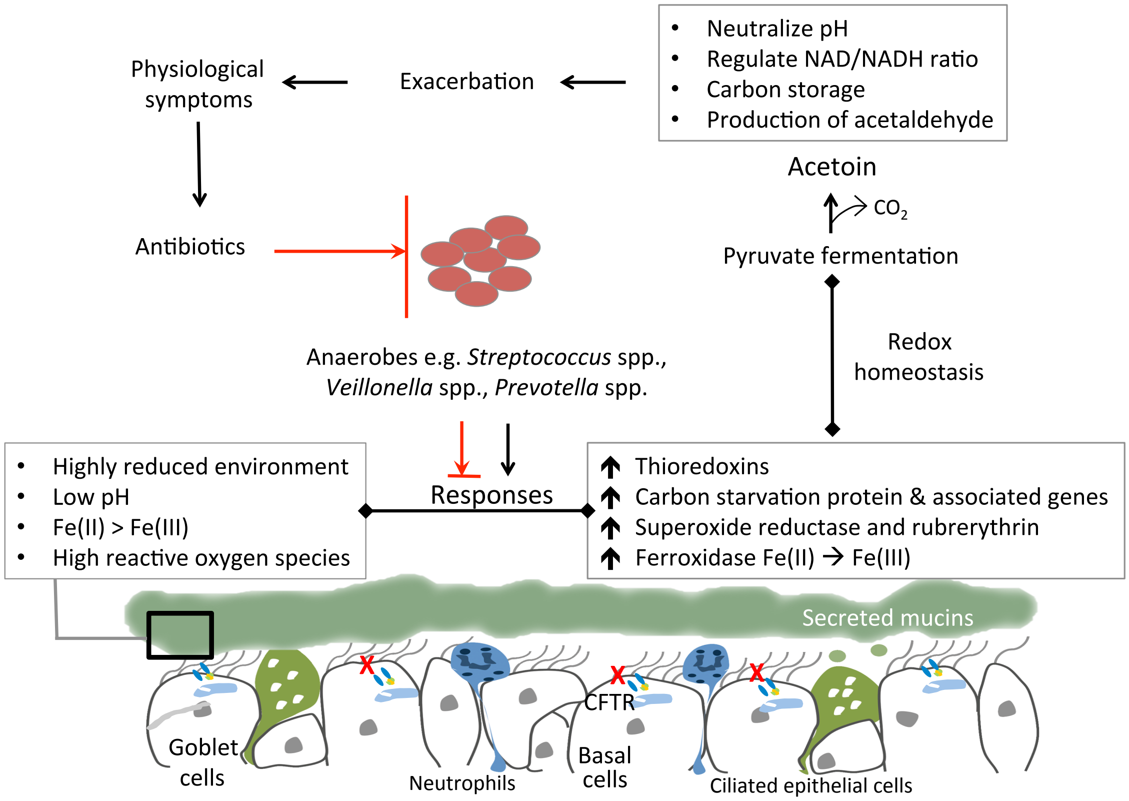

Physiological symptoms
Exacerbation
Neutralize pH
Regulate NAD/NADH ratio
Carbon storage
Production of acetaldehyde
Acetoin
CO_2
Pyruvate fermentation
Antibiotics
Anaerobes e.g. *Streptococcus* spp., *Veillonella* spp., *Prevotella* spp.
Redox homeostasis
Highly reduced environment
Low pH
Fe(II) > Fe(III)
High reactive oxygen species
Responses
Thioredoxins
Carbon starvation protein & associated genes
Superoxide reductase and rubrerythrin
Ferroxidase Fe(II) → Fe(III)
Secreted mucins
CFTR
Goblet cells
Neutrophils
Basal cells
Ciliated epithelial cells

Figure 6.10: A proposed model of the role of anaerobes in CF lungs based on the WinCF transcriptomics analysis. Black lines indicate responses without the pressure of antibiotics. Red lines indicate inhibition due to antibiotics. The mucus plugs within the CF lung is a highly reduced environment with high level of reactive oxygen species. The lack of CFTR function and microbial activities further reduced the pH within the mucus. Anaerobes respond to the "CF lung" environment through regulating genes such as thioredoxins/thiol-dilsulfide oxidoreductase, carbon starvation associated genes, superoxide reductase, rubrerythrin and ferroxidase. Thioredoxins served dual purposes, including antioxidant activity and cell redox homeostasis. In order to avoid acidification and regulate NAD/NADH ratio, anaerobes induced pyruvate fermentation, creating excessive amount of acetoin. High level of acetoin induces the break down of acetoin through acetoin dehydrogenase, producing acetaldehyde. During pyruvate fermentation, CO_2 production served as an indicator in the WinCF model. Accumulation of fermentative products results in a cascade of reactions that subsequently cause exacerbation event in CF patients. The metabolic activities of anaerobes are sensitive to commonly used antibiotics. However, their metabolic activities resumed as soon as the antibiotic stress is lifted.

Acknowledgments

This work was supported by the National Institute of Health (1 R01 GM095384-01) grant awarded to Forest Rohwer, and the Cystic Fibrosis Research, Inc. (15-EN3, 11-006D) grants awarded to Yan Wei Lim and Robert Quinn, respectively. We thank Sean Benler in helping during the process of library preparation and many thoughtful suggestions. We thank The Oxyheal Health Group and University of California, San Diego Hyperbaric Medicine Group for the use of hyperbaric chamber and their technical support.

Chapter 6, in full, is in preparation for journal submission. Yan Wei Lim, Katrine Whiteson, Barbara Bailey, Peter Salamon, Ben Felts, Andreas Haas, Robert Quinn, Mark Hatay, Douglas Conrad, Robert Edwards, and Forest Rohwer; 2015. The dissertation author was the primary investigator and author of this paper.

References

Abubucker S, Segata N, Goll J, Schubert A, Rodriguez-Mueller B, Zucker J, team tHMPMR, Schloss P, Gevers D, Mitreva M, Huttenhower C (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLos Comput Biol. doi: 10:10.1371/journal.pcbi.1002358

Alvarez-Ortega C, Harwood CS (2007) Responses of Pseudomonas aeruginosa to low oxygen indicate that growth in the cystic fibrosis lung is by aerobic respiration. Mol Microbiol 65:153–165. doi: 10.1111/j.1365-2958.2007.05772.x

Aslund F, Berndt KD, Holmgren A (1997) Redox potentials of glutaredoxins and other thiol-disulfide oxidoreductases of the thioredoxin superfamily determined by direct protein-protein redox equilibria. J Biol Chem 272:30780–30786.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. doi: 10.1089/cmb.2012.0021

Bartlett DH (2002) Pressure effects on in vivo microbial processes. Biochim Biophys Acta BBA - Protein Struct Mol Enzymol 1595:367–381. doi: 10.1016/S0167-4838(01)00357-0

Berridge MV, Herst PM, Tan AS (2005) Tetrazolium dyes as tools in cell biology: new insights into their cellular reduction. Biotechnol Annu Rev 11:127–152. doi: 10.1016/S1387-2656(05)11004-7

Bhupathiraju VK, Hernandez M, Landfear D, Alvarez-Cohen L (1999) Application of a tetrazolium dye as an indicator of viability in anaerobic bacteria. J Microbiol Methods 37:231–243.

Breiman L (2001) Random Forests. Mach Learn 45:5–32. doi: 10.1023/A:1010933404324

Brown OR, Silverberg RG, Huggett DO (1968) Synergism between hyperoxia and antibiotics for Pseudomonas aeruginosa. Appl Microbiol 16:260–262.

Byers DM, Gong H (2007) Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. Biochem Cell Biol 85:649–662. doi: 10.1139/o07-109

Cohen TS, Prince A (2012) Cystic fibrosis: a mucosal immunodeficiency syndrome. Nat Med 18:509–519. doi: 10.1038/nm.2715

Cowley ES, Kopf SH, LaRiviere A, Ziebis W, Newman DK (2015) Pediatric Cystic Fibrosis sputum can be chemically dynamic, anoxic, and extremely reduced due to hydrogen sulfide formation. mBio 6:e00767–15.

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044.

Filkins LM, Graber JA, Olson DG, Dolben EL, Lynd LR, Bhuju S, O'Toole GA (2015) Coculture of Staphylococcus aureus with Pseudomonas aeruginosa Drives S. aureus towards Fermentative Metabolism and Reduced Viability in a Cystic Fibrosis Model. J Bacteriol 197:2252–2264. doi: 10.1128/JB.00059-15

Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash MI, Ramsey BW, Rosenstein BJ, Smith AI, Wohl ME (1994) Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with Cystic Fibrosis. The New England Journal of Medicine 331:637–642.

Fung C, Naughton S, Turnbull L, Tingpej P, Rose B, Arthur J, Hu H, Harmer C, Harbour C, Hassett DJ, Whitchurch CB, Manos J (2010) Gene expression of Pseudomonas aeruginosa in a mucin-containing synthetic growth medium mimicking cystic fibrosis lung sputum. J Med Microbiol 59:1089–1100.

Galli F, Battistoni A, Gambari R, Pompella A, Bragonzi A, Pilolli F, Iuliano L, Piroddi M, Dechecchi MC, Cabrini G (2012) Oxidative stress and antioxidant therapy in cystic fibrosis. Biochim Biophys Acta BBA - Mol Basis Dis 1822:690–713. doi: 10.1016/j.bbadis.2011.12.012

Guss AM, Roeselers G, Newton ILG, Young CR, Klepac-Ceraj V, Lory S, Cavanaugh CM (2011) Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. ISME J 5:20–29.

Hare NJ, Soe CZ, Rose B, Harbour C, Codd R, Manos J, Cordwell SJ (2012) Proteomics of Pseudomonas aeruginosa Australian epidemic strain 1 (AES-1) cultured under conditions mimicking the Cystic Fibrosis lung reveals increased iron acquisition via the siderophore pyochelin. J Proteome Res 11:776–795. doi: 10.1021/pr200659h

Hengge-Aronis R (2002) Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. Microbiol Mol Biol Rev MMBR 66:373–395, table of contents.

Hoboth C, Hoffmann R, Eichner A, Henke C, Schmoldt S, Imhof A, Heesemann J, Hogardt M (2009) Dynamics of adaptive microevolution of hypermutable

Pseudomonas aeruginosa during chronic pulmonary infection in patients with Cystic Fibrosis. J Infect Dis 200:118 –130. doi: 10.1086/599360

Hunter RC, Asfour F, Dingemans J, Osuna BL, Samad T, Malfroot A, Cornelis P, Newman DK (2013) Ferrous iron is a significant component of bioavailable iron in Cystic Fibrosis airways. mBio 4:e00557–13. doi: 10.1128/mBio.00557-13

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28:2223–2230. doi: 10.1093/bioinformatics/bts429

Kolter R, Siegele DA, Tormo A (1993) The stationary phase of the bacterial life cycle. Annu Rev Microbiol 47:855–874. doi: 10.1146/annurev.mi.47.100193.004231

Kovacs JA, Brines LM (2007) Understanding how the thiolate sulfur contributes to the function of the non-heme iron enzyme superoxide reductase. Acc Chem Res 40:501–509. doi: 10.1021/ar600059h

Kranke P, Bennet M, Martyn - St James M, Schnabel A, Debus S (2012) Hyperbaric oxygen therapy for treating chronic wounds.

Krause G, Lundström J, Barea JL, Pueyo de la Cuesta C, Holmgren A (1991) Mimicking the active site of protein disulfide-isomerase by substitution of proline 34 in Escherichia coli thioredoxin. J Biol Chem 266:9494–9500.

Létoffé S, Audrain B, Bernier SP, Delepierre M, Ghigo J-M (2014) Aerial exposure to the bacterial volatile compound trimethylamine modifies antibiotic resistance of physically separated bacteria by raising culture medium pH. mBio 5:e00944–00913. doi: 10.1128/mBio.00944-13

Liaw A, Wiener M (2002) Breiman and Cutler's Random Forests for Classification and Regression. Rnews Vol. 2/3:18–22.

Lim YW, Evangelista JS, Schmieder R, Bailey B, Haynes M, Furlan M, Maughan H, Edwards R, Rohwer F, Conrad D (2014a) Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. J Clin Microbiol 52:425–437. doi: 10.1128/JCM.02204-13

Lim YW, Haynes M, Furlan M, Robertson CE, Harris JK, Rohwer F (2014b) Purifying the impure: sequencing metagenomes and metatranscriptomes from complex animal-associated samples. J Vis Exp JoVE. doi: 10.3791/52117

Lim YW, Schmieder R, Haynes M, Furlan M, Matthews TD, Whiteson K, Poole SJ, Hayes CS, Low DA, Maughan H, Edwards R, Conrad D, Rohwer F (2013) Mechanistic model of Rothia mucilaginosa adaptation toward persistence in the

CF lung, based on a genome reconstructed from metagenomic data. PloS One 8:e64285. doi: 10.1371/journal.pone.0064285

Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. J Cyst Fibros Off J Eur Cyst Fibros Soc. doi: 10.1016/j.jcf.2012.07.009

Lipsett J (2002) Analysis of the conducting airway system in the lung: a new method combining morphometry with mathematical modeling for airway classification. Anat Rec 266:51–57.

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clin Microbiol Rev 23:299 –323. doi: 10.1128/CMR.00068-09

López JM, Thomas B, Rehbein H (1975) Acetoin degradation in Bacillus subtilis by direct oxidative cleavage. Eur J Biochem 57:425–430. doi: 10.1111/j.1432-1033.1975.tb02317.x

Mathieu D (2006) Handbook on hyperbaric medicine. Springer, New York

Meganathan R, Marquis RE (1973) Loss of bacterial motility under pressure. Nature 246:525–527.

Muhvich KH, Park MK, Myers RA, Marzella L (1989) Hyperoxia and the antimicrobial susceptibility of Escherichia coli and Pseudomonas aeruginosa. Antimicrob Agents Chemother 33:1526–1530.

Murray JL, Connell JL, Stacy A, Turner KH, Whiteley M (2014) Mechanisms of synergy in polymicrobial infections. J Microbiol Seoul Korea 52:188–199. doi: 10.1007/s12275-014-4067-3

O’Sullivan BP, Freedman SD (2009) Cystic fibrosis. The Lancet 373:1891–1904. doi: 10.1016/S0140-6736(09)60327-5

Oppermann FB, Schmidt B, Steinbüchel A (1991) Purification and characterization of acetoin:2,6-dichlorophenolindophenol oxidoreductase, dihydrolipoamide dehydrogenase, and dihydrolipoamide acetyltransferase of the Pelobacter carbinolicus acetoin dehydrogenase enzyme system. J Bacteriol 173:757–767.

Ould Ali N, Bignon J, Rapoport G, Debarbouille M (2001) Regulation of the acetoin catabolic pathway is controlled by Sigma L in Bacillus subtilis. J Bacteriol 183:2497–2504. doi: 10.1128/JB.183.8.2497-2504.2001

Palmer KL, Aye LM, Whiteley M (2007) Nutritional cues control Pseudomonas aeruginosa multicellular behavior in Cystic Fibrosis sputum. J Bacteriol 189:8079–8087. doi: 10.1128/JB.01138-07

Peng L, He Z, Chen W, Holzman IR, Lin J (2007) Effects of butyrate on intestinal barrier function in a Caco-2 cell monolayer model of intestinal barrier. Pediatr Res 61:37–41. doi: 10.1203/01.pdr.0000250014.92242.f3

Quinn RA, Whiteson K, Lim Y-W, Salamon P, Bailey B, Mienardi S, Sanchez SE, Blake D, Conrad D, Rohwer F (2015) A Winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. ISME J 9:1024–1038. doi: 10.1038/ismej.2014.234

Reid DW, Carroll V, O'May C, Champion A, Kirov SM (2007) Increased airway iron as a potential factor in the persistence of Pseudomonas aeruginosa infection in cystic fibrosis. Eur Respir J 30:286–292. doi: 10.1183/09031936.00154006

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863 –864. doi: 10.1093/bioinformatics/btr026

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814. doi: 10.1038/nmeth.2066

Sibley CD, Parkins MD, Rabin HR, Duan K, Norgaard JC, Surette MG (2008) A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in Cystic Fibrosis patients. Proc Natl Acad Sci 105:15070 –15075. doi: 10.1073/pnas.0804326105

Silva VL, Carvalho M a. R, Nicoli JR, Farias LM (2003) Aerotolerance of human clinical isolates of Prevotella spp. J Appl Microbiol 94:701–707.

Sriramulu DD, Lünsdorf H, Lam JS, Römling U (2005) Microcolony formation: a novel biofilm model of Pseudomonas aeruginosa for the cystic fibrosis lung. J Med Microbiol 54:667–676. doi: 10.1099/jmm.0.45969-0

Sztukowska M, Bugno M, Potempa J, Travis J, Kurtz DM (2002) Role of rubrerythrin in the oxidative stress response of Porphyromonas gingivalis. Mol Microbiol 44:479–488.

Tally FP, Stewart PR, Sutter VL, Rosenblatt JE (1975) Oxygen tolerance of fresh clinical anaerobic bacteria. J Clin Microbiol 1:161–164.

Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS (2008) Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 177:995–1001.

Twomey KB, Alston M, An S-Q, O'Connell OJ, McCarthy Y, Swarbreck D, Febrer M, Dow JM, Plant BJ, Ryan RP (2013) Microbiota and metabolite profiling reveal specific alterations in bacterial community structure and environment in the Cystic Fibrosis airway during exacerbation. PLoS ONE 8:e82432.

Van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW, Carroll MP, Parkhill J, Bruce KD (2011) Partitioning core and satellite taxa from within Cystic Fibrosis lung bacterial communities. ISME J 5:780–791.

Welch TJ, Farewell A, Neidhardt FC, Bartlett DH (1993) Stress response of Escherichia coli to elevated hydrostatic pressure. J Bacteriol 175:7170–7177.

Whiteson KL, Meinardi S, Lim YW, Schmieder R, Maughan H, Quinn R, Blake DR, Conrad D, Rohwer F (2014) Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. ISME J 8:1247–1258. doi: 10.1038/ismej.2013.229

Worlitzsch D, Tarran R, Ulrich M, Schwab U, Cekici A, Meyer KC, Birrer P, Bellon G, Berger J, Weiss T, Botzenhart K, Yankaskas JR, Randell S, Boucher RC, Döring G (2002) Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. J Clin Invest 109:317–325.

Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp RM, Sittler T (2011) Faster and more accurate eequence alignment with SNAP.

Zarei S, Mirtar A, Rohwer F, Conrad DJ, Theilmann RJ, Salamon P (2012) Mucus distribution model in a lung with Cystic Fibrosis. Comput Math Methods Med. doi: 10.1155/2012/970809

Zhang Y-M, Frank MW, Virga KG, Lee RE, Rock CO, Jackowski S (2004) Acyl carrier protein is a cellular target for the antibacterial action of the pantothenamide class of pantothenate antimetabolites. J Biol Chem 279:50969–50975.

Zhang Z, Wei T, Hou J, Li G, Yu S, Xin W (2003) Iron-induced oxidative damage and apoptosis in cerebellar granule cells: attenuation by tetramethylpyrazine and ferulic acid. Eur J Pharmacol 467:41–47.

Zobell CE, Cobet AB (1962) Growth, reproduction, and death rates of Escherichia coli at increased hydrostatic pressures. J Bacteriol 84:1228–1236.

Appendix for Chapter 6

Supplementary Table 6.1: Summary of microbial metagenomic sequence data

Sample ID	Health Status	Total Sequences	Total Hits	Bacteria[1]	Unknown[2]
CF1	Ex	29,316	24,506	14,223 (49%)	4,810
CF9 (July)	Pt	14,689	11,857	7,057 (48%)	2,832
CF9 (Nov)	St	87,321	64,170	43,552 (50%)	23,151
CF8 (August)	Pt	167,448	154,812	133,232 (80%)	12,636
CF8 (Nov)	Ex	784,143	682,956	580,850 (74%)	101,187
CF2	Ex	28,128	19,750	1,788 (6%)	8,378
CF10	Ex	82,337	52,298	8,630 (11%)	30,039
CF11 (August)	Pt	583,394	516,628	412,705 (71%)	66,766
CF11 (Sept)	Ex	113,763	93,139	59,879 (53%)	20,624
CF3	Pt	36,368	25,585	10,139 (28%)	10,783
CF12	St	77,933	65,466	46,596 (60%)	12,467
CF14	Ex	41,194	28,445	18,944 (46%)	12,749
CF13	St	78,901	60,989	41,254 (52%)	17,912
CF4	Ex	17,383	12,678	839 (5%)	4,705
CF5 (Oct)	Tr	16,038	11,598	354 (2%)	4,440
CF5 (Nov)	Pt	20,158	14,336	2,264 (11%)	5,822
CF7	Ex	336,845	322,801	12,461 (4%)	14,044
CF6	Pt	59,491	51,174	35,754 (60%)	8,317

[1] Total number of hits based on only Read 1 of the paired-end reads, BLASTn comparison against the NCBI nucleotide (nt) database. Percentage was calculated against the “Total Sequences”.

[2] Unknown sequences are sequences that do not have any similarity against the NCBI nucleotide (nt) database based on BLASTn comparison at ≥ 60% of sequence length coverage and ≥ 40% nucleotide identity.

Supplementary Table 6.2: Presence of microbial communities that produce gas in the WinCF system across all CF patients. No Exacerbation or Stable samples were collected from patient CF3. Treatment sample from patient CF3 yielded an average of 1.6% of gas bubbles in the WInCF system. Ex: Exacerbation; St: Stable

Sample ID	**Health Status**	**Gas Production**
CF1	Ex	yes
CF9	St	yes
CF8	Ex	yes
CF2	Ex	yes
CF10	Ex	yes
CF11	Ex	yes
CF12	St	slightly
CF14	Ex	slightly
CF13	St	no
CF4	Ex	yes
CF4	St	yes
CF15	St	yes
CF5	Ex	yes
CF7	Ex	yes

Supplementary Table 6.3: The effect of antibiotics on the WinCF community metabolisms and gas bubbles production. Phosphate-buffered saline (PBS) was used as the control of fluid addition to the top of the mucus within the capillary tubes (See Supplementary Figure 3).

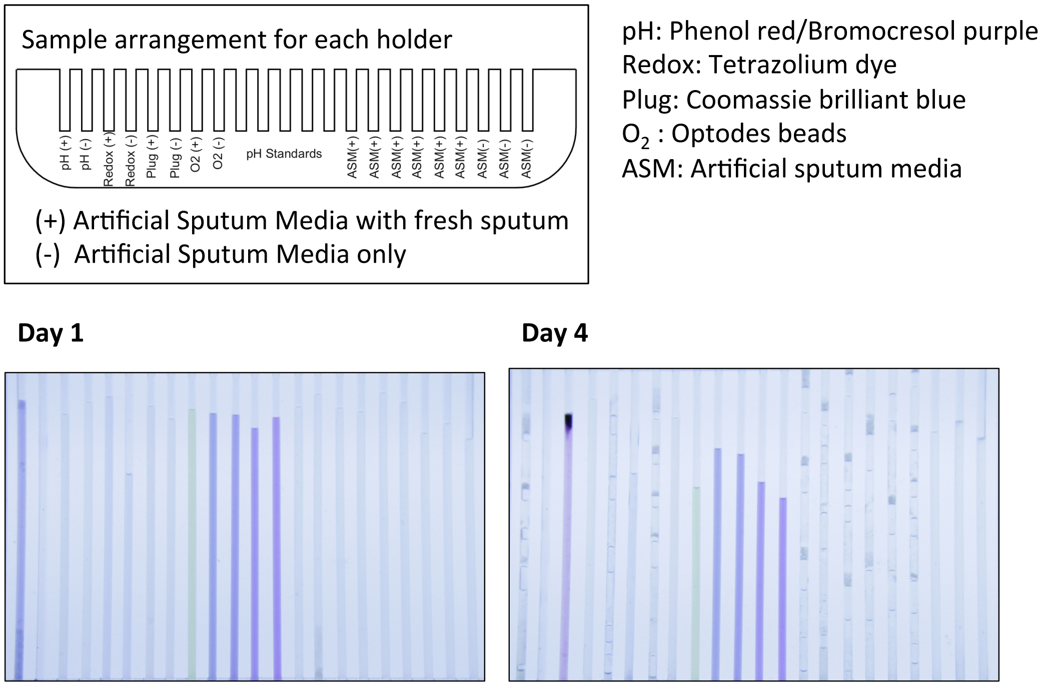
	Percentage of tubes occupied by gas (%)				
	Hyperbaric Oxygen	Hyperbaric Normoxic	Hyperoxic	Control	Hyperbaric Nitrogen
CF11	46.3	46.3	66.3	60.4	n/a
+Ciprofloxacin	18.1	4.6	17.8	17.6	n/a
CF14	3.5	6.1	23.3	20.4	3.1
+PBS	0.0	0.8	8.5	14.4	0.0
+Tobramycin	0.0	0.0	0.0	0.0	n/a
CF14	1.0	2.0	10.4	28.7	0.0
+PBS	0.0	1.9	22.4	17.9	0.0
+Ciprofloxacin	0.0	0.0	1.0	0.0	0.0
CF5	11.8	23.9	49.2	58.9	23.7
+PBS	10.5	17.7	51.4	43.6	3.9
+Tobramycin	0.0	0.8	4.5	4.5	0.0

Supplementary Table 6.4: Metatranscriptomics data characteristics

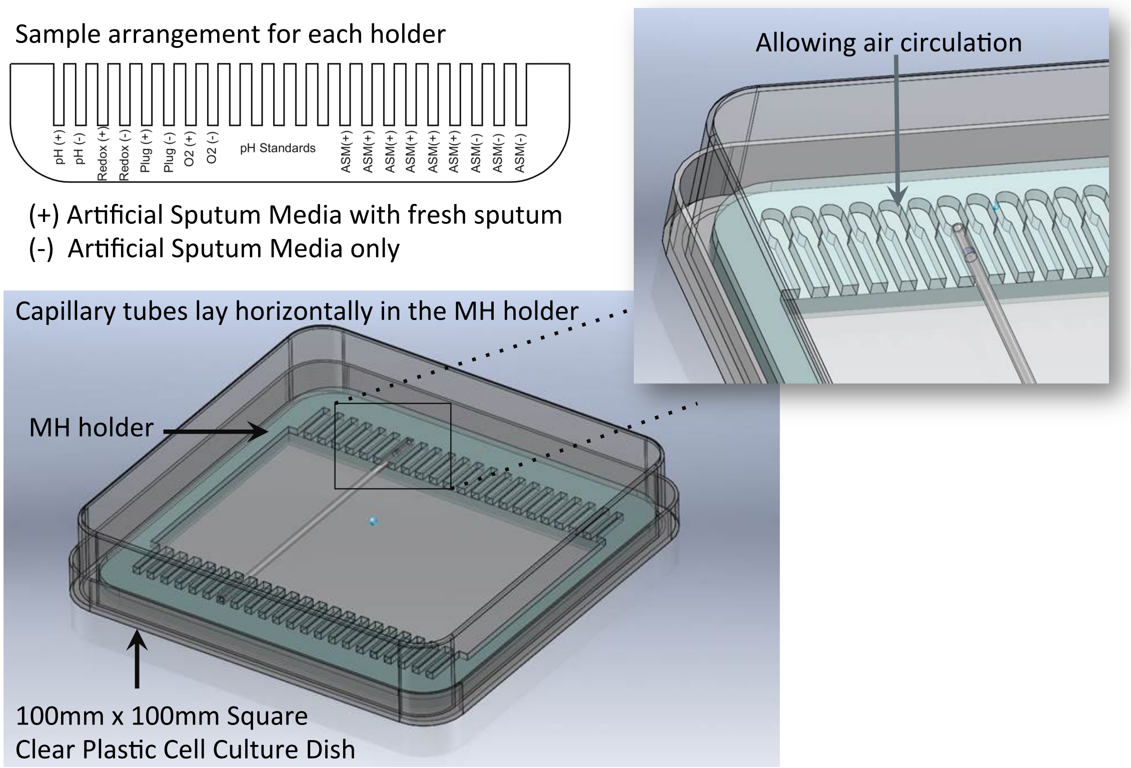
Patient ID	Sample Date	Sample Source	Experiment	% Gas	Raw reads (R1)	Raw reads (R2)	Filtered (R1)	Filtered (R2)
(St) CF9	10/23/14	Capillary tubes	Hyperbaric Oxygen	34.3	1,167,008	1,167,008	999,937	1,072,105
CF9	10/23/14	Capillary tubes	Hyperbaric Normoxic	30.2	863,254	863,254	712,140	788,137
CF9	10/23/14	Capillary tubes	Hyperoxic	41.0	780,661	780,661	649,750	720,885
CF9	10/23/14	Capillary tubes	Control	39.7	1,277,290	1,277,290	1,064,742	1,191,137
(St) CF9	11/20/14	Capillary tubes	Hyperbaric Oxygen	30.8	1,113,318	1,113,318	691,034	793,638
CF9	11/20/14	Capillary tubes	Hyperbaric Normoxic	24.4	1,639,272	1,639,272	899,244	1,346,224
CF9	11/20/14	Capillary tubes	Hyperoxic	36.1	1,092,248	1,092,248	718,973	814,993
CF9	11/20/14	Capillary tubes	Control	33.0	2,326,293	2,326,293	1,440,474	2,052,767
(Ex) CF11	12/15/14	Capillary tubes	Hyperbaric Oxygen	46.3	832,223	832,223	672,416	763,368
CF11	12/15/14	Capillary tubes	Hyperbaric Normoxic	46.3	1,270,126	1,270,126	975,427	1,136,146
CF11	12/15/14	Capillary tubes	Hyperoxic	66.3	1,200,693	1,200,693	989,298	1,090,087
CF11	12/15/14	Capillary tubes	Control	60.4	1,180,331	1,180,331	865,625	966,934
(Ex) CF11	2/9/15	Capillary tubes	Hyperbaric Oxygen	7.4	751,373	751,373	365,194	449,269
CF11	2/9/15	Capillary tubes	Hyperbaric Normoxic	4.5	1,061,056	1,061,056	222,617	304,010
CF11	2/9/15	Capillary tubes	Hyperoxic	12.8	1,286,104	1,286,104	478,822	701,360
CF11	2/9/15	Capillary tubes	Control	12.4	1,081,704	1,081,704	337,793	410,401
CF11	2/9/15	Capillary tubes	Hyperbaric Nitrogen	2.0	1,192,993	1,192,993	533,758	611,308
(Ex) CF14	11/10/15	Capillary tubes	Hyperbaric Oxygen	36.9	1,252,550	1,252,550	945,614	984,863
CF14	11/10/15	Capillary tubes	Hyperbaric Normoxic	36.2	3,820,714	3,820,714	2,425,368	3,563,676
CF14	11/10/15	Capillary tubes	Hyperoxic	52.5	1,082,452	1,082,452	883,398	959,022
CF14	11/10/15	Capillary tubes	Control	47.0	682,926	682,926	517,240	573,991
(Tr) CF14	1/14/15	Capillary tubes	Hyperbaric Oxygen	3.5	654,121	654,121	576,314	553,208
CF14	1/14/15	Capillary tubes	Hyperbaric Normoxic	6.1	1,530,717	1,530,717	1,175,054	1,383,996
CF14	1/14/15	Capillary tubes	Hyperoxic	23.3	1,670,081	1,670,081	1,103,763	1,559,588
CF14	1/14/15	Capillary tubes	Control	20.4	596,611	596,611	402,104	476,208
CF14	1/14/15	Capillary tubes	Hyperbaric Nitrogen	3.1	2,115,033	2,115,033	1,408,123	1,956,701
(Ex) CF14	2/12/15	Capillary tubes	Hyperbaric Oxygen	1.0	950,239	950,239	581,484	644,186
CF14	2/12/15	Capillary tubes	Hyperbaric Normoxic	2.0	679,543	679,543	506,048	583,888
CF14	2/12/15	Capillary tubes	Hyperoxic	10.4	1,202,108	1,202,108	736,822	825,940
CF14	2/12/15	Capillary tubes	Control	28.7	652,210	652,210	366,286	459,873
CF14	2/12/15	Capillary tubes	Hyperbaric Nitrogen	0	889,092	889,092	643,778	745,829
CF9	10/23/14	Sputum	N/A		2,515,843	2,515,843	421,861	1,604,894
CF9	11/20/14	Sputum	N/A		1,740,291	1,740,291	229,834	1,194,769
CF14	1/14/15	Sputum	N/A		891,202	891,202	36,753	779,049
CF14	2/12/15	Sputum	N/A		3,143,606	3,143,606	332,540	2,831,450

Supplementary Table 6.4: Metatranscriptomics data characteristics (continued)

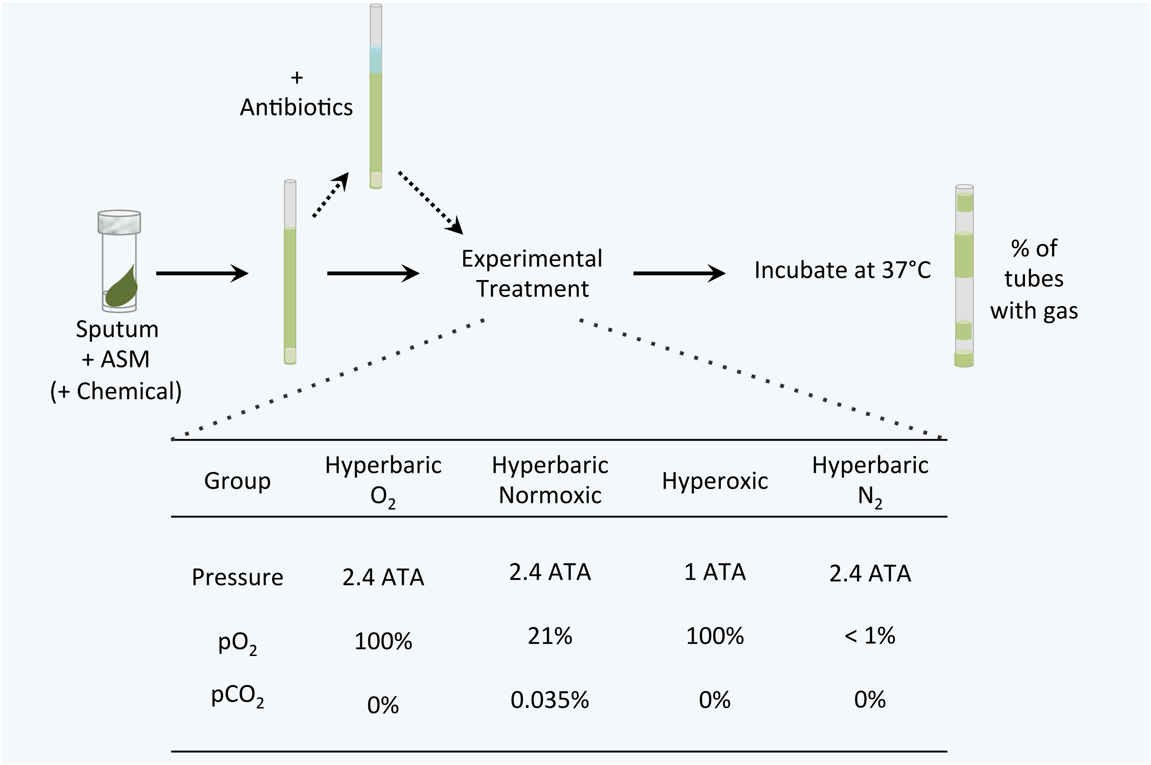
Patient ID	Sample Date	Sample Source	Experiment	non-rRNA (R1)	%	non-rRNA (R2)	%	% R1 Hits to KEGG	% R1 Hits to CFaDB
(St) CF9	10/23/14	Capillary tubes	Hyperbaric Oxygen	1,118,742	97.3	1,028,619	89.6	3.90	70.5
CF9	10/23/14	Capillary tubes	Hyperbaric Normoxic	777,309	92.2	696,385	82.8	3.79	63.7
CF9	10/23/14	Capillary tubes	Hyperoxic	741,329	96.8	686,163	89.9	5.21	60.8
CF9	10/23/14	Capillary tubes	Control	936,347	74.3	796,499	63.4	3.87	63.7
(St) CF9	11/20/14	Capillary tubes	Hyperbaric Oxygen	729,715	83.6	585,445	68.7	3.75	62.8
CF9	11/20/14	Capillary tubes	Hyperbaric Normoxic	843,134	59.1	392,547	27.8	2.43	29.4
CF9	11/20/14	Capillary tubes	Hyperoxic	743,653	83.1	611,428	69.6	4.27	62.0
CF9	11/20/14	Capillary tubes	Control	1,201,005	55.3	566,891	26.4	2.20	29.5
(Ex) CF11	12/15/14	Capillary tubes	Hyperbaric Oxygen	710,593	88.2	632,147	78.7	3.16	69.7
CF11	12/15/14	Capillary tubes	Hyperbaric Normoxic	661,270	54.1	743,008	61.0	4.87	81.2
CF11	12/15/14	Capillary tubes	Hyperoxic	663,319	56.8	694,557	59.6	3.29	82.5
CF11	12/15/14	Capillary tubes	Control	665,956	65.0	701,856	69.0	2.82	81.9
(Ex) CF11	2/9/15	Capillary tubes	Hyperbaric Oxygen	668,628	136.1	247,683	51.7	0.00	49.1
CF11	2/9/15	Capillary tubes	Hyperbaric Normoxic	151,624	42.4	194,459	60.6	3.45	76.5
CF11	2/9/15	Capillary tubes	Hyperoxic	432,325	56.5	183,784	25.0	0.90	25.9
CF11	2/9/15	Capillary tubes	Control	425,882	92.1	355,879	82.3	2.34	56.3
CF11	2/9/15	Capillary tubes	Hyperbaric Nitrogen	577,540	86.2	489,850	75.3	2.83	58.8
(Ex) CF14	11/10/15	Capillary tubes	Hyperbaric Oxygen	933,301	87.2	965,502	91.0	3.41	88.1
CF14	11/10/15	Capillary tubes	Hyperbaric Normoxic	137,380	3.7	285,605	7.7	2.35	56.7
CF14	11/10/15	Capillary tubes	Hyperoxic	870,214	85.7	934,263	92.5	3.77	80.5
CF14	11/10/15	Capillary tubes	Control	431,431	70.6	468,567	77.3	7.16	77.5
(Tr) CF14	1/14/15	Capillary tubes	Hyperbaric Oxygen	561,663	91.4	537,448	87.7	4.96	88.1
CF14	1/14/15	Capillary tubes	Hyperbaric Normoxic	411,792	27.8	456,518	31.0	6.61	75.0
CF14	1/14/15	Capillary tubes	Hyperoxic	156,448	9.5	200,827	12.3	2.89	59.9
CF14	1/14/15	Capillary tubes	Control	361,948	70.9	421,484	83.6	13.61	81.0
CF14	1/14/15	Capillary tubes	Hyperbaric Nitrogen	399,329	19.3	434,245	21.1	3.59	62.5
(Ex) CF14	2/12/15	Capillary tubes	Hyperbaric Oxygen	397,680	57.0	444,933	65.0	5.44	75.5
CF14	2/12/15	Capillary tubes	Hyperbaric Normoxic	543,692	88.4	588,603	96.3	4.43	62.3
CF14	2/12/15	Capillary tubes	Hyperoxic	350,306	38.9	436,206	49.3	4.62	67.0
CF14	2/12/15	Capillary tubes	Control	350,306	71.2	436,206	90.7	14.88	70.6
CF14	2/12/15	Capillary tubes	Hyperbaric Nitrogen	624,202	78.9	720,297	91.8	8.09	68.2
CF9	10/23/14	Sputum	N/A	361,559	19.2	1,495,065	88.8	14.7	-
CF9	11/20/14	Sputum	N/A	200,681	14.4	1,115,970	89.7	19.38	-
CF14	2/12/15	Sputum	N/A	88,682	3.0	233,501	8.2	12.08	-


Sample arrangement for each holder
pH (+)
pH (-)
Redox (+)
Redox (-)
Plug (+)
Plug (-)
O2 (+)
O2 (-)
pH Standards
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(-)
ASM(-)
ASM(-)
(+) Artificial Sputum Media with fresh sputum
(-) Artificial Sputum Media only
pH: Phenol red/Bromocresol purple
Redox: Tetrazolium dye
Plug: Coomassie brilliant blue
O2 : Optodes beads
ASM: Artificial sputum media
Day 1
Day 4

Supplementary Figure 6.1: Sample arrangement in each MH holder and the colorimetric changes observed on Day 4 of incubation at 37°C and 5% CO_2. Tetrazolium dyes were added into tubes 3 and 4 to detect growth. Interestingly, there was no gas production in tube 3 that containing tetrazolium dye and the same mixture of ASM and sputum as the rest of the sample tubes.

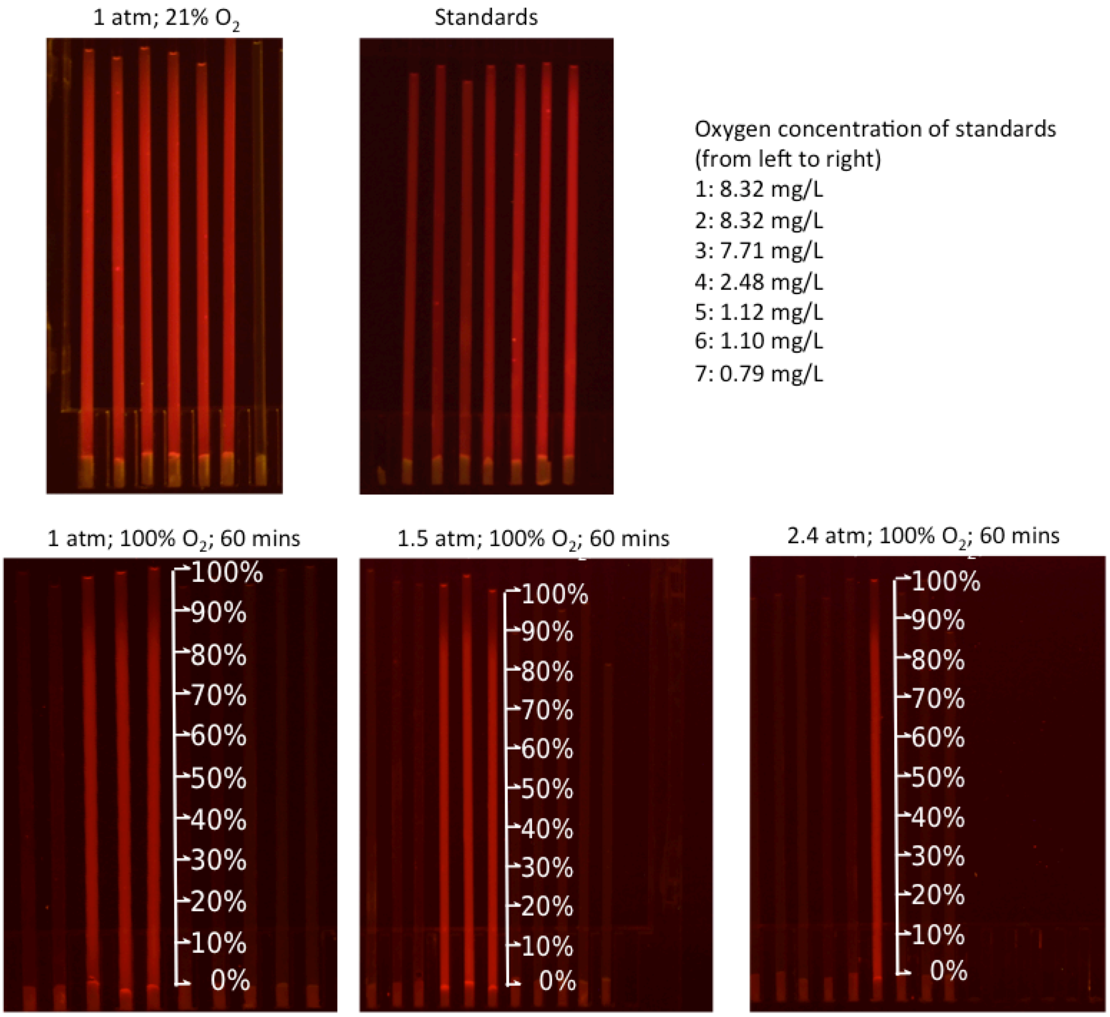

Sample arrangement for each holder
pH (+)
pH (-)
Redox (+)
Redox (-)
Plug (+)
Plug (-)
O2 (+)
O2 (-)
pH Standards
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(+)
ASM(-)
ASM(-)
ASM(-)
(+) Artificial Sputum Media with fresh sputum
(-) Artificial Sputum Media only
Allowing air circulation
Capillary tubes lay horizontally in the MH holder
MH holder
100mm x 100mm Square
Clear Plastic Cell Culture Dish

Supplementary Figure 6.2: Design of an MH holder and sample arrangement in each holder

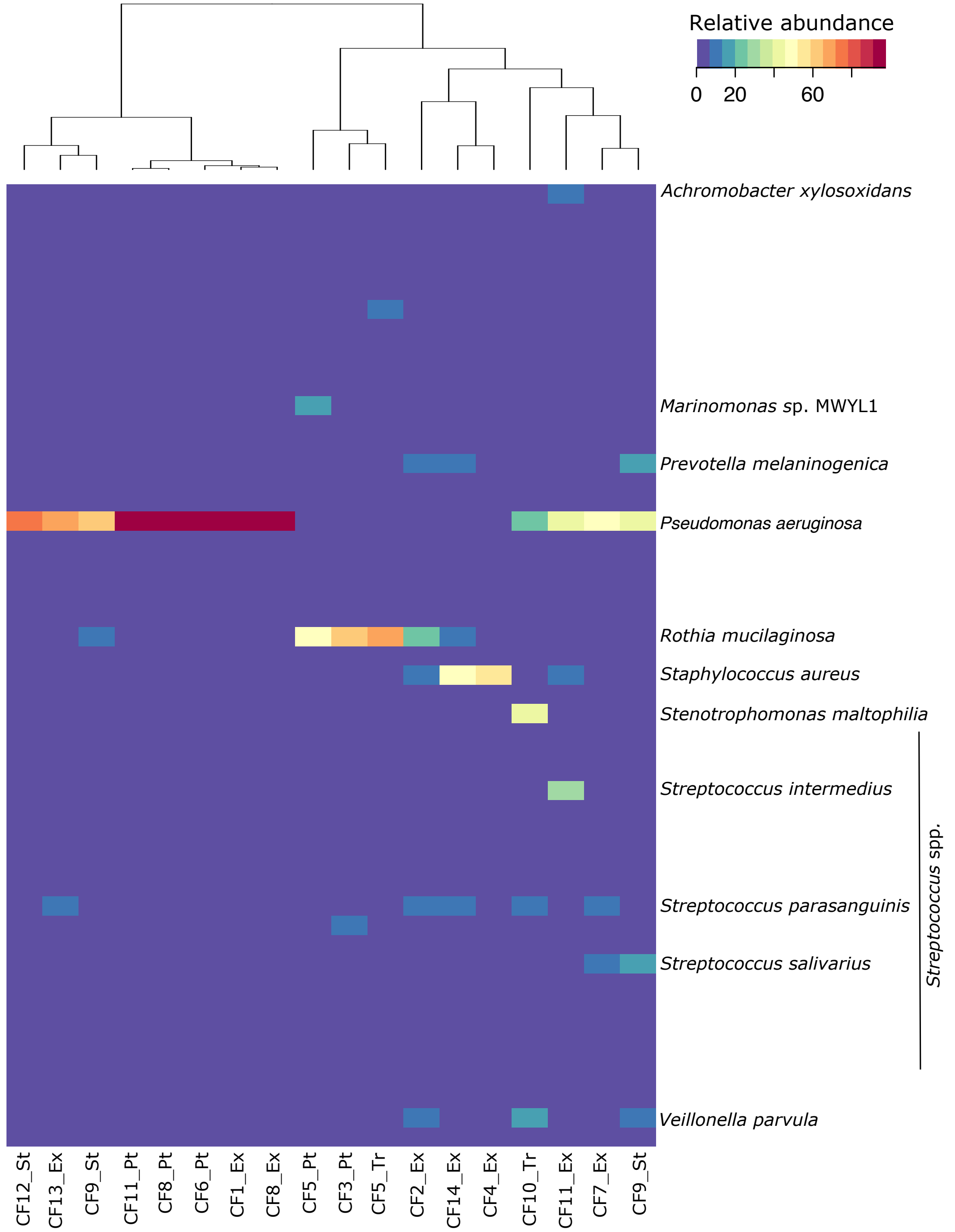


Group	Hyperbaric O_2	Hyperbaric Normoxic	Hyperoxic	Hyperbaric N_2
Pressure	2.4 ATA	2.4 ATA	1 ATA	2.4 ATA
pO_2	100%	21%	100%	< 1%
pCO_2	0%	0.035%	0%	0%

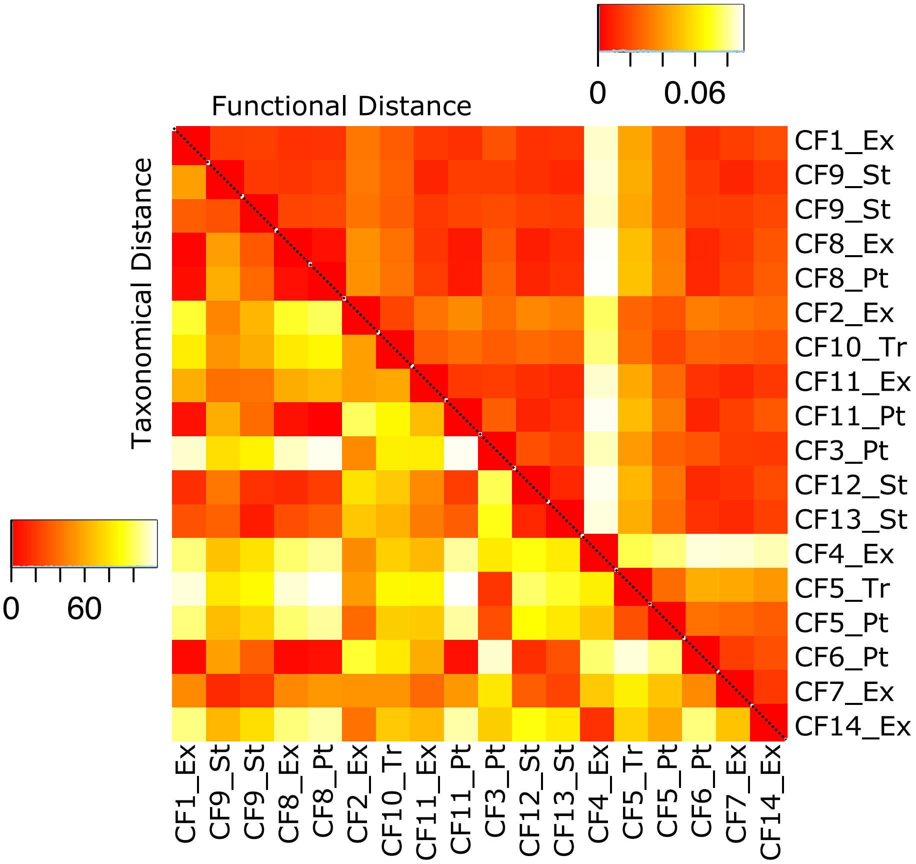
Supplementary Figure 6.3: Perturbations were introduced through increased oxygen and pressure. Experimental treatments were performed prior to the incubation of capillary tubes at 37°C. Antibiotics (Ciprofloxacin or Tobramycin) or phosphate buffered saline (PBS) was added to the top of the mucus within the capillary tubes to mimic addition of compounds onto plugged airways.



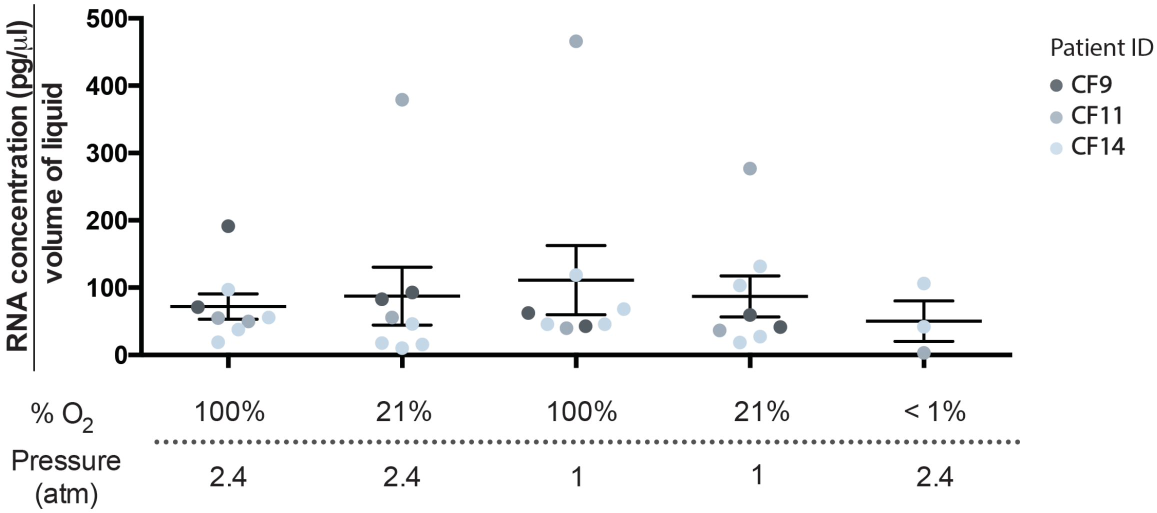
Supplementary Figure 6.4: Oxygen and hyperbaric oxygen treatment increases the amount of dissolved oxygen into mucus solution. Relative O_2 concentration was monitored based on the sensing chemistry platinum (II) 5,10,15,20-tetrakis (2,3,4,5,6-pentafluorophenul)-porphyrin (PtTFPP) entrapped into poly(styrene-block-venulpyrolidone) nanobeads, at which low oxygen is indicated by bright fluorescence and high oxygen is indicated by dark fluorescence (refer to Standards).



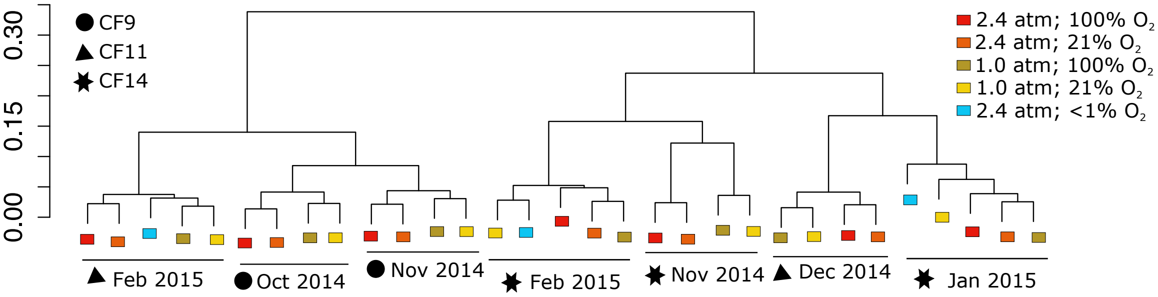
Supplementary Figure 6.5: Microbiomes of all CF patients in this study based on shotgun metagenomic sequencing of microbes-enriched DNA. Sequence data were analyzed based on BLASTn comparison against the NCBI nucleotide database.



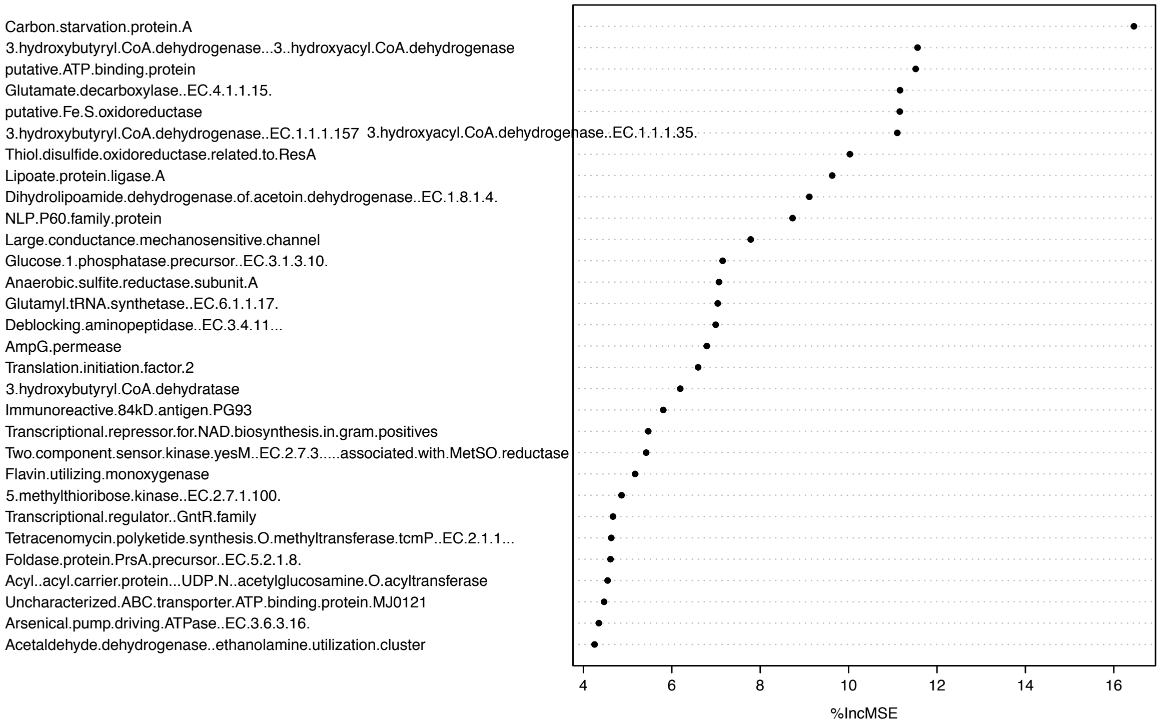
Supplementary Figure 6.6: Similarity of each CF sample based on taxonomical and functional assignment using the metagenomic sequence data.



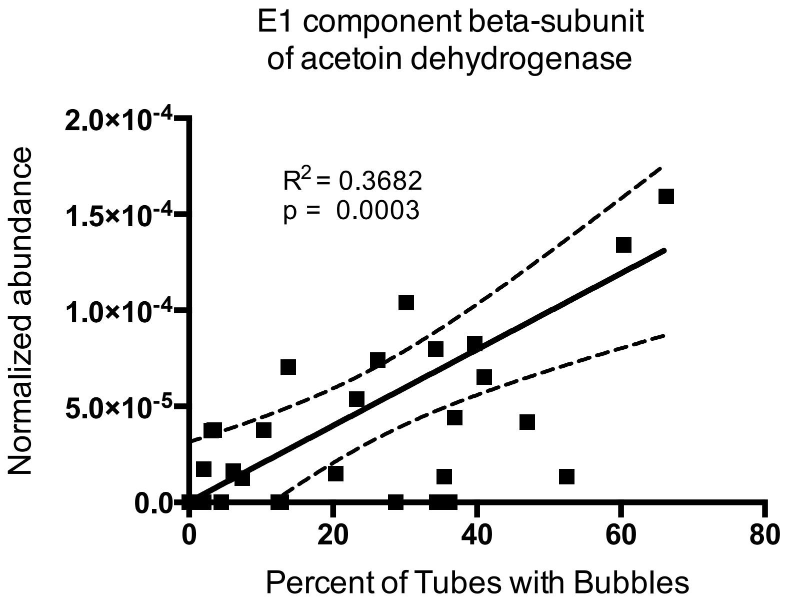
Supplementary Figure 6.7: The amount of RNA per volume of liquid within the capillary tubes. Total RNA was extracted from the contents of 3 tubes and the same tubes were used for the calculation of % of gas within the tubes. RNA concentration was calculated based on the area under-the-curve on Agilent Bioanalyzer 6000 Pico Chip analysis. The total volume of liquid was calculated as (100 - percentage of tubes with gas).



Supplementary Figure 6.8: Hierarchical clustering of samples based on Bray-Curtis distance calculated from the overall normalized gene abundance assigned via CFaDB.



Supplementary Figure 6.9: Variable Importance Plot (VIP) based on the percent mean squared error (MSE) of a supervised random forests analysis. The higher the error, the more important the variables in this model. Therefore, the top 10 genes were chosen for linear regression analysis against the percent gas production.)



Supplementary Figure 6.10: Normalized gene expression level of the E1 components beta-subunit of acetoin dehydrogenase against the `amount of gas production.

CHAPTER 7

The final run

Perspective and Future Directions in CF research

Redefining the Climax and Attack Model published by (Conrad et al. 2013): A model of our current understanding of CF pulmonary disease is presented in Figure 7.1. The lack of functional CFTR proteins on the airway epithelia leads to defective mucus formation and clearance, which provides a platform for the colonization of opportunistic microbes including viruses, bacteria, and fungi (Smith et al. 1996; LiPuma 2010). Many of these organisms come from the environment and the host itself including oral cavity (Fodor et al. 2012; Twomey et al. 2013). Inefficient immune responses result in not only the inability to clear the infection but also tissue damage and permanent scarring. Microbes in the CF lungs are able to adapt to the CF airways through acquiring similar metabolic potentials, virulence factors, and antibiotic resistance genes, or regulating their metabolic activities in response to perturbations (Lim et al. 2012; Lim et al. 2013; Lim et al. 2014; and Chapter 7).

As the disease progresses, mucus continues to build up and microbial activities within the lungs create oxygen, chemical, and nutrient gradients that further select for a specific groups of persistent microbes (Worlitzsch et al. 2002; Quinn et al. 2014; Cowley et al. 2015). In time, the transient initial communities become the persistent *Climax* community that is highly resilient to perturbations by antibiotics, host immune responses, and other microbes (Conrad et al. 2013). The *Climax* community establishes itself within the CF lungs, especially at the area of mucus plugs and damaged tissues, while incoming opportunistic microbes and their metabolic products continue to cause further damage. The disease progression cycle repeats until the patient loses their lung function. This model is in line with the gradual decline in lung function observed in CF patients (VanDevanter 2012). The proposed model supports the observation that in some cases, high concentrations of the classical CF pathogens, such as *P. aeruginosa,* are recovered during exacerbation and that in time, microbial diversity decreases as the patient age (Cox et al. 2010; Zhao et al. 2012). However, to date the clinical significance of community composition fluctuations in the progression of CF pulmonary disease and exacerbations is still unclear.

The role of anaerobes and other upper respiratory tract bacteria such as *Streptococcus* spp. in CF lungs is significant, but yet to be determined. In this dissertation, all facultative anaerobes and obligate anaerobes are collectively defined as anaerobes. The build up of toxic products and the production of secondary metabolites by anaerobes are known to support other microbes within a polymicrobial system (Létoffé et al. 2014; Whiteson et al. 2014), as well as directly and indirectly elicit heighten immune responses (Murray et al. 2014). This dissertation showed that fermentation processes are detected in CF sputum during stable and exacerbation states, but not during treatment, indicating that the drugs administered in response to exacerbation, including antibiotics, inhibit fermentative activity.

Here, I proposed the mechanisms that lead to CF exacerbation:

(i) Microbial activities lower the pH within the mucus plug, generate high level of reactive oxygen species, and create a steep redox potential.

(ii) Anaerobes express stress response genes to alleviate the effect of acidification and reactive oxygen species.

(iii) Anaerobes accomplish redox homeostasis through enzymes (thioredoxins and different oxidoreductases) and metabolites (such as acetaldehyde), as well as induction of pyruvate fermentation.

(iv) Neutral products from fermentation also neutralize acidification.

(v) However, accumulation of toxic products from fermentation and stress-response processes initiate immune responses, which directly or indirectly (through other members of the community) cause exacerbation.

The tipping point of a Climax community: Current clinical therapies for CF lung disease are mainly directed against the symptoms. These treatments often indiscriminately decrease the symptoms but rarely provide long-term resolution in decreasing exacerbation events. In some cases, the steady decline in lung function becomes a downward spiral. Most often lung transplant is required to prolong the patient's life. Therefore, the motivation for future studies is to decrease exacerbation frequency and airway remodeling through the manipulation of microbial metabolisms such as the anaerobes that are associated with exacerbation. In addition, further study is needed to identify the tipping point of the *Climax* community that eventually leads to the death of CF patients.

Applications of multi-omics approaches in other polymicrobial infections. This dissertation presents the first attempt in combining multi-disciplinary approaches in studying complex polymicrobial infections. The combination use of clinical information, metagenomics, metatranscriptomics, and metabolomics has proven to be beneficial. The same approaches presented in this dissertation can also be used in the monitoring of other chronic diseases from obesity, diabetes and chronic heart diseases to chronic wound infections, asthma and chronic obstructive pulmonary diseases (COPD), all of which are thought to be the result of polymicrobial infections. To date, one of the many challenges remains is the process of aggregation and presentation of multi-dimensional data in a concise and meaningful way.

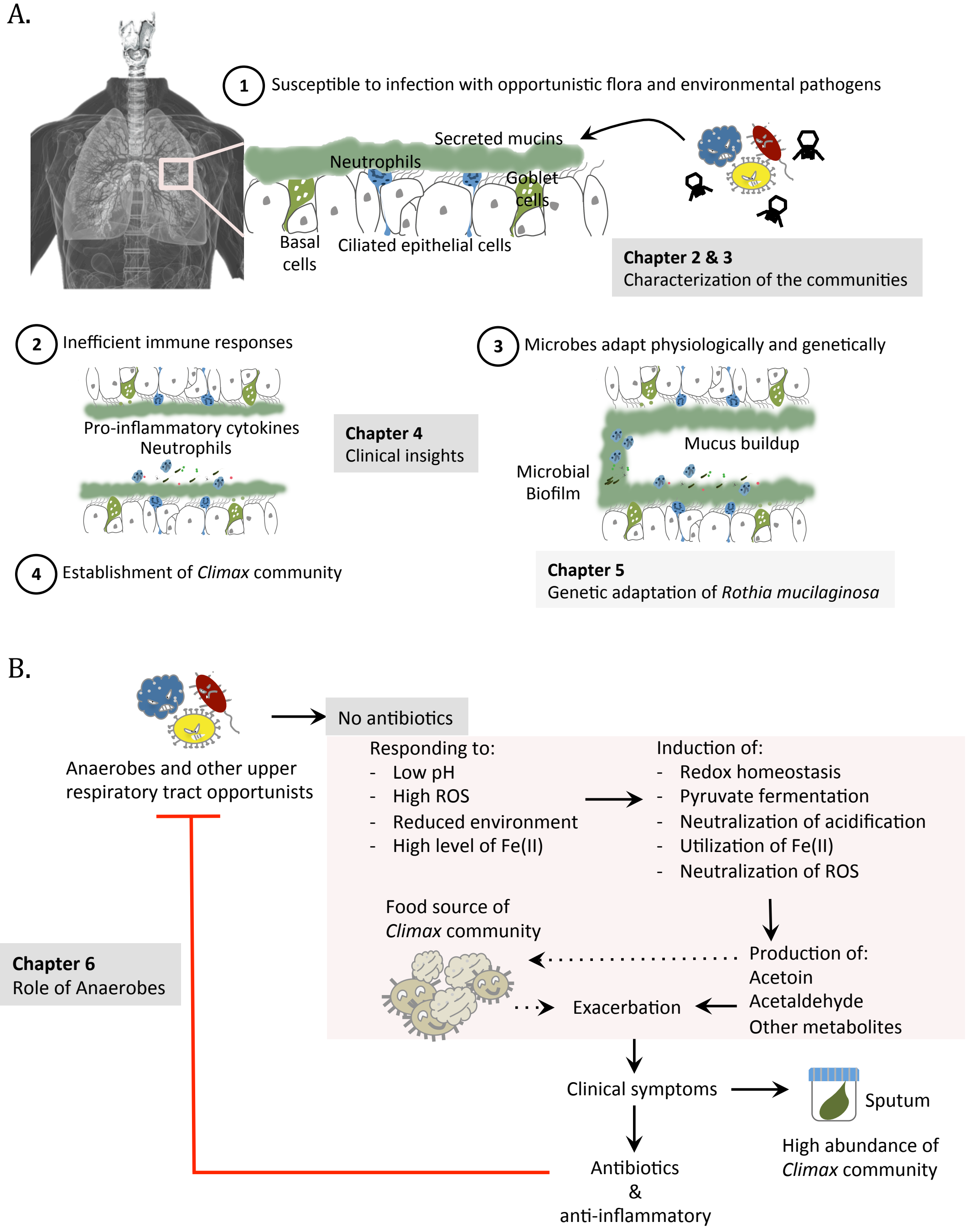

A.
1 Susceptible to infection with opportunistic flora and environmental pathogens
Secreted mucins
Neutrophils
Goblet cells
Basal cells
Ciliated epithelial cells
Chapter 2 & 3
Characterization of the communities
2 Inefficient immune responses
Pro-inflammatory cytokines
Neutrophils
Chapter 4
Clinical insights
3 Microbes adapt physiologically and genetically
Mucus buildup
Microbial Biofilm
4 Establishment of *Climax* community
Chapter 5
Genetic adaptation of *Rothia mucilaginosa*
B.
Anaerobes and other upper respiratory tract opportunists
No antibiotics
Responding to:
- Low pH
- High ROS
- Reduced environment
- High level of Fe(II)
Induction of:
- Redox homeostasis
- Pyruvate fermentation
- Neutralization of acidification
- Utilization of Fe(II)
- Neutralization of ROS
Food source of *Climax* community
Production of:
Acetoin
Acetaldehyde
Other metabolites
Exacerbation
Chapter 6
Role of Anaerobes
Clinical symptoms
Sputum
High abundance of *Climax* community
Antibiotics & anti-inflammatory

Figure 7.1: What's happening in the CF lungs? A. The defective mucociliary clearance in the CF lungs provides a colonization platform to opportunistic flora and environmental pathogens. Inefficient immune responses lead to chronic infection and microbes adapt to the CF lung environment. In time, a resilient *Climax* community establishes and CF patients gradually lose their lung function. B. Current hypothesis in the roles of anaerobes in CF pathogenesis.

Conclusion

Recognition of the CF lung as a polymicrobial ecosystem opens the door to new ways of understanding and treating CF patients. Chapter 2 provides a set of comprehensive molecular solution for studying the viral and microbial communities in complex samples, using CF sputum as an example. Chapter 3 is the first study to simultaneously access microbial metagenomic, viral metagenomic, and community metatranscriptomic data across individuals and temporal scales in any ecosystem. The data also reveals that the dominant player in CF airways, *P. aeruginosa* can be replaced by other opportunistic bacteria such as *Pseudomonas fluorescence* and *Rothia* spp. Chapter 4 describes a new way of coupling metagenomics and clinical information in assessing the patient's health status with the ultimate goal to provide a new and improved way of identifying and characterizing complex disease. Chapter 5 provides an example of how normal flora can evolve into opportunistic pathogens and became an important component of the CF lung microbiome while posing a threat to the CF airways. Metagenomics data were used to reconstruct a near complete CF *Rothia mucilaginosa* genome that was compared to a periodontitis isolate of *R. mucilaginosa*. The CF-derived genome encodes metabolic capabilities and strategies against extracellular stress that allows the organism to survive in the CF lung. It can be concluded that every patient presents a unique microbial community and that these communities adapt to the CF airway environment through the acquisition of similar metabolic potential. Some patients presented a classic CF lung microbiome where *P. aeruginosa*, *S. maltophilia,* or ESBL *E. coli* were one of the main players at the time of sampling. However, other non-typical CF pathogens community members such as *Rothia mucilaginosa*, *Streptococcus* spp., *Prevotella* spp., and *Veillonella* spp. were also common in these patients. An important finding from this dissertation is that the diversity in metabolic potential allows CF microbes to carry out not only aerobic respiration, but also anaerobic respiration and fermentation processes. Chapter 6 provides experimental evidence showing that CF anaerobes are aerotolerant and their metabolism can contribute to CF pathogenesis. These anaerobes are sensitive to perturbations including antibiotic treatments as well as pressure. However, the molecular signatures across the anaerobes in seven samples across three patients showed strong patient- and sample-specific signals. Taken together, this dissertation showed that CF pulmonary disease is complex and highly personalized. The progressive changes in time call for continuous personalized monitoring effort for CF patients.

References

Conrad D, Haynes M, Salamon P, Rainey PB, Youle M, Rohwer F (2013) Cystic fibrosis therapy: a community ecology perspective. Am J Respir Cell Mol Biol 48:150–156. doi: 10.1165/rcmb.2012-0059PS

Cowley ES, Kopf SH, LaRiviere A, Ziebis W, Newman DK (2015) Pediatric Cystic Fibrosis Sputum Can Be Chemically Dynamic, Anoxic, and Extremely Reduced Due to Hydrogen Sulfide Formation. mBio 6:e00767–15. doi: 10.1128/mBio.00767-15

Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified Cystic Fibrosis patients. PLoS ONE 5:e11044. doi: 10.1371/journal.pone.0011044

Fodor AA, Klem ER, Gilpin DF, Elborn JS, Boucher RC, Tunney MM, Wolfgang MC (2012) The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. PLoS ONE 7:e45001. doi: 10.1371/journal.pone.0045001

Létoffé S, Audrain B, Bernier SP, Delepierre M, Ghigo J-M (2014) Aerial exposure to the bacterial volatile compound trimethylamine modifies antibiotic resistance of physically separated bacteria by raising culture medium pH. mBio 5:e00944–00913. doi: 10.1128/mBio.00944-13

Lim YW, Evangelista JS, Schmieder R, Bailey B, Haynes M, Furlan M, Maughan H, Edwards R, Rohwer F, Conrad D (2014) Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. J Clin Microbiol 52:425–437. doi: 10.1128/JCM.02204-13

Lim YW, Schmieder R, Haynes M, Furlan M, Matthews TD, Whiteson K, Poole SJ, Hayes CS, Low DA, Maughan H, Edwards R, Conrad D, Rohwer F (2013) Mechanistic model of Rothia mucilaginosa adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data. PLOS ONE 8:e64285. doi: 10.1371/journal.pone.0064285

Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2012) Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. J Cyst Fibros Off J Eur Cyst Fibros Soc. doi: 10.1016/j.jcf.2012.07.009

LiPuma JJ (2010) The changing microbial epidemiology in Cystic Fibrosis. Clin Microbiol Rev 23:299 –323. doi: 10.1128/CMR.00068-09

Murray JL, Connell JL, Stacy A, Turner KH, Whiteley M (2014) Mechanisms of synergy in polymicrobial infections. J Microbiol Seoul Korea 52:188–199. doi: 10.1007/s12275-014-4067-3

Quinn RA, Lim YW, Maughan H, Conrad D, Rohwer F, Whiteson KL (2014) Biogeochemical Forces Shape the Composition and Physiology of Polymicrobial Communities in the Cystic Fibrosis Lung. mBio 5:e00956–13. doi: 10.1128/mBio.00956-13

Smith, Travis S, E Greenberg, Welsh M (1996) Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. Cell 85:229–236.

Twomey KB, Alston M, An S-Q, O'Connell OJ, McCarthy Y, Swarbreck D, Febrer M, Dow JM, Plant BJ, Ryan RP (2013) Microbiota and Metabolite Profiling Reveal Specific Alterations in Bacterial Community Structure and Environment in the Cystic Fibrosis Airway during Exacerbation. PLoS ONE 8:e82432. doi: 10.1371/journal.pone.0082432

VanDevanter DR (2012) Epidemiology of cystic fibrosis lung disease progression in adolescence.

Whiteson KL, Meinardi S, Lim YW, Schmieder R, Maughan H, Quinn R, Blake DR, Conrad D, Rohwer F (2014) Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. ISME J 8:1247–1258. doi: 10.1038/ismej.2013.229

Worlitzsch D, Tarran R, Ulrich M, Schwab U, Cekici A, Meyer KC, Birrer P, Bellon G, Berger J, Weiss T, Botzenhart K, Yankaskas JR, Randell S, Boucher RC, Döring G (2002) Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. J Clin Invest 109:317–325. doi: 10.1172/JCI13870

Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, Young VB, LiPuma JJ (2012) Decade-Long Bacterial Community Dynamics in Cystic Fibrosis Airways. Proc Natl Acad Sci 109:5809–5814. doi: 10.1073/pnas.1120577109