FROM GENOMES TO METAGENOMES: BIG DATA ANALYSIS OF MICROBES RELATED TO PUBLIC HEALTH

A Dissertation
Presented to
The Academic Faculty

by

Maria Juliana Soto-Girón

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology
December 2018

FROM GENOMES TO METAGENOMES: BIG DATA ANALYSIS OF MICROBES RELATED TO PUBLIC HEALTH

Approved by:

Dr. Konstantinos T. Konstantinidis, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Brian Hammer
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Karen Levy
Rollins School of Public Health
Emory University

Dr. Frank Stewart
School of Biological Sciences
Georgia Institute of Technology

Date Approved: November 6, 2018

To my parents Maria Teresa Girón and Jose Maria Soto.

Thanks for your unconditional support.

I love you

ACKNOWLEDGEMENTS

I would like to thank exceptional people from around the world who made this dissertation possible. Specially, I want to thank my advisor, Dr. Konstantinidis, for his constant support and guidance during this journey as a scientist. I appreciate all his scientific contributions, insightful discussions, and guidance throughout this experience. My sincere thanks also go to my fellow lab mates, it has been an honor to work with them. Thanks for the stimulating discussions, advice, and feedback I have gotten over the years. Special thanks to Coto, Roth, Minjae, Eric, Lizbeth, and Juan.

Special mention goes to my committee members, Dr. Levy, Dr. Stewart, and Dr. Hammer for their constructive feedback during the meetings and for their time and support towards the completion of this dissertation. Profound gratitude goes to King Jordan, who has been a truly mentor since I was an undergraduate student in Colombia. Thanks for his unconditional support and constant encouragement throughout these years.

I will forever be grateful for my second family in Atlanta, my amazing friends, thanks for keeping me sane, for the adventures, parties, birthdays, and for being always there, in the most difficult times. A special acknowledgement goes to Juan Pablo Aragon, Catalina Rivera, Andres Caballero, Victor Rodriguez, Angela Peña, Fabrizio Falasca, Giuliana Salazar, Filippos Tagklis, Abner Ayala, Kizee Etienne, Chris Gaby, Natasha De Leon, Monica Rojas, Giuseppe Trainiti, Melisa Alvarado, Fernando Patiño, Sebastian Ortega, and Carlos Ruiz. Also, to my friends in Colombia, Laura Rodriguez, Leidy Salamanca, Isabel Quiceno, Catalina Gutierrez, Cesar Giraldo, Joel Panay, Mario Ceron, and Oscar Rodriguez.

Words cannot express how lucky I am for having the best support to my side since I decided to study abroad. My parents Maria Teresa and Jose Maria, my brother Sebastian, my grandfather Joaquin Renol, my lovely aunt Lili, uncles Jaime and Robert, cousins, and all my family for their unconditional love, support, and care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Advances in high-throughput sequencing techniques have substantially increase our understanding on how microbes interact among them and with their host. Commensal bacteria are indispensable in host physiology and homeostasis. Conversely, pathogenic bacteria can infect their human host, which leads to several important worldwide diseases. Bacterial infections represent one of the greatest public health concern mainly in young children in the developing world and immunocompromised individuals. Despite advances in clinical microbiology, our understanding of the genomic and ecological mechanisms underlying the pathogen-host-environment interplay especially in clinical settings, remain challenging to elucidate.

In this work, we applied cutting-edge laboratory and bioinformatics techniques to profile microbial communities obtained from diverse sources and that might pose a risk for human health. As biofilms are considered likely reservoirs of pathogens in clinical settings, in **Chapter 2** we characterize the composition of microbial communities growing on hospital shower hoses using shotgun metagenomics. We also evaluate the genetic diversity and resistance profile of the microbial communities associated with their ability to survive under high doses of disinfectants and chlorine residuals. The composition of the human gut microbiota is influenced by multiple extrinsic (lifestyle) and intrinsic (host health) factors. In **Chapter 3**, we explore the role of lifestyle on microbial diversity and functional potential of the gut microbiota in a rural-to-urban gradient in Northern Ecuador and evaluate whether urbanization plays a role in the gut microbial response during diarrheal infection.

Besides diversity and functional potential, quantifying genetic exchange across bacterial species is also important in order to understand how adaptable bacterial species are to environmental fluctuations and for the identification of critical phenotypic properties such as the emergence of antibiotic resistance. To address this question, in **Chapter 4**, we develop a mathematical model to systematically detect and quantify recent gene transfer events between closely related genotypes from large genomic datasets.

INTRODUCTION

Microbes inhabit almost every imaginable environment on earth. Their ability to colonize diverse environments and adapt to changing and harsh conditions is encoded in their genomes. The use of whole-genome sequencing (WGS) technologies, has allowed us to unravel the mechanisms and properties of population genomes related to genetic variability (gene gain and loss), flexibility (horizontal gene transfer), and lifestyle (clonal vs. recombinogenic) (1–4). At the same time, culture-independent techniques have extended the genomic approaches and contributed to the identification of microbes that have not been cultured yet (approx. 99% of the total diversity for Prokaryotes) (5).

Thus, microbial WGS has become an important tool for effectively and rapidly analyzing hundreds of bacterial genomes from different environments and with special relevance for human health (6, 7). The study of bacterial genomes from multiple isolation sources has increased our knowledge of their ecological roles in different ecosystems, led to the identification of novel species, and the (successful) tracking of disease outbreaks (8, 9). For instance, the identification of different genes or even single nucleotide polymorphisms (SNPs) among genotypes from the same species can be used to distinguish pathogens or commensals (10).

On the other hand, the rise of shotgun metagenomics (sequencing of the total DNA from an environmental sample) has allowed the study of microbes beyond traditional lab-based microbiological techniques and transformed our understanding of the physiology and ecology of communities from diverse ecosystems (7, 11). By applying metagenomics, the total microbial populations that co-habit the same environment and their entire gene collection can be characterized and the phylogenetic relationships among community members can be assessed (12–14).

The accelerated increase of genomic information and low sequencing cost have brought new and inherent computational challenges related to processing, storing, and handling large volume of genomic data. Genomics is considered as the “four-headed beast” because of its highly computational demands when compared to other Big Data domains

(i.e., YouTube, Google, Twitter) (15). Therefore, the development and application of bioinformatics/computational methods for the efficient analysis of large genomic data, remain challenging. Moreover, extracting meaningful information and elucidating the underlying biological mechanisms from sequencing data is an essential task in life sciences, especially in public health and clinical microbiology. Thus, decoding genomic information from microbial communities will increase our understanding of how bacteria exchange information (e.g. antibiotic resistant genes), adapt/respond to modern human lifestyles, and impact host health (e.g., infectious diseases).

In this thesis, we applied cutting-edge laboratory and computational algorithms to profile and quantify microbes in diverse ecosystems with a profound impact on public health. Specifically, In **Chapter 2**, we used shotgun metagenomics to catalog biofilm-associated microbial communities growing on showerheads in a hospital in Ohio and evaluate their relevance for nosocomial infections and the emergence of antibiotic-resistant bacteria. In **Chapter 3**, we applied shotgun metagenomics and 16S rRNA gene sequencing to study the relationship between lifestyle and the gut microbial composition and its response during acute diarrheal disease in a rural-to-urban gradient in Northern Ecuador. Finally, in **Chapter 4**, we introduced an alternative mathematical model to quantify recent genetic flow between closely related genotypes from a collection of hundreds of bacterial genomes. We applied this model to estimate the fraction of recent gene exchange of several opportunistic pathogens and identify the functional profile of the recent imports.

1.1 Application of high-throughput sequencing technologies in bacteria growing on hospital surfaces

The identification of bacterial populations that cause infections represents a crucial initial step in clinical microbiology and public health, in order to develop effective control strategies and pathogen surveillance. Bacterial infections are one of the greatest public health concern mainly in low- and middle-income countries causing morbidity and mortality, particularly in children (16, 17). Alarming, the World Health Organization (WHO) has recognized multi-drug resistant bacteria as one of the major and potentially

most dangerous threats worldwide (WHO, 2018). In particular, bacteria present in healthcare centers are a serious health risk of hospital-acquired infections (HAI). These infections occur in 10% in developing and 7% in developed countries (19) and account for 4%–56% of mortality in neonates with a high of 75% in South-East Asia and Sub-Saharan Africa (20). In the US, approximately 75,000 patients die because of these infections according to the Centers for Disease Control and Prevention (CDC) (21).

Despite efforts to control pathogen transmission within medical settings by using different cleaning and disinfecting protocols, the presence of opportunistic pathogens and bacteria remains the biggest problem of HAIs. Medical devices including urinary tract and central venous catheters (22), ventilator-associated pneumonia, and surgical site infections (19) are known to be a source of pathogenic bacteria. Additionally, potential pathogens can be found in beds, floor, windows, soap dispensers, and even in water distribution systems (23, 24).

Given the elevated number of nosocomial infections and spread of antibiotic resistant bacteria within health care settings, there is an urgent need to develop more efficient and rapid detection protocols to characterize and monitor the major pathogen reservoirs (e.g., medical devices, surfaces) and the mechanisms of microbial transfer inside hospitals. Recently, high-throughput culture-independent approaches have been added to screening protocols in health care settings allowing an unbiased detection of the whole microbial community and tracking of pathogens. For instance, the gut microbiome of neonates in intensive care units (ICU) is colonized by microbes residing in ICU, specifically room surfaces (25). Further, Greninger and collaborators (26) applied metagenomics to monitor in real time the progression of parainfluenza 3 virus infections at a children's hospital and identify the common source in a medical unit. Recently, the diversity and dynamics of the microbiome of healthcare settings and its interaction with abiotic factors, building materials, and even medical devices are only beginning to be explored (27–30). New, quantitative insights with respect to these issues will allow us to unravel the reservoirs and new sources of emerging and unrecognized pathogens to ultimately reduce the incidence of nosocomial infections and hospital outbreaks.

In Chapter 2, we characterized the taxonomic composition and functional potential of biofilm microbial communities growing on showerheads at a major U.S. hospital using shotgun metagenomics. We were able to recover the draft genome of a novel *Mycobacterium* species, closely related to opportunistic pathogenic nontuberculous mycobacteria. Additionally, we identified genes related to disinfectant tolerance, virulence determinants involved in colonization and evasion of the host immune system, and genes potentially conferring resistance to several antibiotics.

Collectively, our results highlight the need to understand the microbiome of drinking water biofilms using metagenomic approaches and its potential links to public health. Our data suggested that although water supply systems and surfaces in hospitals are constantly treated with disinfectants, showerhead biofilms represent a potential reservoir for HAIs and antibiotic resistance genes. Therefore, better cleaning practices should be applied in order to significantly minimize the risk of biofilm-associated infections in susceptible populations.

1.2 The gut microbiota from a rural-to-urban gradient: diversity and resilience unraveled by metagenomics

Commensal microbiota co-evolve with their host and is indispensable in multiple metabolic functions, host physiology, and the development of the immune system (31, 32). Microbes associated with the digestive-tract are referred to as the gut microbiota. Given its tight relationship with the host, previous studies have reported high inter-subject variability of the microbiota composition from healthy subjects from different geographical locations (33–35). Even individual-specific patterns have been observed such as differences in gene content of the same species among individuals (36) as well as a stable and unique microbial profile over time (37). Among the factors that modulate the composition of the gut microbiota include host genetics, immune alterations, antimicrobials, and diet (38–40). Moreover, environmental and social factors such as mode of delivery, breastfeeding, pets at home, and tobacco smoke significantly impact gut microbial communities, especially their development during early life (41, 42).

In recent years, major research efforts have focused on studying how lifestyle factors have shaped our gut microbiota during a transition from rural populations located in farms and agricultural settings to westernized populations living in cities. To date, urbanization has globally increased with more than 53% of the total human population living predominantly in cities (43). Massive movement of populations from rural to urban areas is occurring rapidly in developing countries (44). Previous reports have shown that pre-agriculture rural populations (Yanomami, Venezuela (45), Malawian, Amazon (46), Hadza, Tanzaia (47), and Matses, Peru (48)) harbor higher fecal bacterial diversity than urban/industrialized populations (i.e., USA, Europe).

The impact of urbanization (changes in dietary structure and lifestyle) on microbial diversity together with genetic and environmental factors has been associated with the increased incidence of gastrointestinal alterations including metabolic disorders, inflammatory bowel disease, and obesity mainly in western populations (49–51). At the same time, metabolic syndrome shows a close link with the increase of type 2 diabetes and cardiovascular diseases throughout the world, leading to morbidity and mortality (52). Metabolic syndrome encompasses an increased fasting plasma glucose, reduced HDL cholesterol, hyperlipidemia and hypertension, and obesity (53). Nonetheless, how the shifts in gut microbiome diversity in rural settings are exactly related (i.e., what the underlying mechanisms are) to the abovementioned disease or unhealthy states remains essentially unknown and is the subject of intense research currently.

Changes in gut microbial composition have been recognized as a key factor on the development of metabolic diseases and chronic inflammation with abnormal production of multiple inflammatory mediators, impaired fat accumulation, insulin action, and immunity (54). For instance, the ratio *Bacteroidetes*/*Firmicutes* has been linked to obesity (55), albeit the exact underlying mechanism remains unclear. Type 2 diabetes has been associated with a reduction of butyrate-producing and endotoxins-producing Gram-negative bacteria in the gut, and an increase of opportunistic pathogens (56–58).

Characterizing microbial communities from human populations with different lifestyles and traditions is essential for our understanding on how urbanization processes have shaped the gut microbiota and influenced the development of host disorders/alterations and clinical outcomes. In this context, globalization of the Western lifestyle has brought new changes and perturbations in the microbial ecology of the gut including its capacity to tolerate stress or perturbation before changing to a different state (e.g., resilience) (59, 60). Based on this concept, in **Chapter 3** we were interested in evaluating whether urbanization influences the gut microbiota response during infectious diarrhea by comparing fecal samples from a rural-to-urban gradient during disease and health states.

Despite the increasing number of studies cataloguing gut microbial communities in human populations worldwide, most of the phylogenetic diversity analyses are based on 16S rRNA gene, limiting our knowledge to low-level taxonomic resolution and leaving several unanswered questions such as the metabolic resilience and diversity of the community. Moreover, most of these studies aimed to compare populations from distinct geographical regions and cultural backgrounds and few studies have evaluated differences in microbial composition along rural-to-urban gradient in the same geographical area (e.g., same country or region) (61, 62). Focusing on the microbial dynamics in a lifestyle gradient within the same country undergoing urbanization, our study sidestepped confounding factors such different cultural and social preferences and provided new insights in the microbial response.

Specifically, in **Chapter 3** we applied shotgun metagenomics to compare the gut microbiota of subjects living in Quito (Ecuador's capital) and rural populations from villages in Northern Ecuador and profiled the gut microbiota during acute diarrheal disease (ADD). Our data indicated differences in the abundance of community members and metabolic functions between the two populations most likely driven by lifestyle. When healthy microbiomes were compared to those during ADD, urban subjects showed larger shifts in abundance of multiple taxa and metabolic pathways than those from rural populations, indicating a less resilient gut community in the former ones. Our data indicated that local environmental and geographical factors seem to play a role in the gut

ecosystem response during diarrheal infection, which has important implication for treating diarrheal infection and for human wellbeing.

1.3 Bioinformatics algorithms for analysis of large bacterial genome datasets: detecting recently gene transfer events

Next generation sequencing technologies have allowed us to study bacterial populations at the whole-genome scale, increasing our understanding on how bacteria interact with others and the environment as well as their genomic adaptations (i.e., modify, acquire, or loss gene content) during selective pressures (4, 63, 64). The systematic comparison of hundreds of bacterial genomes from different species has become a highly interesting task with applications in epidemiology, biodefense, biotechnology, among others.

Genetic innovations such as mutations and gene transfer events confer advantage to the bacterial populations that acquire new physiological and metabolic capabilities to colonize new ecological niches and co-evolve with their host (65, 66). Horizontal gene transfer (HGT) is frequent among bacterial populations and can alter phenotypic properties substantially. This process has played a fundamental role on bacterial evolution, genomic diversification, and speciation (67, 68). Previous studies based on genomic data have reported that a considerable fraction of genes in prokaryotic genomes have been derived from HGT (69, 70) and through HGT, divergent populations (phylogenetically unrelated bacteria) can share adaptive traits such as antibiotic resistant genes (71).

While several speciation scenarios have been postulated for bacteria (sexual vs. asexual), the bacterial species concept remains a controversial issue and our understanding on how clusters of closely related genotypes emerge and are maintained under high rates of gene transfer, is far from complete (72–74). Sexual populations are considered those in which the recombination rate “r” (polymorphisms shuffle by recombination during the same time interval) is greater than the mutation rate “m” (i.e., new polymorphisms

introduced by mutation). Under this scenario, recombination can act as a cohesive force that counteracts genetic divergence on the populations (74, 75).

Despite the important effects of HGT on bacterial species, its effect on population structure (e.g., sexual vs. asexual speciation) and diversity remains to be fully understood. This is in part because detecting and quantifying HGT is still limited to small datasets, mainly due to the computationally expensive phylogeny-based approaches available. Thus, the quantification of HGT rates in a large collection of genomes is a noteworthy computational problem to address. On the other hand, quantify HGT between genomes of the same species has been challenging especially due to the high sequence identity of core genes at this level (e.g., low signal-to-noise ratio). Moreover, most of the methods employed to date for this purpose are based on assumptions that are frequently violated by the data analyzed, limiting the broad applicability of the derived conclusions.

To help meet these challenges, in **Chapter 4** we introduced an alternative mathematical model to estimate recent genetic events based on the genome-aggregate Average Nucleotide Identity (ANI) concept (76). Our model quantifies recent gene exchanges in a genome pair by comparing the fraction of shared genes at the 100% nucleotide identity to the average of hundreds of genomes from different bacterial species with similar ANI values. The fact that our approach is not based on a specific method, and its assumptions, represents a distinguishing strength compared to previous approaches and can be computationally scalable to thousands of bacterial genomes. We applied this model to compare the fraction of recent imports within and across bacterial species with distinct lifestyles (e.g., symbiotic vs. free-living) and ecological niches (e.g., fluctuating or more stable environments) and provided insights into the role of HGT on bacterial speciation.

Comparison of the rates among commensal and free-living bacteria revealed that opportunistic pathogens, including *Campylobacter jejuni*, *Campylobacter coli*, and *Neisseria meningitids,* showed the highest fraction of genetic exchange. Annotations of the recent gene imports from these opportunistic pathogens indicated an enrichment of functions that allow efficient interactions with host cells as well as antibiotic resistant factors. A second set of genes was comprised by sequences related to the activation or

inactivation of virulence genes and also to the variation in the envelope structure. Collectively, our methodology and associated model offer an important addition to the toolbox for studying recent gene transfer and gene content adaptation on bacterial genomes.

1.4 REFERENCES

1. Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610.

2. Field D, Wilson G, van der Gast C. 2006. How do we compare hundreds of bacterial genomes? Curr Opin Microbiol 9:499–504.

3. Segerman B. 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol 2.

4. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.

5. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82:6955–6959.

6. Wilson DJ. 2012. Insights from Genomics into Bacterial Pathogen Populations. PLOS Pathog 8:e1002874.

7. Gasc C, Ribière C, Parisot N, Beugnot R, Defois C, Petit-Biderre C, Boucher D, Peyretaillade E, Peyret P. 2015. Capturing prokaryotic dark matter genomes. Res Microbiol 166:814–830.

8. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. 2008. Microbiology in the post-genomic era. Nat Rev Microbiol 6:419–430.

9. Castelle CJ, Banfield JF. 2018. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell 172:1181–1197.

10. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin Microbiol Infect 24:350–354.

11. Streit WR, Schmitz RA. 2004. Metagenomics – the key to the uncultured microbes. Curr Opin Microbiol 7:492–498.

12. Kembel SW, Eisen JA, Pollard KS, Green JL. 2011. The Phylogenetic Diversity of Metagenomes. PLOS ONE 6:e23214.

13. Matsen FA. 2015. Phylogenetics and the Human Microbiome. Syst Biol 64:e26–e41.

14. Turaev D, Rattei T. 2016. High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. Curr Opin Biotechnol 39:174–181.

15. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big Data: Astronomical or Genomical? PLoS Biol 13.

16. Deen J, Seidlein L von, Andersen F, Elle N, White NJ, Lubell Y. 2012. Community-acquired bacterial bloodstream infections in developing countries in south and southeast Asia: a systematic review. Lancet Infect Dis 12:480–487.

17. Seale AC, Blencowe H, Zaidi A, Ganatra H, Syed S, Engmann C, Newton CR, Vergnano S, Stoll BJ, Cousens SN, Lawn JE, Neonatal Infections Estimation Team. 2013. Neonatal severe bacterial infection impairment estimates in South Asia, sub-Saharan Africa, and Latin America for 2010. Pediatr Res 74 Suppl 1:73–85.

18. WHO. 2018. WHO releases its first report on global antibiotic resistance. CIDRAP.

19. Khan HA, Baig FK, Mehboob R. 2017. Nosocomial infections: Epidemiology, prevention, control and surveillance. Asian Pac J Trop Biomed 7:478–482.

20. WHO. 2016. The burden of health care-associated infection worldwide. WHO.

21. CDC. 2018. Healthcare-associated infections (HAIs). HAI Data and Statistics.

22. Aly NYA, Al-Mousa HH, Asar ESMA. 2008. Nosocomial Infections in a Medical-Surgical Intensive Care Unit. Med Princ Pract 17:373–377.

23. Buffet-Bataillon S, Rabier V, Bétrémieux P, Beuchée A, Bauer M, Pladys P, Le Gall E, Cormier M, Jolivet-Gougeon A. 2009. Outbreak of Serratia marcescens in a neonatal intensive care unit: contaminated unmedicated liquid soap and risk factors. J Hosp Infect 72:17–22.

24. Decker BK, Palmore TN. 2014. Hospital water and opportunities for infection prevention. Curr Infect Dis Rep 16:432.

25. Brooks B, Firek BA, Miller CS, Sharon I, Thomas BC, Baker R, Morowitz MJ, Banfield JF. 2014. Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. Microbiome 2:1.

26. Greninger AL, Zerr DM, Qin X, Adler AL, Sampoleo R, Kuypers JM, Englund JA, Jerome KR. 2017. Rapid Metagenomic Next-Generation Sequencing during an

Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. J Clin Microbiol 55:177–182.

27. Arnold C. 2014. Rethinking Sterile: The Hospital Microbiome. Environ Health Perspect 122:A182–A187.

28. Chen C-H, Lin Y-L, Chen K-H, Chen W-P, Chen Z-F, Kuo H-Y, Hung H-F, Tang CY, Liou M-L. 2017. Bacterial diversity among four healthcare-associated institutes in Taiwan. Sci Rep 7.

29. Lax S, Sangwan N, Smith D, Larsen P, Handley KM, Richardson M, Guyton K, Krezalek M, Shogan BD, Defazio J, Flemming I, Shakhsheer B, Weber S, Landon E, Garcia-Houchins S, Siegel J, Alverdy J, Knight R, Stephens B, Gilbert JA. 2017. Bacterial colonization and succession in a newly opened hospital. Sci Transl Med 9:eaah6500.

30. Lax S, Smith D, Sangwan N, Handley K, Larsen P, Richardson M, Taylor S, Landon E, Alverdy J, Siegel J, Stephens B, Knight R, Gilbert JA. 2017. Colonization and Succession of Hospital-Associated Microbiota. Sci Transl Med 9.

31. Ivanov II, Littman DR. 2011. Modulation of immune homeostasis by commensal bacteria. Curr Opin Microbiol 14:106–114.

32. Reid G, Younes JA, Van der Mei HC, Gloor GB, Knight R, Busscher HJ. 2011. Microbiota restoration: natural and supplemented recovery of human microbial communities. Nat Rev Microbiol 9:27–38.

33. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the Human Intestinal Microbial Flora. Science 308:1635–1638.

34. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694–1697.

35. Jones RB, Zhu X, Moan E, Murff HJ, Ness RM, Seidner DL, Sun S, Yu C, Dai Q, Fodor AA, Azcarate-Peril MA, Shrubsole MJ. 2018. Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab, and mucosal samples. Sci Rep 8:4139.

36. Zhu A, Sunagawa S, Mende DR, Bork P. 2015. Inter-individual differences in the gene content of human gut bacterial species. Genome Biol 16:82.

37. Rajilić-Stojanović M, Heilig Hans G. H. J., Tims Sebastian, Zoetendal Erwin G., Vos Willem M. 2012. Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 15:1146–1159.

38. Conlon MA, Bird AR. 2014. The impact of diet and lifestyle on gut microbiota and human health. Nutrients 7:17–44.

39. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

40. Munyaka PM, Khafipour E, Ghia J-E. 2014. External Influence of Early Childhood Establishment of Gut Microbiota and Subsequent Health Implications. Front Pediatr 2.

41. Levin AM, Sitarik AR, Havstad SL, Fujimura KE, Wegienka G, Cassidy-Bushrow AE, Kim H, Zoratti EM, Lukacs NW, Boushey HA, Ownby DR, Lynch SV, Johnson CC. 2016. Joint effects of pregnancy, sociocultural, and environmental factors on early life gut microbiome structure and diversity. Sci Rep 6:31775.

42. Martin R, Makino H, Cetinyurek Yavuz A, Ben-Amor K, Roelofs M, Ishikawa E, Kubota H, Swinkels S, Sakai T, Oishi K, Kushiro A, Knol J. 2016. Early-Life Events, Including Mode of Delivery and Type of Feeding, Siblings and Gender, Shape the Developing Gut Microbiota. PloS One 11:e0158498.

43. King GM. 2014. Urban microbiomes and urban ecology: how do microbes in the built environment affect human sustainability in cities? J Microbiol Seoul Korea 52:721–728.

44. Eckert S, Kohler S. 2014. Urbanization and health in developing countries: a systematic review. World Health Popul 15:7–20.

45. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, Magris M, Hidalgo G, Contreras M, Noya-Alarcón Ó, Lander O, McDonald J, Cox M, Walter J, Oh PL, Ruiz JF, Rodriguez S, Shen N, Song SJ, Metcalf J, Knight R, Dantas G, Dominguez-Bello MG. 2015. The microbiome of uncontacted Amerindians. Sci Adv 1:1–12.

46. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

47. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, Fiori J, Gotti R, De Bellis G, Luiselli D, Brigidi P, Mabulla A, Marlowe F, Henry AG, Crittenden AN. 2014. Gut microbiome of the Hadza hunter-gatherers. Nat Commun 5:3654.

48. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villarroel O, Foster M, Guija-Poma E, Troncoso-Corzo L, Warinner C,

Ozga AT, Lewis CM. 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun 6.

49. Peterson DA, Frank DN, Pace NR, Gordon JI. 2008. Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. Cell Host Microbe 3:417–427.

50. Ley RE. 2010. Obesity and the human microbiome. Curr Opin Gastroenterol 26:5–11.

51. Dugas LR, Fuller M, Gilbert J, Layden BT. 2016. The obese gut microbiome across the epidemiologic transition. Emerg Themes Epidemiol 13:2.

52. Wellen KE, Hotamisligil GS. 2005. Inflammation, stress, and diabetes. J Clin Invest 115:1111–1119.

53. He M, Shi B. 2017. Gut microbiota as a potential target of metabolic syndrome: the role of probiotics and prebiotics. Cell Biosci 7:54.

54. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, Thomas LV, Zoetendal EG, Hart A. 2016. The gut microbiota and host health: a new clinical frontier. Gut 65:330–339.

55. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006. Microbial ecology: Human gut microbes associated with obesity. Nature 444:1022–1023.

56. Larsen N, Vogensen FK, Berg FWJ van den, Nielsen DS, Andreasen AS, Pedersen BK, Al-Soud WA, Sørensen SJ, Hansen LH, Jakobsen M. 2010. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. PLOS ONE 5:e9085.

57. Wu X, Ma C, Han L, Nawaz M, Gao F, Zhang X, Yu P, Zhao C, Li L, Zhou A, Wang J, Moore JE, Millar BC, Xu J. 2010. Molecular characterisation of the faecal microbiota in patients with type II diabetes. Curr Microbiol 61:69–78.

58. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60.

59. Gillings MR, Paulsen IT, Tetu SG. 2015. Ecology and Evolution of the Human Microbiota: Fire, Farming and Antibiotics. Genes 6:841–857.

60. Sommer F, Anderson JM, Bharti R, Raes J, Rosenstiel P. 2017. The resilience of the intestinal microbiota influences health and disease. Nat Rev Microbiol 15:630–638.

61. Winglee K, Howard AG, Sha W, Gharaibeh RZ, Liu J, Jin D, Fodor AA, Gordon-Larsen P. 2017. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. Microbiome 5.

62. Stagaman K, Cepon-Robins TJ, Liebert MA, Gildner TE, Urlacher SS, Madimenos FC, Guillemin K, Snodgrass JJ, Sugiyama LS, Bohannan BJM. 2018. Market Integration Predicts Human Gut Microbiome Attributes across a Gradient of Economic Development. mSystems 3:e00122-17.

63. Bobay L-M, Ochman H. 2017. Biological Species Are Universal across Life's Domains. Genome Biol Evol 9:491–501.

64. Chaguza C, Bentley SD. 2017. Adaptation... that's what you need? Nat Rev Microbiol. News.

65. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976.

66. Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 6:8924.

67. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

68. Cohan FM, Koeppel AF. 2008. The Origins of Ecological Diversity in Prokaryotes. Curr Biol 18:R1024–R1034.

69. Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–766.

70. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. Science 330:50.

71. Fondi M, Fani R. 2010. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. Environ Microbiol 12:3228–3242.

72. Hanage WP, Fraser C, Spratt BG. 2006. The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol 239:210–219.

73. Narra HP, Ochman H. 2006. Of What Use Is Sex to Bacteria? Curr Biol 16:R705–R710.

74. Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the Nature of Bacterial Speciation. Science 315:476–480.

75. Fraser C, Hanage WP, Spratt BG. 2005. Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci 102:1968–1973.

76. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci 102:2567–2572.

CHAPTER 2. CHARACTERIZATION OF BIOFILMS DEVELOPING ON HOSPITAL SHOWER HOSES AND IMPLICATIONS FOR NOSOCOMIAL INFECTIONS

2.1 ABSTRACT

Although the source of drinking water (DW) used in hospitals is commonly disinfected, biofilms forming on water pipelines are a refuge to bacteria, including possible pathogens, which survive different disinfection strategies. These biofilm communities are only beginning to be explored by culture-independent techniques that circumvent the limitations of conventional monitoring efforts. Hence, the frequency of opportunistic pathogens in DW biofilms and how biofilm members withstand high doses of disinfectants and/or chlorine residuals in the water supply remain speculative. The aim of this study was to characterize the composition of microbial communities growing on five hospital shower hoses using both 16S rRNA gene sequencing of bacterial isolates and whole-genome shotgun metagenome sequencing. The resulting data revealed a *Mycobacterium*-like population, closely related to *M. rhodesiae* and *M. tusciae*, to be the predominant taxon in all five samples, and its nearly complete draft genome was recovered. In contrast, the fraction recovered by culture was mostly affiliated to *Proteobacteria*, such as members of the genera *Sphingomonas*, *Blastomonas,* and *Porphyrobacter.* The biofilm community harbored genes related to disinfectant tolerance (2.34% of the total annotated proteins), and a lower abundance of virulence determinants related to colonization and evasion of the host immune system. Additionally, genes potentially conferring resistance to beta-lactam, aminoglycoside, amphenicol, and quinolone antibiotics were detected. Collectively, our results underscore the need to understand the microbiome of DW biofilms using

metagenomic approaches. This information could lead to more robust management practices that minimize risks associated with exposure to opportunistic pathogens in hospitals.

2.2 INTRODUCTION

Despite the use of disinfectants in drinking water distribution systems (DWDS), bacteria are able to colonize different parts of DWDS such as building plumbing systems and fixtures (e.g., sinks, showerheads, and faucets) (1–3). Previous studies have shown that several organisms associated with DWDS can tolerate the effects of disinfectant compounds because of their ability to form biofilm (4, 5). Unlike planktonic forms, bacteria in biofilms are more resistant to sterilization procedures and antimicrobial exposure, showing in some cases a minimal inhibitory concentration (MIC) up to 1000-fold higher than their planktonic counterpart (6). Hence, biofilm formation in response to disinfectant treatment can increase the resistance to common cleaning protocols and promote the transfer of antibiotic-resistance genes among the biofilm members, producing multidrug resistance bacteria (7, 8).

Although the frequency of nosocomial infections caused by bacteria located in hospital water supplies is traditionally thought to be low (9), this infection route has regained attention due to the increase of hospital-acquired infections in recent years and the presence of opportunistic pathogens in biofilms located in hospital premise plumbing and medical device (10–12). Microorganisms forming a biofilm can detach and be transferred to surfaces, medical equipment, and human individuals (13). Biofilms located on hospital showerheads can as such be an important reservoir for nosocomial infections (1, 8, 14). Previous studies of the microbial community composition of showerhead biofilms have identified nontuberculous mycobacteria (NTM), some of which are considered opportunistic pathogens that are commonly found in natural environments (i.e., soil and water) as well as in the built environment including hospitals (10, 15). Some NTM species have been linked to hypersensitivity pneumonitis, cervical lymphadenitis, allergies, and respiratory problems mainly in immuno-compromised individuals (16, 17). NTM growing

in biofilms have been identified in drinking water systems, on PVC surfaces, and on showerheads from hospitals, houses, and workplaces (18–20). Their frequent occurrence in such habitats may be explained because of their ability to survive stressors commonly found in distribution systems such as oligotrophic conditions, chlorination, and hot temperatures (21, 22). However, most previous surveys reporting the occurrence and prevalence of mycobacteria in DW have been restricted to 16S ribosomal RNA (rRNA) gene fragment analysis, lacking resolution at the species level (23), and to culture-based techniques (23, 24), which often provide a biased representation of the sample due to the selective lab media, culture conditions, and volume of the sample processed.

The gene functions that underlie the ecological success of most DW bacteria in DWDS remain poorly described, in part due to the lack of genetic information on microbial groups commonly inhabiting DWDS. Hence, metagenome sequencing (i.e., random sequencing of total community DNA extracts) has been used recently to examine the functional network of complex microbial communities (25, 26). In spite of the rise of infections by opportunistic premise plumbing pathogens, relatively few studies have assessed the diversity of biofilms growing in DWDS at the metagenome level, especially in health care units (27–30). Most previous reports are based on 16S rRNA gene amplicon surveys that are limited in scope as far as accurately predicting exposure risks. Therefore, in this study, we characterized the biofilm microbial communities of shower hoses in a hospital using shotgun metagenome sequencing and evaluated the genetic diversity and relative abundance of antibiotic and disinfectant resistance present. We also compared the metagenomics findings to those obtained by a substantial collection of genome sequences of isolates (n=94) recovered from the same samples and those of previous studies from other hospitals and the built environment.

2.3 MATERIALS AND METHODS

2.3.1 Sample collection

The samples used in this study were collected during four consecutive days in 2012 from 40 showerheads located in different rooms within an Ohio hospital. Drinking water in this building normally contains a free chlorine residual of 0.8 mg/L, and the average water temperature and pH are 20°C and 8.4, respectively. In addition, the concentrations of several metals (Cr, Cu, Fe, Ni, Sr, Sn, Pb) were measured using a Inductively Coupled Plasma-Mass Spectrometry (ICP-MS) according to U.S EPA method 200.8 (31) and were found to be below regulatory thresholds (e.g., Al: 52 to 65 ug/L, P: 155 to 170 ug/L, S: 20 to 22 mg/L, K: 0.5 to 2.4 mg/L), with limited variation from room to room.

To minimize collection time, the entire showerheads were removed with the shower hoses, water was discarded and the showerheads transferred to sterile plastic bags, which were then placed in coolers containing ice packs. Hoses were removed and split open with a sterile knife to expose the inner luminal surfaces. Biofilms from the shower hoses were collected by scrapping the inner surfaces with sterile spatulas. The biomass was then transferred to sterile conical tubes and re-suspended in phosphate buffer. Five of the samples were randomly selected for metagenomic studies while all samples were used for conventional microbiological culture. Samples were processed within four hours of collection time.

2.3.2 Culturing and Identification of Isolates

For isolation, an aliquot (1 ml) of the re-suspended biomass was used to grow heterotrophic bacteria. Biofilm samples were diluted and processed in duplicate, spotted onto R2A agar plates (32), which were then incubated at 25°C for 5-7 days. Colonies were re-streaked onto R2A agar plates to obtain single isolated colonies. Using sterilized toothpicks, more than 2000 pure colonies were carefully scrapped from the R2A agar plates and re-suspended in 30 µl of sterile molecular grade water. Re-suspended cells (2 µl) were used to partially amplify the 16S rRNA gene using universal primers 8F and 787F. Amplification conditions and sequencing analysis conducted were the same as described elsewhere (33).

2.3.3 High throughput sequencing

The five samples used for the metagenomic analyses were filtered onto polycarbonate membranes and stored at -20°C until further processed. Total DNA was extracted from these filters using a Ultra Clean Soil DNA kit (MoBio Laboratories) as previously described (34). A subset of all strains isolated in this study were subjected to whole-genome sequencing. This subset represented strains with the most common colony morphotypes and included strains from the samples used in metagenome sequencing.

Total DNA extracted from polycarbonate filters and from selected isolates was normalized to 5ng/µl and libraries were constructed using Illumina TruSeq preparation protocol and sequenced on an Illumina HiSeq 2000 using a 100 bp paired-end read approach, following the instructions of the manufacturer (Illumina, San Diego, CA).

2.3.4 Read trimming and de novo assembly

Raw reads from the metagenomes and isolate genomes were trimmed using SolexaQA with a Q = 20 Phred score cut-off (35); sequences shorter than 50 bp after trimming and/or with Illumina adaptors at the 3' end were discarded. The assembly of the metagenomes was performed using the hybrid protocol previously described (36), which combines Velvet (37), SOAP *de novo* (38) , and Newbler 2.0 (39) assemblers using k-mer values from 31 to 63. Table S1 shows the statistics of the shower hose metagenomes. For the isolate genomes, trimmed reads were assembled using SPAdes assembler with "--sc --careful" and error correction options (40).

2.3.5 Taxonomic classification of the biofilm microbial communities

Taxonomic classification of assembled metagenomic contigs was carried out using MyTaxa with default parameters (likelihood score ≥ 0.5) (41). In addition, the taxonomic affiliation of 16S rRNA gene fragments recovered from metagenomes and isolate genome reads was determined using the Ribosomal Database Project (RDP) classifier (42) with the RDP 16S rRNA gene database release 11.3 (43) at 97% nucleotide sequence identity level.

2.3.6 Metagenomic functional gene assignment and abundance analysis

Protein-coding genes in assembled contigs longer than 5 kbp were identified by MetaGeneMark using default parameters (44). Functional annotation was based on BLASTp (45) searches of the predicted amino acid sequences against the UniProt/SwissProt database (46) using a cut-off for a match of at least 30% identity and 50% of the length of the query protein sequence covered in the alignment. The abundance of protein functions in each dataset was calculated as the number of (assembled) protein sequences assigned to the function above the cut-off divided by the total number of annotated proteins predicted in the respective sample.

Predicted proteins associated with antibiotic resistance mechanisms were identified by BLASTp searches against the antibiotic resistance database (ARDB) (47) composed of 23,137 antibiotic resistance genes (ARG) with a threshold e-value of 1e-10 and at least 70% of the query sequence covered by the BLAST alignment [higher stringency compared to above in order to reduce the frequency of false positive matches, as previously suggested (48)].

Genome equivalents in the metagenomic datasets were calculated as follows: HMM (Hidden Markov Model) searches of 101 universally conserved single-copy genes (49) against the individual, unassembled metagenomic reads were performed using HMMER3 version 3.1 (http://hmmer.janelia.org/) (50) with default settings. Ten models, which represented more than one family or extremely conserved families at the sequence level (*rpoC*, *rpoC1, pheT*-bacteria*, pheT*-archaea, *proS*-bacteria, *proS*-archaea, *glyS*, alpha-*glyS, era* and, tRNA synthase class I), were excluded from further analysis. The median sequencing depth (number of reads/bp) of the remaining 91 HMM models was determined and was taken as a proxy of 1 genome equivalent (i.e., the corresponding proteins should be encoded by every genome in the sample). The number of copies per cell of a target gene was estimated as the sequencing depth of that gene (number of reads/bp) divided by the normalizing factor, *i.e.*, the median number of reads/bp of the 91 universal genes.

ORF prediction and functional annotation of protein-coding genes in the isolate or population (bin) genomes (see also below) were performed as described above for metagenomes. Proteins were assigned to the functional categories using Gene Ontology

terms (51). In addition, genome completeness was estimated by the recovery of the 91 universal single copy genes based on HMM searches. Contamination rate was defined as the percentage of the universal genes found in multiple copies in an isolate or population genome.

2.3.7 Recovery of genomes from metagenomes (Binning)

Assembled contigs for each dataset were clustered using MaxBin (52), an expectation-maximization based-algorithm that combines differential coverage and tetranucleotide compositional information to bin contigs into population genomes. Additionally, population genomes (bins) were visually inspected for uniform coverage across the genome sequence and consistent phylogenetic signal of universal genes, and confirmed using CONCOCT (53). Taxonomic affiliation of bins was based on MyTaxa analysis, and the results were further validated by inspecting the results of BLASTp searches of universal genes predicted in the bins against the NCBI refseq database using the LCA algorithm of MEGAN (54), essentially as previously performed (55).

Potential virulence factors in the *Mycobacterium* bin were identified by BLASTp searches of its predicted proteins against the Virulence Factors of Pathogenic Bacteria (25) and PATRIC databases (56) using a cutoff e-value of 1e-10 and at least 70% of the query aligned sequence. All raw sequence datasets were deposited in the Sequence Read Archive database at NCBI under the number SRP065069 and binned genome sequences are available at http://enve-omics.ce.gatech.edu/data/showerheads.

2.4 **RESULTS**

2.4.1 Composition of the microbial community of shower hose biofilms

The taxonomic assignment based on 16S rRNA gene-encoding metagenomic reads showed that shower hose biofilms contained *Actinobacteria* closely related to the genus *Mycobacterium* (average relative abundance 42.2 ± 13% of total; 13% represents the standard deviation observed among the five samples), *Proteobacteria* closely related to the

genera *Erythrobacter* (average 9.4 ± 3%), *Sphingomonas* (average 6.6 ± 2.6%), *Novosphingobium* (average 4.2 ± 1.4%), and *Bradyrhizobium* (average 5.2 ± 3.2%), and other, less abundant bacterial genera affiliated with the phyla *Bacteriodetes* (4.1 ± 3) and *Firmicutes* (1.2 ± 1) (Fig. 2.1, Table A.1). Similar results were obtained based on best match analysis of predicted protein sequences recovered in the assembled metagenomic contigs against complete available genome sequences (Table A.1).

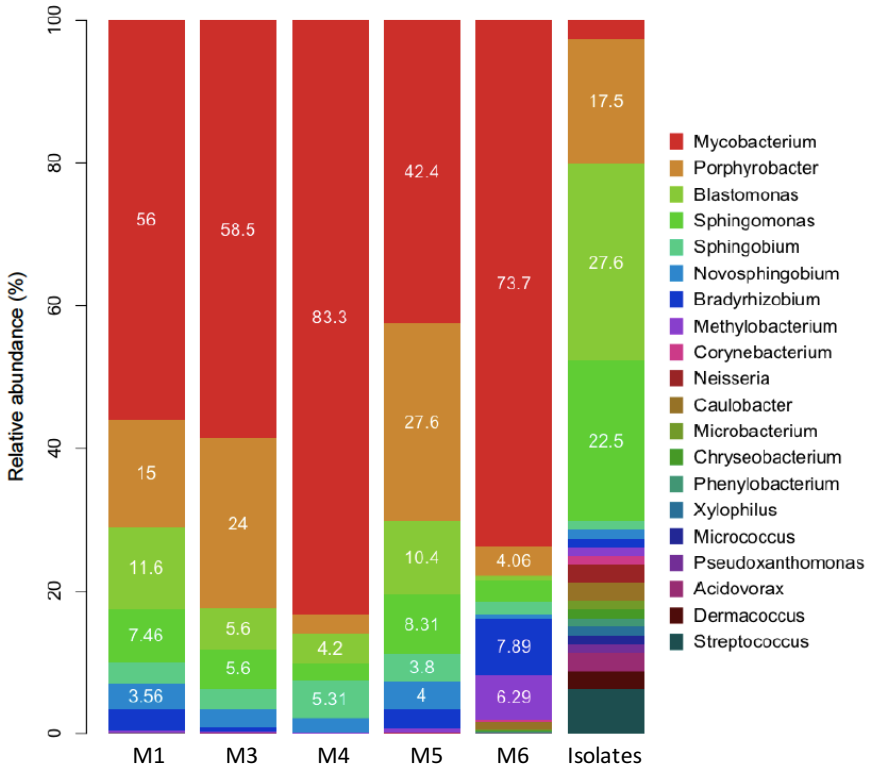


Figure 2.1. Taxonomic composition of the shower hose biofilms based on 16S rRNA gene fragments recovered from the metagenomes and isolates.

The relative abundance (y axis) of the 16S rRNA gene-encoding reads recovered from the metagenomes (normalized by the total number of classified 16S rRNA gene-encoding reads in each metagenome) and the cultured fraction (normalized by the number of isolates; last column) for the major genera present in each sample (x-axis) is shown.

Overall, in all five shower hose metagenomes the dominant population corresponded to a previously unclassified *Mycobacterium* sp., most closely related to *Mycobacterium rhodesiae* and *M. tusciae*, showing ~85% genome-aggregate average

nucleotide identity or ANI (60). The second most abundant population genome was affiliated with *Blastomonas*, which shared 77% of its proteins at ~84% Average Amino Acid Identity (AAI) to the closely related *Blastomonas* sp. AAP53 reference genome (Fig. 2.2, Table A.3).

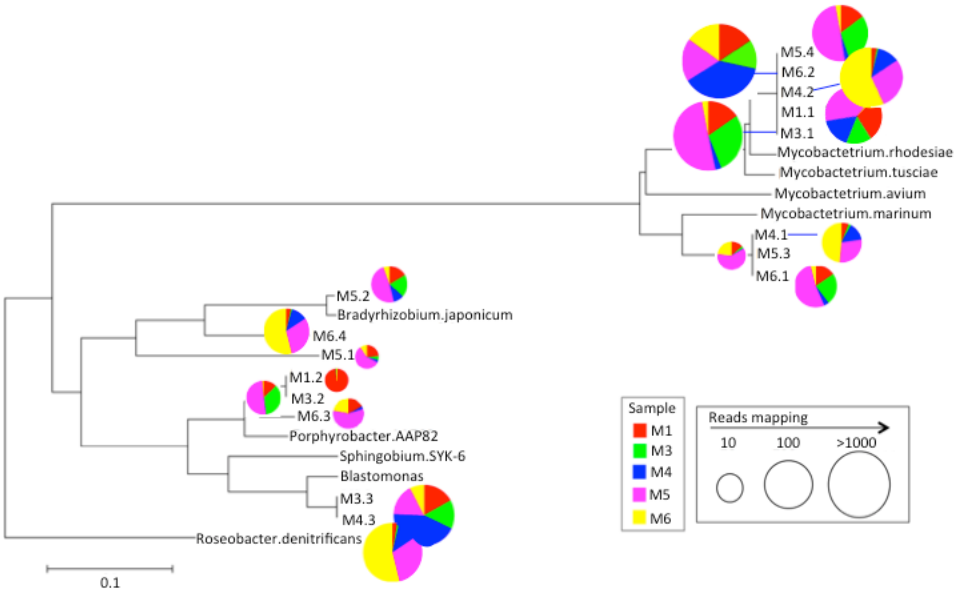


Figure 2.2. Phylogenetic relationships and relative abundance of the populations recovered in the shower hose metagenomes.

The tree shows all 30S ribosomal protein S9 sequences assembled from the metagenomes and selected reference sequences from publicly available genomes (denoted by complete species names). The radius of the pie charts indicates the number of reads mapping to the specific protein sequence related to the node, and the colors represent the five different datasets (see figure key). *Roseobacter denitrificans* was used as an out-group. The phylogenetic tree was constructed using Neighbor-Joining algorithm with 1000 bootstrap replicates in MEGA V.5 (91). Scale bar represents substitutions per site.

Analysis of partial 16S rRNA gene sequences of over 1850 R2A isolates revealed that the vast majority (>74%) belonged to the *Proteobacteria* phylum (data not shown). Specifically, 23% (22/94) of the isolates whose genomes were fully sequenced as part of

this study were affiliated with the genera *Blastomonas*, followed by *Sphingomonas* (18%), and *Porphyrobacter* (14%) (Table A.2). Several isolates were assigned to the genera *Streptococcus* (n=4), *Dermacoccus* (n=2), *Acidovorax* (n=4), *Neisseria* (n=3), and *Mycobacterium* (n=2) (Fig. 2.1).

A comparison of the recovered *Blastomonas* sp. population genome against the *Blastomonas* isolates showed an ANI of 99.9% (s.d 0.01) sharing approximately 91% of its protein sequences. These results suggest that the *Blastomonas* isolates are representatives of the population recovered in the metagenomes, presumably representing members of the same population (57). In contrast, the average ANI of the recovered *Mycobacterium* sp. population and the two isolates classified as *Mycobacterium* sp. indicated that indeed belong to the same genus but represent distinct populations and presumably species (ANI: 82.41%, s.d 0.02), and are low-abundance members of the biofilm community.

The discrepancy between the taxonomic profiles of the culture-dependent and culture-independent results was presumably attributable to the culture medium and growth conditions used, which favored the recovery of *Blastomonas sp.* (58). While many mycobacterial species can grow on R2A, it should be noted that some mycobacteria are slow growers and can take up to 8 weeks (or longer) to grow on media typically used for the propagation of mycobacteria (59). Nonetheless, the genome isolates were useful as reference genomes in evaluating genome coverage and confirming species identification (Fig. A.2). Our results also showed that a substantial fraction (20% or more) of drinking water microbial communities growing on the shower hose surfaces can be cultured with the described medium, contrasting with the 1-2% or less for several complex natural environments such as soils (60).

2.4.2 Presence of opportunistic pathogens

The taxonomic classification of metagenomic sequences, as well as genomes of isolates (see above), revealed the presence of potential opportunistic pathogens in shower hose biofilms (e.g., members of the *Sphingomonas*, *Rhizobium, Mycobacterium mucogenicum*, and *Neisseria perflava*). Notably, the most abundant population recovered

in the metagenomes (average relative abundance among samples 66.7% ± 8.21 of the total) represented a close relative of *M. rhodesiae* and *M. tusciae.* These two mycobacterial species are considered potential opportunistic pathogens since they have been identified as the causing agent of pulmonary and disseminated infections in immuno-compromised individuals (10, 61–63).

Phylogenetic analysis showed that the assembled protein sequences of this *Mycobacterium*-like population genome are linked to a novel species based on relatively low ANI values (~85%; Fig. A.2) to known mycobacteria species (57). Remarkably, the recruitment of metagenomic reads against the recovered mycobacterial genome revealed that this population was the most abundant and distinct from rare (less-abundant) co-occurring relatives in the samples. (Fig. A.4). Further, reads with more than 99% nucleotide identity to the reference represented around 62.5% of the total *Mycobacterium*-like sequences in the metagenomes and overlapping reads sampling the same part of the genome produced a star-like phylogeny (Fig. A.4), suggesting that this is an abundant and homogenous, clonal (or nearly clonal) population. Predicted proteins from this population shared 85.6% AAI (77% of the total number of proteins in the population bin) with *M. tusciae* and 86.2% AAI (76% of the total number of predicted proteins) with *M. rhodesiae*.

Functional annotation of the recovered *Mycobacterium* sp. population genome revealed a number of proteins related to virulence and host colonization previously identified in other NTM species, including *M. rhodesiae, M. smegmatis*, and *M. bovis* (Table A.3). In particular, our analysis identified several key proteins for i) biogenesis and central metabolism inside host cells such as pantothenate synthetase (*panC*), aspartate-1-decarboxylase (*panD*), and superoxide dismutase (*sodC*), together with genes for ii) insertion into the host cell via complement-mediated phagocytosis, including fibronectin-binding protein C (*fbpC2*) and the fibrinogen-binding protein (*fbpA*), and iii) for protection against oxygen-free radicals delivered by host cells, such as catalase-peroxidase (*katG*) and the sigma factor (*sigF*) (Table 2.1).

Table 2.1. Description of the proteins present in the metagenomes associated with biofilm formation, antibiotic and disinfectant resistance mechanisms and virulence.

Standard deviation (4th column) represents the variation observed among the five metagenomes. Relative abundance was based on the number of predicted proteins assigned to a particular function divided by the total number of annotated metagenomic proteins, previously mentioned on the methods.

Mechanism	Protein name	Gene	Average abundance (± standard deviation)
Biocide Resistance			
Attachment, invasion, and peroxide resistance	DNA binding protein	*dps*	0.27 (0.16)
Protective role, oxidative stress defense	Thioredoxin reductase	*trxB*	0.51 (0.05)
	Copper/zinc superoxide dismutase	*sodC*	0.07 (0.01)
	Putative alcohol dehydrogenase D	*adhD*	0.09 (0.08)
	Redox-sensitive transcriptional regulator	*soxR*	0.04 (0.01)
	Alkyl hydroperoxide reductase protein	*ahpF*	0.06 (0.02)
	Glutathione reductase	*gorA*	0.01 (0.008)
	Manganese superoxide dismutase	*sodA*	0.03 (0.02)
	RNA polymerase sigma factor	*rpoS*	0.1 (0.07)
	Hydrogen peroxide-inducible genes activator	*oxyR*	0.14 (0.04)
DNA repair	exodeoxyribonuclease III	*xthA*	0.12 (0.07)
Resistance to copper and silver	Cation efflux system protein	*cusA*	0.09 (0.03)
Multidrug efflux pump systems	Resistance nodulation division (RND) family	*acrB, mdtB*	0.02 (0.01)
	Multidrug resistance protein	*emrK,*	0.02 (0.01)
	ABC transporter ATPase	*PGP3*	0.01 (0.01)
Biofilm formation			
	Heat shock protein	*GroEL1*	0.27 (0.06)
Biosynthesis	Glutamate synthase	*gltB*	0.08 (0.02)
Growth	Putative membrane protein	*mmpL4*	0.07 (0.05)
Biofilm detachment	Glutathione synthetase	*ghsB*	0.08 (0.02)
Carbon metabolism	Phosphoenolpyruvate carboxykinase	*pckA*	0.05 (0.04)
Metabolism	Mycocerosic acid synthase	*mas*	0.05 (0.04)
	extracellular polymeric substance	*EPS*	0.07 (0.04)

Table 2.1 continued

Exopolysaccharide biosynthesis and Biofilm development	GDP mannose dehydrogenase	*algD*	0.09 (0.09)
Virulence and antigenic variation			
Possible role in virulence and antigenic variation	Uncharacterized PE-PGRS family protein	*PE_PGRS33*	0.22 (0.16)
Required for virulence	cholesterol oxidase	*choD*	0.06 (0.04)
	ABC transporter ATP-binding/permease protein	*Rv1747*	0.11 (0.09)
	Serine/threonine-protein kinase	*pknF*	0.09 (0.08)
	Probable cation-transporting ATPase	*ctpG*	0.04 (0.04)
	Probable copper-exporting P-type ATPase V	*ctpV*	0.03 (0.05)
Known Virulence factors			
Protection against oxygen free radicals	Peroxidase/catalase	*katG*	0.23 (0.04)
Increased resistance to reactive oxygen intermediates	Sigma factor	*sigF*	0.11 (0.08)
Secreted protein and virulence determinant factor	Glutamine synthase	*glnA1*	0.07 (0.04)
Facilitate the adhesion of bacteria to the mucosal surface	fibronectin binding proteins	*fbpC2* and *fbpA*	0.02 (0.01)
ESX-1 secretion system and DNA conjugation	Extracellular mycosin protease	*mycP1*	0.05 (0.03)
Transposition	Insertion element IS6110	*MRA_0012*	0.06 (0.04)
	Transposase for insertion sequence element IS1081	*YIA3_RHISP*	0.09 (0.06)
	Uncharacterized protein y4hP	*NGR_a03*	0.18 (0.13)
	Transposase for insertion sequence element IS6120	*PUV_09480*	0.11 (0.07)
	Uncharacterized protein y4jA/y4nE/y4sE	*NGR_a03150*	0.07 (0.04)
	Insertion element ISR1	*YIA3_RHISP*	0.06 (0.04)

We also identified members of the *Sphingomonas* genus in the shower hose biofilms. Members of this genus that have been previously isolated from hospital water sources and associated with urinary tract infections and peritonitis (12, 64). Notably, cases of bacteremia have been reported including one in a hospital in Taiwan (65) and other in a cardiovascular ICU in a hospital in Turkey (66). Since these reports are based on non-sequencing methods (e.g., pulse field gel electrophoresis and blood cultures), it was not possible to perform a more detailed comparison to the isolates and populations recovered in the present study. 16S rRNA gene sequence analysis showed that the closest relative for several of our isolates was *S. koreensis* (99% nucleotide identity), which has been identified as the causative agent of meningitis in at least one previous study (67). Taken together, it is likely that the *Sphingomonas* isolates recovered here could represent opportunistic pathogens.

2.4.3 Disinfectant resistance mechanisms

Several genes associated with resistance to disinfectants applied in municipal water treatment were recovered in both metagenome and isolate genomes. For instance, we recovered genes encoding proteins with participation in SoxR, OxyR, and SOS systems that have previously been experimentally identified as conferring protection against oxidative stress (68, 69). These functions were at least 10 times more frequent (i.e., number of distinct gene alleles detected) in the metagenomes relative to all completed bacterial genomes with similar genome sizes available in NCBI as of January 2016 (number of genomes used = 442; genome size in the range of 2 to 4Mb; t-test p-value 0.00065, on average), indicating that the shower hose environment selects for the functions. In addition, we identified multidrug efflux pump genes, including the ABC, SMR, and RND systems, which can confer resistance to disinfectants as well as antibiotics (for the latter, see below) in Gram-negative biofilm members affiliated with *Sphingomonas*, *Porphyrobacter*, and *Blastomonas*. All corresponding protein sequences showed high amino acid identity (> 40%) and conservation of functional domains with their experimentally verified homologs (Fig. 2.3, Table A.4).

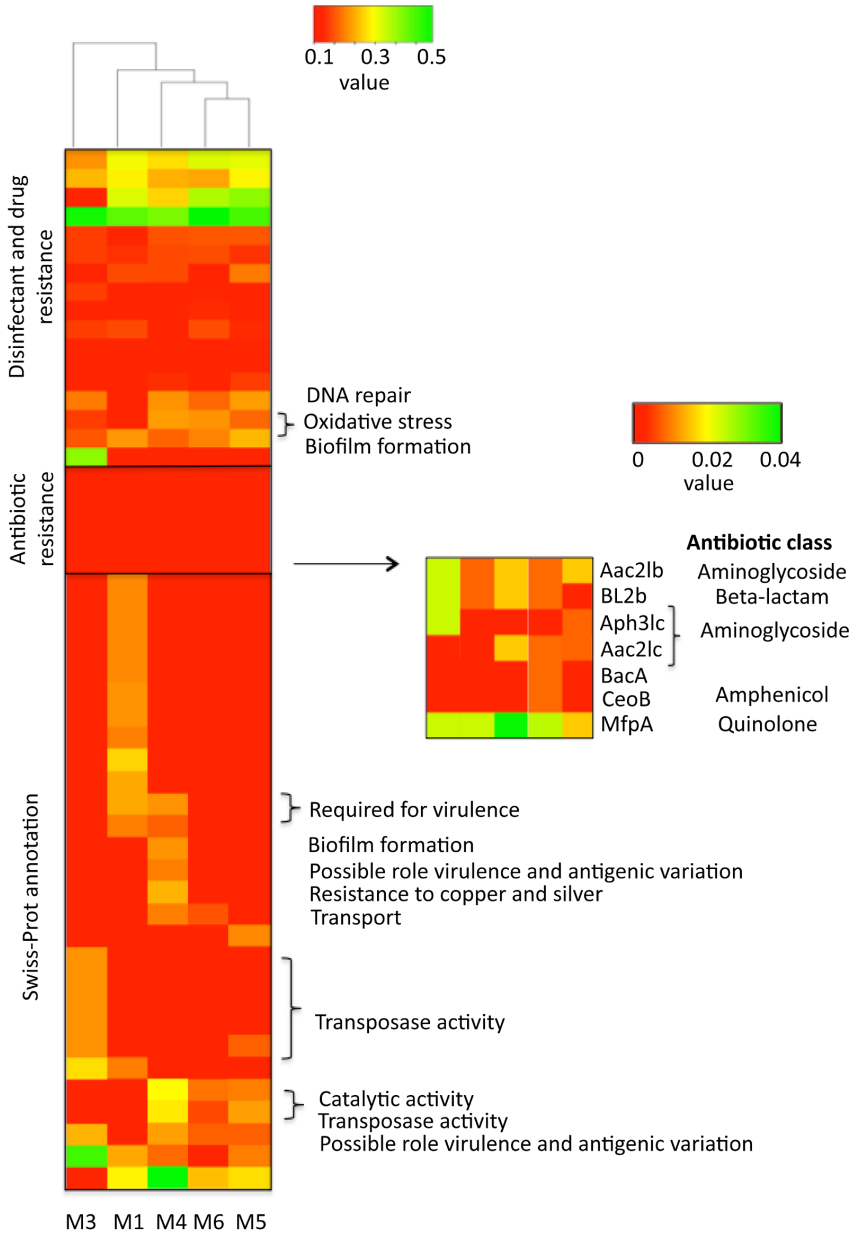


Figure 2.3 Relative abundance of functional genes in the shower hose metagenomes.

(From top to bottom) The heat map on the left is composed of 7 proteins involved in antibiotic resistance, 12 proteins involved in disinfectant resistance mechanisms and EPS production, and the 30 most abundant proteins annotated with UniProt DB (rows) for each sample (columns). The small heat map on the right represents a magnification of the main heat map, focusing on the antibiotic resistance genes (note the difference in scale). Antibiotic class denotes the classification of the antibiotics based on WHO ATC code J01

(WHO Collaborating Centre for Drug Statistics Methodology, available through http://www.whocc.no/atc_ddd_index/). The cladogram was constructed using complete linkage hierarchical clustering with Euclidean distance as implemented on gplot package in R (70). A detailed description of all the proteins plotted in the heat map is in the table S4.

2.4.4 Antibiotic resistance mechanisms

A BLAST analysis of the metagenomic proteins against the Antibiotic Resistance Genes Database (ARDB) revealed that the sampled organisms likely have proteins that underlie resistance to at least four distinct antibiotic classes: beta-lactamase, quinolone, aminoglycoside, and amphenicol. Overall, the M3 metagenome presented the highest percentage of cells encoding ARG, specifically beta-lactamase (*bl2B*) (23.1% of total), mycobacterial fluoro-quinolone resistance protein A (*mfpA*) (28.3%) that is involved in DNA mimicry mechanisms (67), and aminoglycoside 2'-N-acetyltransferase (*aac2Ib*) that acetylates aminoglycoside antibiotics preventing their binding to the bacterial ribosome (71) (Table 2.2). Thus, the dominance of *Mycobacterium* in sample M3 (53%) was also reflected in the antibiotic resistance profile of this sample since 66.6% of the contigs encoding ARG were phylogenetically affiliated with this genus.

Table 2.2. Abundance of antibiotic resistance genes recovered from the shower hose metagenomes.

The values represent the genome equivalents of each gene calculated using its sequencing depth divided by the normalizing factor of the corresponding dataset outlined in the Material and Methods section

ARG symbol	Function	Antibiotic resistance	Genome equivalents (%) M1	M3	M4	M5	M6
BL2b	Beta-lactamase	Penicillin, cephalosporin	10.9	23.1	6.2	0	2.5
MfpA	*Mycobacterium* fluoro-quinolone resistance protein A	Fluoroquinolone, ciprofloxacin and sparfloxacin	4.2	17.6	4.3	16.7	5.5
Aac2Ib	Aminoglycoside 2'-N-acetyltransferase	Netilmicin, tobramycin, dibekacin, gentamicin	1.3	38.3	2.7	0	2.8
Aph3Ic	Aminoglycoside O-phosphotransferase	Paromomycin, neomycin, kanamycin, lividomycin, ribostamycin, gentamincin_b	0	1.9	0	0.2	0
Aac2Ic	Aminoglycoside N-acetyltransferase	Gentamicin, netilmicin, tobramycin, dibekacin	0	0	0	18.7	31.6
CeoB	Resistance-nodulation-cell division transporter system. Multidrug resistance efflux pump	Chloramphenicol	0	0	0	0	0.5
BacA	Undecaprenyl pyrophosphate phosphatase	Bacitracin	0	0	0	0	0.4

The second most abundant population genome recovered from the metagenomes, which was also well represented among the isolates (unlike the abundant mycobacterial population), was assigned to *Blastomonas* and encoded genes likely conferring resistance to aminoglycoside, macrolide, and bacitracin antibiotics. Indeed, a comparison between the 16S RNA gene sequences obtained from the shower hose *Blastomonas* isolates in this study and those obtained from *Blastomonas* strains isolated from a tap water in Portugal (GenBank: HF930725.1) (61) revealed high sequence identity (> 97%); therefore, these two isolates likely represent the same or highly related species. The Portuguese tap water

isolate was highly resistant to antibiotics, based on an ATB PSE EU (bioMérieux) susceptibility test, mostly to the aminoglycoside antibiotic class, including gentamicin and tobramycin. This finding was consistent with the gene content predicted in the *Blastomonas* isolates of our study. In addition, other genes conferring resistance to penicillin, cephalosporin, paromomycin, neomycin, lividomycin, ribostamycin, and chloramphenicol were detected in the metagenomes, albeit at much lower abundances (present in <5% of the genome equivalents).

2.4.5 Comparisons to other similar environments

We compared 16S rRNA gene fragments recovered from the shower hose metagenomes against 16S rRNA gene sequences available from DWDS pipes located in Florida (72) and a surface in the intensive care unit (ICU) of a hospital ward in Spain (73). This analysis revealed distinct taxonomic profiles between these and our shower hose metagenomes (Fig. A.1). Most notably, the shower hose datasets presented higher abundances of sequences related to *Mycobacteriaceae* (an average of 38% in shower hose versus 0.03% in the ICU surface and 6% in pipes), followed by *Sphingomonadaceae* (18% in shower hose versus 5% in the ICU surface and 0.03% in pipes), and *Erythrobacteraceae* (13% in shower hose versus 0.09% in the ICU surface and 0% in pipes). Distinctively, members of the *Methylococcaceae* order dominated the DWDS pipe sample (83% of the total) but were essentially absent in the other two datasets. In addition, *Staphylococcaceae* and *Enterobacteriaceae* dominated the ICU ward surface of the hospital (22% and 20% of the total, respectively) but were in low abundance in the other datasets.

A comparison of the shower hose metagenomes with available metagenomes from diverse natural water ecosystems in similar temperate geographic regions indicated that the former were enriched in virulence factors (3.64% of total metagenomic reads) and antibiotic resistance functions (0.032%) compared to metagenomes from the Pearl River (China) (0.072% and 0.011% of total reads annotated as virulence factors and antibiotic resistance genes, respectively), and wintertime (0.071% and 0.010%, respectively) and summertime (0.340% and 0.072%, respectively) samples from Lake Lanier (Georgia,

USA). Compared to a drinking water treatment plant located in the Pearl River Delta in China (0.008%), the showerheads were enriched in these two functions (0.17%) (Fig. A.3).

2.5 DISCUSSION

This study analyzed biofilms of shower hoses in a hospital and found that most metagenomic sequences were associated with members of the genera *Mycobacterium, Erythrobacter, Sphingomonas*, and *Novosphingobium*. These findings were consistent with those from previous studies showing that mycobacterial populations are frequently abundant in DWDS because of their high resistance to chlorine, monochloramine, and other disinfectant compounds in water systems (1, 20, 74). The high abundance of mycobacterial populations in the shower hose biofilms contrasted with their low abundance or absence in microbial communities on the surface of the ICU in a hospital in Spain, which consisted predominantly of *Staphylococcaceae* and *Enterobacteriaceae* (Fig. A.1). The abundance of these two bacterial groups in the Spanish hospital could be the result of these organisms being continuously shed by incoming patients and hospital staff and therefore may not be waterborne in nature.

Other possible explanation for the dominance of mycobacteria-like sequences is related to the particular physicochemical features of the shower hose, such as pipes material either galvanized (zinc coated) or made of copper, and the disinfectants and low organic carbon content of the water, that selectively favor the growth of some mycobacterial populations (75, 76). Because of the identification of several pathogenicity factors and antibiotic resistance genes (Fig. 2.3), as well as its high relatedness to characterized NTMs (i.e., in terms of both gene content and amino acid similarity), the recovered *Mycobacterium* sp. population might represent an opportunistic pathogen. Therefore, our findings revealed that microbial biofilms in hospital shower hoses are characterized by distinct composition, including previously non-described species, which require more attention due to their potential implications for health (see also below). Nonetheless, it should be noted that this *Mycobacterium* sp. genome encoded, in general, fewer virulence factors compared to its close relatives *M. tusciae* (66.6% of total VFs of *M. tusciae* were

present in the *Mycobacterium* sp. population) and *M. rhodesiae* (80.9% of total VFs shared), indicating that this population might represent a member of NTM with comparatively less public health implications.

In addition, some of the isolates were affiliated with disease-causing bacteria. The isolates CCH10-H12 and CCH6-A12 were most closely affiliated with *Neisseria perflava* (98% and 100% 16S rRNA gene identity, respectively). This bacterium is a common oral commensal of the human upper respiratory tract, but occasionally, can cause endocarditis, peritonitis, and complicated bacteremia, mainly in individuals with immune suppression (77). Further, two *Mycobacterium* isolates were below the detection limit of our metagenomic effort (rare members of the biofilm) and most closely assigned to *M. mucogenicum* (100% 16S rRNA gene identity). Compared to other mycobacterial species, this is a fast-growing organism and is commonly involved in catheter-related infections and nosocomial outbreaks caused by contaminated hospital equipment and water sources (78, 79). The divergence between the *Mycobacterium* species recovered by culture-dependent and -independent methods was probably due to the fact that incubation time and culture media were not suitable for isolating the most abundant *Mycobacterium*, which was recovered with the herein used metagenomic approach (close relative to *M. rhodesiae* and *M. tusciae*). Therefore, even though the frequency that the aforementioned organisms cause infections is probably lower compared to some other most commonly encountered opportunistic pathogens based such as members of the *Burkholderia* and *Ralstonia* genera, it is quite likely that they represent a health risk, especially for immuno-compromised patients. Collectively, these findings suggest that more attention needs to be given to biofilms growing on shower hoses and other surfaces in clinical settings due to their potential to represent a health risk. Current and future studies held by the Hospital Microbiome and the Indoor Environment Projects (30, 80), analyzing hundreds of samples and from various hospital settings, would add to the picture of the microbial communities presented here, and the assessment of the associated risk for public health.

In addition to mycobacteria, members of other abundant genera present in the shower hose biofilms, namely *Porphyrobacter, Blastomonas*, and *Sphingomonas*, have also been frequently found in water-related environments such as swimming pools, bulk water, and

faucets, presumably because of their ability to survive disinfection regimes (3, 81). In particular, these bacterial groups are considered to play an important role in the formation and dynamics of biofilms because of their high production potential for EPS and ability to colonize surfaces (82). Members of these genera also have the ability to co-aggregate with other community members, contributing to an effective colonization and expansion of biofilms (81). In view of the frequent occurrence of *Sphingomonadaceae* in hospital tap water and their high survival in the air of the indoor environment, this group has been identified as frequent contaminant of medical devices (64, 66, 73). Although these organisms were less abundant in shower hose biofilms than mycobacteria (Fig. A.2), their occurrence in these environments may be linked to resistance to cleaning and disinfection due to known adaptive mechanisms and biofilm-forming ability.

Biocide agents have a strong influence on the bacterial community structure and may increase the frequency of antibiotic resistance bacteria (83). Exposure to chlorine can stimulate the expression of efflux pumps and drug resistance operons as well as induce mutations in some genes leading to increased antibiotic resistance (84). Some of the antibiotic resistance signatures observed in the shower hose metagenomes have been reported to be triggered by biocide exposure; these include the chloramphenicol-, kanamycin-, and penicillin-resistance genes (84, 85). Further, previous studies have observed that several *Mycobacterium* species can modify the cell membrane fatty acid composition in response to stress conditions, producing an altered permeability to biocide and antibiotic compounds (86, 87). Several of the known proteins that underlie the latter phenotype such as those involved in lipid metabolism and mycolic acid biosynthesis, e.g., long-chain-fatty-acid ligase (Facl), membrane protein (MmpL3), mycolic acid methyltransferase (MmaA), and GroEL, were encode in the shower hose metagenomes. Accordingly, the acquisition of the antibiotic resistance profile identified in the biofilm community may, to a certain extent, be directly influenced by chlorine exposure. However, directly testing this hypothesis and quantifying the effect of chlorine exposure would require additional experiments.

The bacterial populations recovered from the metagenomes were validated through analysis of presence/absence (completeness) and phylogenetic identity (contamination) of

single copy genes. These binned populations represented consistent biological units with a limited -if any- contaminating sequences from other populations based on the phylogenetic analysis of single copy genes (e.g., Fig. A.2). Also, the genome sequence of the isolates recovered from the same samples was used to validate several of the bins at almost complete, high-draft genomes (Table A.3). For example, the binned *Blastomonas* population genome showed high nucleotide identity values (ANI 99.9%, s.d 0.01) and remarkable synteny with the *Blastomonas* isolate genomes (Fig. A.5). In contrast to *Blastomonas*, the recovery of an abundant uncultivated *Mycobacterium* population, without known sequenced representatives and 100% completeness was achieved using binning approaches. The fact that a number of functional gene sequences were recovered using culture-dependent and culture independent approaches (i.e., both genome isolates and metagenomes) as well as the high relative abundance in-situ (e.g, *Mycobacteria* sp. and *Blastomonas* populations) suggest that many of the bacteria in these biofilms were alive, further highlighting their ability to withstand the harsh conditions within DW systems. Finally, although the variation in abundance of the dominant populations among the samples was, in general, limited, certain populations such as the *Mycobacteria sp.* showed substantial differences in abundance (e.g., Fig. 2.1). These differences were not attribute to the physicochemical parameters of the water of the shower hoses measured, which typically do not vary much among samples, or some characteristics (e.g., floor) of the hospital rooms sampled and thus, are likely due to random sampling events.

Altogether, the results reported here revealed novel metagenomic information relevant to microbial exposure in the built environment. As some of the identified mycobacterial populations are related to previously identified pathogens they may represent an uncharacterized pool of potential nosocomial pathogens, growing in biofilms attached to the showerhead surfaces. While further evidence is needed to determine if the abundant *Mycobacterium* sp. and some of other less abundant biofilm populations represent a high risk to patients and healthcare workers, the data suggest that they should be carefully examined due to their chlorine-resistant phenotype and presence of several important antibiotic resistance genes in their genomes. Because of the persistence of several community members across samples, the potential for release from the biofilm and adhesion to medical devices, and the presence of antibiotic resistance genes in the biofilm

community, our findings call for more attention to the biofilms growing on showerheads, as they might constitute a public health risk. In conclusion, our findings further highlight the increasing importance of metagenomic surveys to better understand the functional genetic network (or microbiome) in clinical settings and in DW distribution systems (20).

2.6 CONCLUSIONS AND PERSPECTIVES

In this study, we described the microbial communities found in shower hoses at a major U.S. hospital using cutting-edge metagenomic techniques. We identified potential pathogenic bacteria living inside the water supply pipes as well as genes for resistance to antibiotic and water disinfectants. The resulting insights are of practical importance for pathogen surveillance, epidemiologic investigations, and characterization of resistant determinants in health care settings.

Our data provide a foundation for new research into the microbiome network in indoor environments, especially in hospitals, where selective pressure of cleaning disinfectants and daily use of antibiotics can increase the prevalence of resistance. Besides, the chlorine compounds used in public drinking water may not provide sufficient protection for water supplies in these facilities. Further studies on profiling microbial communities from more hospitals are needed in order to evaluate whether similar biofilm communities would be found in other medical settings, how mechanical and chemical monitoring should be done, and how often shower heads and hoses should be replaced on a regular basis.

We recommend adding sequencing data and bioinformatics analyses to routine surveillance protocols in hospitals, where the presence of opportunist pathogens poses a threat to immunocompromised patients. High resolution protocols like shotgun metagenomics will help to increase precision on bacterial identification, monitor bacteria difficult to grow by traditional culture-based methods, and quantify and monitor genetic markers associated with resistance and virulence that represent a public health concern.

2.7 ACKNOWLEDGEMENTS

This study was supported by the U.S. Environmental Protection Agency. Soto-Giron M.J was supported by Colciencias - Colombian Government -doctoral scholarship. HR was the recipient of National Research Council Senior Research Fellowship. We thank Mark Rodgers for helping during sample collection and for providing comments on early drafts. We also thank Simoni Triantafyllidou for sharing physico-chemical data and related discussions. The U.S. Environmental Protection Agency, through its Office of Research and Development, partially funded, managed, and collaborated in the research described herein. This work has been subjected to the agency's administrative review and has been approved for external publication. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the agency; therefore, no official endorsement should be inferred. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

2.8 REFERENCES

1. Feazel LM, Baumgartner LK, Peterson KL, Frank DN, Harris JK, Pace NR. 2009. Opportunistic pathogens enriched in showerhead biofilms. Proc Natl Acad Sci 106:16393–16399.

2. Poitelon J-B, Joyeux M, Welté B, Duguet J-P, Prestel E, Lespinet O, DuBow MS. 2009. Assessment of phylogenetic diversity of bacterial microflora in drinking water using serial analysis of ribosomal sequence tags. Water Res 43:4197–4206.

3. Liu R, Yu Z, Guo H, Liu M, Zhang H, Yang M. 2012. Pyrosequencing analysis of eukaryotic and bacterial communities in faucet biofilms. Sci Total Environ 435–436:124–131.

4. Berry D, Xi C, Raskin L. 2006. Microbial ecology of drinking water distribution systems. Curr Opin Biotechnol 17:297–302.

5. Revetta RP, Gomez-Alvarez V, Gerke TL, Curioso C, Santo Domingo JW, Ashbolt NJ. 2013. Establishment and early succession of bacterial communities in monochloramine-treated drinking water biofilms. FEMS Microbiol Ecol 86:404–414.

6. Araújo P, Lemos M, Mergulhão F, Melo L, Simões M. 2011. Antimicrobial resistance to disinfectants in biofilms.

7. Schwartz T, Kohnen W, Jansen B, Obst U. 2003. Detection of antibiotic-resistant bacteria and their resistance genes in wastewater, surface water, and drinking water biofilms. FEMS Microbiol Ecol 43:325–335.

8. Falkinham JO, Pruden A, Edwards M. 2015. Opportunistic Premise Plumbing Pathogens: Increasingly Important Pathogens in Drinking Water. Pathog Basel Switz 4:373–386.

9. Williams MM, Armbruster CR, Arduino MJ. 2013. Plumbing of hospital premises is a reservoir for opportunistically pathogenic microorganisms: a review. Biofouling 29:147–162.

10. Shin JH, Lee EJ, Lee HR, Ryu SM, Kim HR, Chang CL, Kim YJ, Lee JN. 2007. Prevalence of non-tuberculous mycobacteria in a hospital environment. J Hosp Infect 65:143–148.

11. Szymańska J. 2007. Bacterial contamination of water in dental unit reservoirs. Ann Agric Environ Med AAEM 14:137–140.

12. Kilic A, Senses Z, Kurekci AE, Aydogan H, Sener K, Kismet E, Basustaoglu AC. 2007. Nosocomial outbreak of *Sphingomonas paucimobilis* bacteremia in a hemato/oncology unit. Jpn J Infect Dis 60:394–396.

13. Guinto CH, Bottone EJ, Raffalli JT, Montecalvo MA, Wormser GP. 2002. Evaluation of dedicated stethoscopes as a potential source of nosocomial pathogens. Am J Infect Control 30:499–502.

14. Vornhagen J, Stevens M, McCormick DW, Dowd SE, Eisenberg JNS, Boles BR, Rickard AH. 2013. Coaggregation occurs amongst bacteria within and between biofilms in domestic showerheads. Biofouling 29:53–68.

15. Parker BC, Ford MA, Gruft H, Falkinham JO. 1983. Epidemiology of infection by nontuberculous mycobacteria. IV. Preferential aerosolization of *Mycobacterium intracellulare* from natural waters. Am Rev Respir Dis 128:652–656.

16. Primm TP, Lucero CA, Falkinham JO. 2004. Health impacts of environmental mycobacteria. Clin Microbiol Rev 17:98–106.

17. Wallace RJ, Brown BA, Griffith DE. 1998. Nosocomial outbreaks/pseudo-outbreaks caused by nontuberculous mycobacteria. Annu Rev Microbiol 52:453–490.

18. Carter G, Wu M, Drummond DC, Bermudez LE. 2003. Characterization of biofilm formation by clinical isolates of *Mycobacterium avium*. J Med Microbiol 52:747–752.

19. Nishiuchi Y, Tamura A, Kitada S, Taguri T, Matsumoto S, Tateishi Y, Yoshimura M, Ozeki Y, Matsumura N, Ogura H, Maekura R. 2009. *Mycobacterium avium* complex organisms predominantly colonize in the bathtub inlets of patients' bathrooms. Jpn J Infect Dis 62:182–186.

20. Gomez-Alvarez V, Revetta RP, Santo Domingo JW. 2012. Metagenomic analyses of drinking water receiving different disinfection treatments. Appl Environ Microbiol 78:6095–6102.

21. Joseph O. Falkinham. 2003. Mycobacterial Aerosols and Respiratory Disease. Emerg Infect Dis J 9:763.

22. Amoils S. 2009. Showering with bacteria. Nature 461:360.

23. Hilborn ED, Covert TC, Yakrus MA, Harris SI, Donnelly SF, Rice EW, Toney S, Bailey SA, Stelma GN. 2006. Persistence of nontuberculous mycobacteria in a drinking water system after addition of filtration treatment. Appl Environ Microbiol 72:5864–5869.

24. Hussein Z, Landt O, Wirths B, Wellinghausen N. 2009. Detection of non-tuberculous mycobacteria in hospital water by culture and molecular methods. Int J Med Microbiol IJMM 299:281–290.

25. Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, Tu Q, Deng Y, He Z, Shi JZ, Yuan MM, Sherry RA, Li D, Luo Y, Schuur EAG, Chain P, Tiedje JM, Zhou J, Konstantinidis KT. 2014. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Appl Environ Microbiol 80:1777–1786.

26. Rodriguez-R LM, Overholt WA, Hagan C, Huettel M, Kostka JE, Konstantinidis KT. 2015. Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. ISME J 9:1928–1940.

27. Furuhata K, Kato Y, Goto K, Saitou K, Sugiyama J-I, Hara M, Fukuyama M. 2007. Identification of yellow-pigmented bacteria isolated from hospital tap water in Japan and their chlorine resistance. Biocontrol Sci 12:39–46.

28. McLean JS, Lombardo M-J, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, Tesler G, Nurk S, Lesin V, Brami D, Hall AP, Edlund A, Allen LZ, Durkin S, Reed S, Torriani F, Nealson KH, Pevzner PA, Friedman R, Venter JC, Lasken RS. 2013. Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. Genome Res 23:867–877.

29. McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS. 2013. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci 110:E2390–E2399.

30. Lax S, Gilbert JA. 2015. Hospital-associated microbiota and implications for nosocomial infections. Trends Mol Med 21:427–432.

31. U.S EPA. 1994. Method 200.8: Determination of Trace Elements in Waters and Wastes by Inductively Coupled Plasma-Mass Spectrometry,. 5.4. Analytical Method, Environmental Protection Agency, Cincinnati.

32. Reasoner DJ, Geldreich EE. 1985. A new medium for the enumeration and subculture of bacteria from potable water. Appl Environ Microbiol 49:1–7.

33. Ryu H, Henson M, Elk M, Toledo-Hernandez C, Griffith J, Blackwood D, Noble R, Gourmelon M, Glassmeyer S, Santo Domingo JW. 2013. Development of quantitative PCR assays targeting the 16S rRNA genes of Enterococcus spp. and their application to the identification of enterococcus species in environmental samples. Appl Environ Microbiol 79:196–204.

34. Revetta RP, Matlib RS, Santo Domingo JW. 2011. 16S rRNA gene sequence analysis of drinking water using RNA and DNA extracts as targets for clone library development. Curr Microbiol 63:50–59.

35. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

36. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J 6:898–901.

37. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829.

38. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272.

39. 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, 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.

40. 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 J Comput Mol Cell Biol 19:455–477.

41. Luo C, Rodriguez-R LM, Konstantinidis KT. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res 42:e73.

42. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267.

43. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141-145.

44. Zhu W, Lomsadze A, Borodovsky M. 2010. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38:e132–e132.

45. Altschul SF, Lipman DJ. 1990. Protein database searches for multiple alignments. Proc Natl Acad Sci U S A 87:5509–5513.

46. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191.

47. Liu B, Pop M. 2009. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res 37:D443-447.

48. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, Pan Y, Li J, Zhu L, Wang X, Meng Z, Zhao F, Liu D, Ma J, Qin N, Xiang C, Xiao Y, Li L, Yang H, Wang J, Yang R, Gao GF, Wang J, Zhu B. 2013. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat Commun 4:2151.

49. Haft DH, Selengut JD, Brinkac LM, Zafar N, White O. 2005. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21:293–306.

50. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29-37.

51. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

52. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26.

53. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.

54. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560.

55. Albertsen HM, Chettier R, Farrington P, Ward K. 2013. Genome-wide association study link novel loci to endometriosis. PloS One 8:e58257.

56. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJC, Yoo HS, Zhang C, Zhang Y, Sobral BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:D581-591.

57. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91.

58. Narciso-da-Rocha C, Vaz-Moreira I, Manaia CM. 2014. Genotypic diversity and antibiotic resistance in Sphingomonadaceae isolated from hospital tap water. Sci Total Environ 466–467:127–135.

59. Tsukamura M. 1983. Numerical classification of 280 strains of slowly growing mycobacteria. Proposal of *Mycobacterium tuberculosis* series, *Mycobacterium avium* series, and *Mycobacterium nonchromogenicum* series. Microbiol Immunol 27:315–334.

60. Torsvik V, Øvreås L. 2002. Microbial diversity and function in soil: from genes to ecosystems. Curr Opin Microbiol 5:240–245.

61. Tortoli E, Kroppenstedt R, Bartoloni A, Caroli G, Jan I, Pawlowski J, Emler S. 1999. *Mycobacterium tusciae* sp. nov. Int J Syst Bacteriol 49 Pt 4:1839–1844.

62. Curry EM, Yehia M, Roberts S. 2008. CAPD peritonitis caused by *Mycobacterium rhodesiae*. Perit Dial Int J Int Soc Perit Dial 28:97–99.

63. Rahman SA, Singh Y, Kohli S, Ahmad J, Ehtesham NZ, Tyagi AK, Hasnain SE. 2014. Comparative analyses of nonpathogenic, opportunistic, and totally pathogenic mycobacteria reveal genomic and biochemical variabilities and highlight the survival attributes of *Mycobacterium tuberculosis*. mBio 5:e02020.

64. Hsueh P-R, Teng L-J, Yang P-C, Chen Y-C, Pan H-J, Ho S-W, Luh K-T. 1998. Nosocomial Infections Caused by Sphingomonas paucimobilis: Clinical Features and Microbiological Characteristics. Clin Infect Dis 26:676–681.

65. Lin J-N, Lai C-H, Chen Y-H, Lin H-L, Huang C-K, Chen W-F, Wang J-L, Chung H-C, Liang S-H, Lin H-H. 2010. *Sphingomonas paucimobilis* bacteremia in humans: 16 case reports and a literature review. J Microbiol Immunol Infect Wei Mian Yu Gan Ran Za Zhi 43:35–42.

66. Meric M, Willke A, Kolayli F, Yavuz S, Vahaboglu H. 2009. Water-borne Sphingomonas paucimobilis epidemic in an intensive care unit. J Infect 58:253–255.

67. Marbjerg LH, Gaini S, Justesen US. 2015. First report of *Sphingomonas koreensis* as a human pathogen in a patient with meningitis. J Clin Microbiol 53:1028–1030.

68. Pagán-Ramos E, Song J, McFalone M, Mudd MH, Deretic V. 1998. Oxidative stress response and characterization of the oxyR-ahpC and furA-katG loci in *Mycobacterium marinum*. J Bacteriol 180:4856–4864.

69. Dhandayuthapani S, Mudd M, Deretic V. 1997. Interactions of OxyR with the promoter region of the oxyR and ahpC genes from *Mycobacterium leprae* and *Mycobacterium tuberculosis*. J Bacteriol 179:2401–2409.

70. R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

71. Aínsa JA, Martin C, Gicquel B, Gomez-Lus R. 1996. Characterization of the chromosomal aminoglycoside 2'-N-acetyltransferase gene from *Mycobacterium fortuitum*. Antimicrob Agents Chemother 40:2350–2355.

72. Kelly JJ, Minalt N, Culotti A, Pryor M, Packman A. 2014. Temporal variations in the abundance and composition of biofilm communities colonizing drinking water distribution pipes. PloS One 9:e98542.

73. Poza M, Gayoso C, Gómez MJ, Rumbo-Feal S, Tomás M, Aranda J, Fernández A, Bou G. 2012. Exploring bacterial diversity in hospital environments by GS-FLX Titanium pyrosequencing. PloS One 7:e44105.

74. Perkins SD, Mayfield J, Fraser V, Angenent LT. 2009. Potentially pathogenic bacteria in shower water and air of a stem cell transplant unit. Appl Environ Microbiol 75:5363–5372.

75. Falkinham JO. 1996. Epidemiology of infection by nontuberculous mycobacteria. Clin Microbiol Rev 9:177–215.

76. Wang H, Edwards M, Falkinham JO, Pruden A. 2012. Molecular survey of the occurrence of *Legionella* spp., *Mycobacterium* spp., *Pseudomonas aeruginosa*, and amoeba hosts in two chloraminated drinking water distribution systems. Appl Environ Microbiol 78:6285–6294.

77. Lavigne JP, Le Bayon A, Michaux-Charachon S, Arich C, Bouziges N, Campello C, Sotto A. 2004. [*Neisseria subflava* subsp. perfiava bacteremia: a case study and literature review]. Med Mal Infect 34:331–332.

78. Adékambi T. 2009. Mycobacterium mucogenicum group infections: a review. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis 15:911–918.

79. Kline S, Cameron S, Streifel A, Yakrus MA, Kairis F, Peacock K, Besser J, Cooksey RC. 2004. An outbreak of bacteremias associated with *Mycobacterium mucogenicum* in a hospital water supply. Infect Control Hosp Epidemiol 25:1042–1049.

80. Arnold C. 2014. Rethinking Sterile: The Hospital Microbiome. Environ Health Perspect 122:A182–A187.

81. Rickard AH, Leach SA, Hall LS, Buswell CM, High NJ, Handley PS. 2002. Phylogenetic relationships and coaggregation ability of freshwater biofilm bacteria. Appl Environ Microbiol 68:3644–3650.

82. Bereschenko LA, Stams AJM, Euverink GJW, van Loosdrecht MCM. 2010. Biofilm formation on reverse osmosis membranes is initiated and dominated by *Sphingomonas* spp. Appl Environ Microbiol 76:2623–2632.

83. Webber MA, Whitehead RN, Mount M, Loman NJ, Pallen MJ, Piddock LJV. 2015. Parallel evolutionary pathways to antibiotic resistance selected by biocide exposure. J Antimicrob Chemother 70:2241–2248.

84. Karumathil DP, Yin H-B, Kollanoor-Johny A, Venkitanarayanan K. 2014. Effect of chlorine exposure on the survival and antibiotic gene expression of multidrug resistant *Acinetobacter baumannii* in water. Int J Environ Res Public Health 11:1844–1854.

85. Huang J-J, Hu H-Y, Tang F, Li Y, Lu S-Q, Lu Y. 2011. Inactivation and reactivation of antibiotic-resistant bacteria by chlorination in secondary effluents of a municipal wastewater treatment plant. Water Res 45:2775–2781.

86. Steed KA, Falkinham JO. 2006. Effect of growth in biofilms on chlorine susceptibility of *Mycobacterium avium* and *Mycobacterium intracellulare*. Appl Environ Microbiol 72:4007–4011.

87. Armstrong JL, Calomiris JJ, Seidler RJ. 1982. Selection of antibiotic-resistant standard plate count bacteria during water treatment. Appl Environ Microbiol 44:308–316.

88. Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res 40:D641-645.

CHAPTER 3. THE STRUCTURE OF THE HUMAN GUT MICROBIOME ACROSS A RURAL-TO-URBAN GRADIENT IN NORTHERN ECUADOR

3.1 ABSTRACT

The gut microbiota plays a key role in modulating gut homeostasis and prevention of invasion by pathogens. Previous studies have reported lower fecal bacterial diversity in urban compared to rural settings. However, most of these studies compare geographically distant populations (e.g., different countries, even continents). Focusing on how gut microbial communities differ along a rural-to-urban gradient in the same country undergoing urbanization may provide new insights and avoid confounding factors such as different cultural context or diet. Furthermore, how the bacterial diversity pattern along the gradient plays a role, if any, during diarrheal diseases remain poorly understood. In an attempt to provide new insights into these issues, we sampled the gut microbiome of subjects living in Quito (capital city) and nearby villages in Northern Ecuador and profiled the gut microbiota during acute diarrheal disease (ADD). Metagenomic analyses from young children revealed distinct taxa in rural vs. urban populations, including higher abundance of *Prevotella* on average (49.4% vs. 11.4%, respectively, $P < 0.05$), and lower abundance of *Bacteroides* (16.1% vs. 36.2%, $P < 0.05$) and *Alistipes* (2.1% vs. 8.5%, $P < 0.05$). Metagenomic samples during ADD showed greater shifts in functional pathways, taxon abundance, and the predicted number of taxon-taxon interactions in urban samples than the rural ones. Collectively, our data indicate differentially abundant microbial taxa and metabolic pathways between individuals from rural and urban populations that may play a role in the response to ADD.

3.2 INTRODUCTION

To date, urbanization has globally increased with more than 53% of the total human population living predominantly in large cities (1). Massive movement of populations from rural to urban areas is occurring rapidly in developing countries. Moreover, the increase of westernized urban practices including eating behavior (diet rich in fat, simple sugars, and animal proteins), reduction in physical activity, increased use of antibiotics, and hygiene practices has led to an elevated prevalence of metabolic diseases including obesity, type 2 diabetes, and immunological complications (2). These factors have profoundly impacted the ecology, diversity, and functionality of the gut microbiome (3–7).

In order to understand the relationship between lifestyle and gut microbiota, several investigations have compared the taxonomic profile of gut microbes between traditional agriculturalist societies (e.g., Malawi, Yanomami, Burkina Faso, and hunter-gatherers from Tanzania) and western populations (e.g., United States) (4–6, 8–14). The emerging picture form these previous studies is that rural populations around the world harbor higher fecal bacterial diversity than urban-industrialized populations. Further, it has been shown that microbiota diversity is quickly lost, in just a few generations, in industrialized populations, where some low abundant community members can become extinct (15–17). For instance, members of the genus *Treponema* have been found in the gut of rural traditional populations but not in urban-industrialized gut microbiota (5). On the other hand, urbanization had influenced the acquisition of specific bacterial taxa associated with a diet high in protein and animal fat (12).

Many of the aforementioned studies showing the importance of geographical factors in determining gut microbiome composition are based on 16S rRNA gene amplicon sequencing. However, this approach provides insufficient genetic resolution to capture intraspecific variation and whole-genome functional potential. Additionally, most studies have been focused on comparing populations with from distinct geographical regions, without taking into account confounding factors such as different cultural practices or diet preferences. A small number of recent studies have found lifestyle factors to be important in the microbial composition and functional metabolic properties along an urbanization

gradient in the same geographical area, e.g., same country or region (8, 16, 18). Nonetheless, whether or not changes in the gut microbiome diversity due to urbanization are associated with lower resilience to enteric infections and/or higher frequency of disease remains essentially unknown.

Changes in microbial composition and reduction in bacterial diversity have also been associated with decreased resilience (7), facilitating the colonization of the gut by pathogenic bacteria and the loss of keystone species, altering the homeostasis of the gut microbiota, and producing a dysbiotic system (19, 20). Ecological theory (21, 22) predicts that high gut microbial diversity may confer increased resilience to perturbation and colonization by enteric pathogens, especially in less-urbanized populations with more diverse gut microbiomes. However, this hypothesis remains to be fully tested in the gut environment.

In this study, we explored the functional implications of differences in gut microbial community composition and attempted to test the hypothesis that microbial diversity may confer increased resilience to perturbation especially in less-urbanized populations. We examined metagenomic profiles of fecal samples from individuals living in rural and urban areas of Ecuador and characterized the value of this diversity for resilience, defined as the level of taxonomic or functional shifts during perturbation compared to the control (non-ADD) state, to acute diarrheal disease (ADD). Our analysis reveals distinctive taxonomic and metabolic features between individuals living in Quito (Ecuador's capital city) and in villages located in a rural area of Esmeraldas Province. Comparison between non-ADD and ADD states in each group suggests a distinctive response to ADD between urban and rural subjects showing greater functional and taxonomic shifts in subjects from urban areas (less resilience).

3.3 MATERIALS AND METHODS

3.3.1 Study population

Initially, 800 fecal samples were collected between April– September 2015 from approximately 200 individuals (100 ADD and 100 non-ADD) living in Quito (Ecuador's capital), Esmeraldas, the town of Borbón, and nearby rural communities (Villages) along the Onzole, Cayapas, and Santiago Rivers (Fig. B.1.A). From this, a subset of 411 samples were subjected to 16S rRNA gene amplicon sequencing (Quito: 87, Esmeraldas: 74, Borbón: 128, and villages: 122) including cases of acute diarrheal disease that resulted PCR positive for the presence of marker genes specific for DAEC (Diffusely Adherent *E. coli*), EPEC (Enteropathogenic *E. coli*), ETEC (Enterotoxigenic *E. coli),* EIEC (Enteroinvasive *E. coli*), EAEC (Enteroaggregative *E. coli*). These diarrhea samples were age-matched with control individuals where no pathogenic *E. coli* was detected by PCR. Samples from Quito and villages were used to compare urban and rural areas since these two groups represent the start and end points of the urbanization gradient. From this, a subset, 31 out of 87 samples from Quito and 32 out of 122 samples from the villages from individuals between one and six years old were subjected to shotgun metagenomics.

The ages of the participants from both locations ranged between 0 months to 78 years. In this study, participants were categorized by age according to the following criteria: new born: 0 to 6 months, babies: 7-12 months, young children: 13 months to 3 years, children 4 to 7 years, pre-adolescents: 8 to 17 years, adults: 18 to 74 years, and senior: > 74 years. Subjects from the villages generally presented low educational attainment levels and limited economic resources (23). More details about the region of study can be found in (23–25). Quito represents the urban area with a population size of ~2.67 millions and approximately 43% of the population lives under national poverty. The city presents an elevation of 2,580 m above sea level. The annual mean temperature is 13.4°C (26). On the other hand, villages represent the rural area with an estimated population between 10-500 inhabitants and an elevation of 15m above sea level. The average temperature monthly is 30°C ± 2°C (27).

The inclusion criteria comprised individuals visiting the clinic with acute diarrhea (more than 3 loose stools in a 24-hour period; ADD samples), and controls (non-ADD samples) were individuals visiting the same clinic for any other reason, who did not have symptoms of diarrhea or vomiting in the prior week. Both cases and controls were excluded

if they had taken antibiotics in the prior week, or if they had not lived in the study location for at least six months prior. Written informed consent and assent (for children) were obtained from each participant. This study was supported by NIAID Grant number K01AI103544 and approved by the Institutional Review abroad (IRB) of Emory University and the Universidad San Francisco de Quito (USFQ). Participants were administered a survey to collect information about lifestyle and demographic factors including water consumption (source of drinking water, treatment), sanitation practices, contact with animals, recent travel, and other factors.

3.3.2 Sample collection

Fecal samples from participants were collected in two cryo-conservation tubes and stored in a liquid Nitrogen dewar until being transferred to a -80ºC freezer at the USFQ laboratory. DNA was extracted using the Wizard Genomic DNA Purification kit (Promega, Madison, WI). Amplicon sequencing of the V4 region of the 16S rRNA gene was performed using primers 515F and 806R tailed with Illumina adapters P5 and P7 (28). The tagged amplicons were submitted onto the MiSeq instrument and sequenced on a 2 x 250 bp run. For shotgun metagenomic sequencing, libraries were prepared using the Illumina Nextera XT DNA library prep kit and an equimolar mixture of the libraries was sequenced on an Illumina HiSeq instrument on a 2 x 150 bp paired end run.

3.3.3 16S rRNA gene sequence analysis

Quality control and processing of raw paired-end reads were performed using DADA2 (29) incorporated in Qiime2 version 2017.9 (30). DADA2 denoise-paired plugin was used to trim low quality regions of the sequences (less than Q30), remove chimeras, dereplicate sequences, and finally produce an amplicon sequence variant (ASV) table (hereinafter referred to as Operational Taxonomic Unit (OTU) because of its analogy with the OTU table) correcting for amplicon errors and identifying single-nucleotide differences. To taxonomically classify the ASVs, QIIME2 q2-feature-classifier plugin and the Naive Bayes classifier was used together with the Greengenes13.8 99% OTUs database (31). QIIME2 q2-diversity module was used to calculate alpha and beta diversity indexes

based on a sampling depth of 8000 reads/sample for all samples. This number of reads was used because coverage curves of randomly selected samples suggested that more than 99% of the community diversity was covered/sampled at this level.

For alpha diversity, the number of observed OTUs and Faith's Phylogenetic Diversity index were calculated. Shannon diversity index with the Chao Shen correction (32) was calculated using the entropy package v1.2.1 (33) available in R v3.3.1. For beta diversity, Jaccard and Bray-Curtis distances were calculated and the distance matrices were the input for Principal Coordinate Analyses (PCoA). PCoA plots were visualized with EMPeror (34). Permutational Multivariate Analysis of Variance (PERMANOVA) was performed at the OTU level on the abundance table of control samples to evaluate the effect of geographical factors on the microbial composition. PERMANOVA was performed with the vegan package (Adonis function) (35) in R v3.3.1 using the Bray-Curtis dissimilarities among samples and 1,000 permutations. Significant associations between microbial and geographical variables were identified by applying a multivariate lineal model, MaAsLin (Multivariate microbial Association by Linear models) (36) to the OTU table of control samples (non-ADD) from each location. The *r*-coefficient and Q-value were calculated at different taxonomic levels, and associations were considered as significant at Q-values below 0.1, after correcting for multiple testing.

3.3.4 Microbial network analysis

Network analysis was conducted on the OTU table of samples from rural and urban subjects using SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference). SPIEC-EASI combines two algorithms for neighborhood selection and sparse inverse covariance selection in order to estimate an interaction graph from the data (37). Networks were visualized with Cytoscape v3.6.1, an interactive platform (38).

Network topology analysis including clustering coefficient, average node connectivity, number of edges, number of nodes, average path length, network diameter, network density, among others were calculated using NetworkAnalyzer (39). The degree distribution was calculated using igraph v1.2.1 (40) package available in R v3.3.1. This

function is defined as the fraction of nodes with degree *k*, where *ki*: node degree and n_k: number of nodes with degree *k*:

$$P(k_i = k) = P(k) = \frac{n_k}{\sum_k n_k} = \frac{n_k}{n}$$

3.3.5 Metagenomic sequence analysis

Raw reads from the metagenomes were analyzed in MIGA (Microbial Genome Atlas) (https://microbial-genomes.org) (41) for trimming, removing of Illumina adaptors, assembling, and predicting genes on assembled reads. Nonpareil v3.0 (42) with default parameters was used to estimate the average coverage and diversity (similar to Shannon index) for each sequenced library. Mash distances (43) were calculated using a kmer=25 and visualized in an NMDS (Non-metric multidimensional scaling) plot using the ecodist (44) and vegan (35) packages in R v3.3.1. Taxonomic classification of the short-read metagenomes was determined using MetaPhlan2 with default parameters (45) and the functional profile using HUMAnN2 with default parameters (46).

Statistical analyses of taxonomic and functional profiles between samples for each group and during non-ADD and ADD states were performed with STAMP v2.1.3 software (47). Welch's t-test was used to compare relative abundances between the two locations and the Tukey-Kramer post-hoc for pairwise comparisons and identifying which category differs. Correction for multiple comparisons was adjusted using the Benjamini-Hochberg FDR method (q-value).

3.3.6 Recovery of genome populations in the metagenomes

Assembled contigs larger than 1Kb from each sample were binned into metagenome-assembled genomes (MAGs) using MaxBin2 with default parameters (48). Completeness and contamination of MAGs were estimated using CheckM v1.0.5 with the lineage_wf parameter (49). MAGs with >85% completeness and <8% contamination were selected for subsequent analyses. Phylogenetic reconstruction of 114 high quality MAGs was based on universal single copy proteins identified using the HMM.essential.rb script (50). Proteins

were aligned using MUSCLE v3.8.31 (51) and concatenated using the Aln.cat.rb script (50). Maximum likelihood phylogeny of the concatenated alignment was built using RAxML v8.0.19 (PROTGAMMAAUTO, -f a, -N 100) (52). MAGs were annotated using Prokka v1.10 with default parameters (53) and predicted genes were mapped to the UniProt/SwissProt database (54) using BLASTp v2.2.29+ (55) (minimum amino acid identity, ≥ 40% and query aligned length, ≥70% for a match). UniProt ids were cross-reference with Gene Ontology (GO) terms (56, 57) for assigning biological processes.

3.3.7 Identification of pathogenic E. coli in ADD metagenomes through bioinformatics

E. coli was identified as the probably etiological agent of diarrhea based on the integration of four criteria:

1. The *in-situ* metagenomic abundance of the pathogenic *E. coli* isolate should be higher in ADD vs. non-ADD samples, after one accounts for reads representing commensal *E. coli* populations; the latter reads identified by a competitive search against the isolate genome and that of the commensal *E. coli* strain HS (NC_009800.1). To estimate abundance, metagenomics reads were mapped to the *E. coli* isolate genome of the metagenome assembled genome (MAG) or to a reference commensal *E. coli* genome (strain HS). Recruitment of the mapped reads to the *E. coli* genome was performed using the scripts and the workflow previously described in (50). The average sequencing depth of *E. coli* was calculated using the read recruitment output and the enveomic.R v1.3 package with default parameters in R v3.3.1. The abundance of *E. coli* was estimated as the average sequencing depth multiplied by the genome size and divided by the metagenome size.

2. The pathotype-specific toxins and virulence factors should be detectable in the metagenomes at similar (or higher) abundances than pathogenic *E. coli* and/or present in the *E. coli* MAG recovered from the metagenome. To identify virulence genes, BLASTn searches of metagenomics reads against the nucleotide sequence of virulence genes considered as marker for pathogenic *E. coli* were used to

calculate the average gene sequence depth using the BlastTab.seqdepth_ZIP.pl script. Genes with sequencing depth values ≥1X and query aligned length ≥70% were consider as a presence.

3. The degree of intra-population diversity (or clonality) of the pathogenic *E. coli* population should be lower (more clonal) compared to the *E. coli* population in non-ADD (control) samples. Clonality was measured as follows: the function enve.reploct2.ANIr in the enveomic.R v1.3 package was used to calculate ANI based on metagenomics reads (ANIr) that mapped *E. coli* reference genome with higher nucleotide identity that 95%. ANIr values between 99% and 100% were considered high clonality zone.

4. Epidemiology of clonal complex the isolate was assigned to, i.e., whether other isolates in the same complex were associated more strongly with ADD vs. non-ADD samples (Table B.8).

3.4 **RESULTS**

3.4.1 Geographic location has an effect on the gut microbiota composition

Analyses of microbial richness and diversity based on 16S rRNA gene amplicons (Shannon index, Phylogenetic diversity, and Observed OTUs) indicated that rural and urban populations harbor a similar microbial composition, with minor, mostly insignificant differences between the two groups (Fig. 3.1A). Nonetheless, when Bray-Curtis dissimilarity distances were plotted using Principal Coordinate Analysis (PCoA), bacterial communities from non-ADD subjects living in urban areas segregate from the rural ones (PERMANOVA, $P < 0.05$) (Fig. 3.1B). Since several samples were intermixed between the two groups, the travel pattern of the corresponding human subjects, including the number of times reported to have traveled to the nearest town (Borbón), to a city (Esmeraldas), and other communities in the past year, was evaluated to test whether or not traveling could explain the intermixing. No significant association between travel pattern

and the overlap of the samples was found (PERMANOVA, P>0.05). Comparison with non-ADD subjects living in the USA (58) indicated that the bacterial communities from subjects living in Ecuador (Quito and villages) were strongly distinct (PERMANOVA, P= 0.01) (Fig. 3.1C) and present higher microbial richness (average Chao1 271.6 ±61.3 vs. 102± 36.9). A linear discriminant analysis (LEfSe) (59) indicated differential taxa underling this separation such as higher proportions of *Bacteroides*, *Clostridium*, *Coprococcus*, and *Faecalibacterium* in U.S subjects while *Prevotella*, *Teponema*, *Desulfobrivio*, and *Fusobacterium* in Ecuadorian subjects (Fig. B.2).

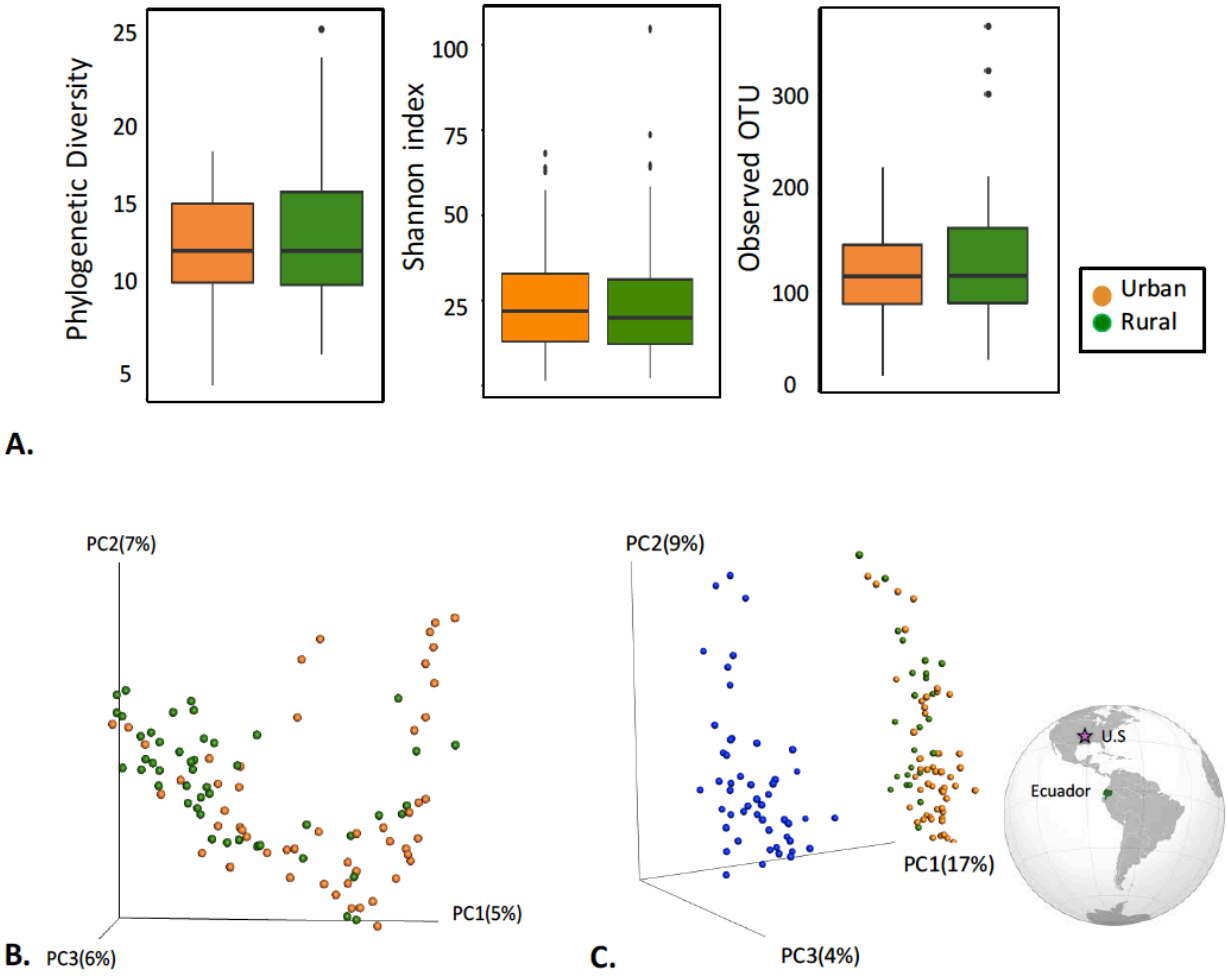


Figure 3.1 Diversity comparisons between rural and urban microbiomes based on 16S rRNA gene amplicon sequences.

A. Alpha diversity boxplots at the OTU level (Shannon index, phylogenetic diversity, and Observed number of OTUs) for each location. B. Beta-diversity PCoA plot based on Bray–Curtis distances of OTU similarities among samples (PERMANOVA, $P = 0.001$). Dots are colored by urban/rural status. C. Beta-diversity PCoA plots based on unweigthed Unifrac distances (PERMANOVA, $P = 0.001$) comparing Ecuadorian populations against the USA (blue dots). Levels of significance: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

PERMANOVA confirmed that location (rural vs. urban) showed a significant effect on the composition of the gut microbiota (Adonis $R^2 = 0.02$, $P = 0.001$), along with age and race (Adonis $R^2 = 0.06$, $P = 0.001$, $R^2 = 0.02$, $P = 0.027$, respectively) (Table B.2). The remaining factors analyzed (delivery mode, education, house sanitation, water treatment, water treatment type, gender, $P > 0.05$) did not present significant correlations with the microbial community structure (Table B.2).

In order to identify specific microbial taxa that were mainly responsible for the differences observed between rural vs. urban populations, we applied a multivariate association with linear model (36) in samples from non-ADD subjects controlling for age and race (Table B.3). These results suggested that *Prevotella copri* (average abundance 37.5% ± 23.6 vs. 19.5% ±20) and members from the *Comamonas* genus (average abundance 0.02% ± 0.04 vs. 0) and the *Elusimicrobiaceae* family (average abundance 1.1% ± 3 vs. 0.0.5% ±0.06) (lowest taxonomic classification provided by DADA2) were positively associated with rural settings, while members from the *Rikenellaceae* family (average abundance 2% ± 2.3 vs. 0.3% ± 0.5) were inversely associated with this location.

3.4.2 OTU networks in non-ADD rural vs. urban microbiomes

We additionally explored the inter-microbial relationships in the gut microbiota from non-ADD subjects living in rural and urban settings using OTU networks (37). We

found that the rural network presented a higher number of nodes (OTUs, 395 vs. 324, respectively) and edges (connections among taxa, 801 vs. 652, respectively) than the one from urban subjects (Table B.4). The number of positive associations among taxa was higher than that of negative ones in both networks (85.4% vs. 14.6% in rural and 83% vs. 17% in urban of the total edges).

Taxonomic affiliations of the 10 most connected OTUs differed between the two networks (Table B.5). In the rural network, *Bacteroides uniformis* showed the highest number of connections (n=12), followed by members of the family S24-7 of the order of *Bacteroidales* (n=10), and *Oscillospira* (n=10). On the other hand, OTUs affiliated with *Oscillospira* presented the highest number of edges (n=14), followed by *Bifidobacterium* (n=10), and members of the *Erysipelotrichaceae* family (n=10) in the urban network. The OTU classified as *Oscillospira* was the only highly-connected OTU found in both groups. This taxon is a butyrate producer and able to metabolize glucuronate, an animal-derived sugar, offering beneficial effects on human health (60).

3.4.3 Metagenome-based resolution of differences between urban and rural microbiomes

Taxonomic differences: To get a higher resolution of the microbiome structure associated with rural and urban lifestyles, whole-genome shotgun metagenomics was applied to a subset of samples (31 samples from Quito and 32 from villages) from subjects between one and six years old, in order to constrain the effect of age, living in Quito and the villages. Most of the metagenomics samples from Quito clustered together while those from the villages were more spread in the ordination plot based on Mash similarity distances (PERMANOVA, $P<0.05$) (Fig. B.3), somewhat consistent with the 16S rRNA gene-based results reported above.

Inspection of taxonomic profiles of non-ADD subjects indicated the presence of taxa with differential abundance between rural and urban microbiomes (Fig. 3.2A-2B), despite a high inter-person microbiome variation. Overall, *Prevotella* was the most prevalent taxon detected in rural samples, with variations in abundance among subjects

ranging from 7.43% up to 90.2% and encompassing more than 50% of the total community in more than half (9/17) of the samples. *Prevotella*'s relative abundance was also significantly higher in the rural vs. urban microbiomes (49.4% ± 28.6 vs. 11.4% ± 22.1, respectively; Tukey-Kramer post-hoc test, $P < 0.05$), consistent with 16S rRNA gene-based surveys mentioned above. At the species level, the fraction of the *Prevotella* signal among the rural metagenomes was dominated by *P. copri* (e.g., up to 100% of the total fraction in 8/17 samples) and *P. stercorea* (e.g., 93% in one sample). The other two taxa (*Comamonas* and *Elusmicrobiaceae*) that were associated with rural subjects based on 16S rRNA gene data did not show signatures of differential abundance in the metagenomes. This pattern was likely attributable to the fact that these taxa were mostly abundant in adult samples while the metagenomes were derived from young children (from 1 to 6 y.o).

In contrast, *Bacteroides* was more abundant in urban samples (Tukey-Kramer post-hoc test, $P < 0.05$). Among the 19 identified *Bacteroides* species, *B. dorei* was the most abundant (e.g., covering more than 93% of the total *Bacteroides* population in two samples), followed by *B. caccae* (e.g., more than 60% in two samples), and *B. vulgatus* (49% in one sample). *Alistipes* was also more abundant in urban metagenomes. Among the eight identified *Alistipes* species, *A. shahii* was the most abundant (e.g., 95% of the total *Alistipes* population in one sample), followed by *A. finegoldii* (e.g., 100% in one sample), and *A. putredinis* (e.g., more than 70% in three samples).

Functional gene differences: Functional annotation of metagenomic reads from all samples was performed using HUMAnN2 and the KEGG database (Fig. 3.2C). The predicted microbial functions highlighted differences in 35 KEGG pathways associated with non-ADD samples between the two groups. In particular, biosynthesis pathways including eight nucleotide, seven amino acid (L-lysine biosynthesis and S-adenosyl-L-methionine), five co-factors (flavin, folate, coeznyme A, and biotin), four secondary metabolites pathways (methylerythritol phosphate and chorismate), and eight glycolysis and carbohydrate metabolism pathways showed significant differences in abundances. Among these, pathways related to ribonucleotide (average relative abundance in the metagenome 0.0012 ± 0.0004 vs. 0.0008 ±0.0002), lysine (average 0.0008 ± 0.0003 vs. 0.0006 ± 0.0002), methionine (average 0.0008 ± 0.0003 vs. 0.0005 ± 0.0002), and aromatic

amino acid biosynthesis showed significantly higher abundance in the microbiota of rural than the urban samples. Some of these pathways may play key roles in maintaining an intestinal homeostasis. For instance, chorismate is a precursor for many bacterial metabolic pathways (61) and amino acid metabolism by bacteria is thought to be an important modulator of diverse physiological processes (62). Correlation analysis between bacterial taxonomic and functional richness indicated that in rural samples the number of predicted genes is significantly correlated with the number of OTUs present in the metagenomes (Pearson's $r = 0.71$, $P < 0.01$) (Fig. B.4).

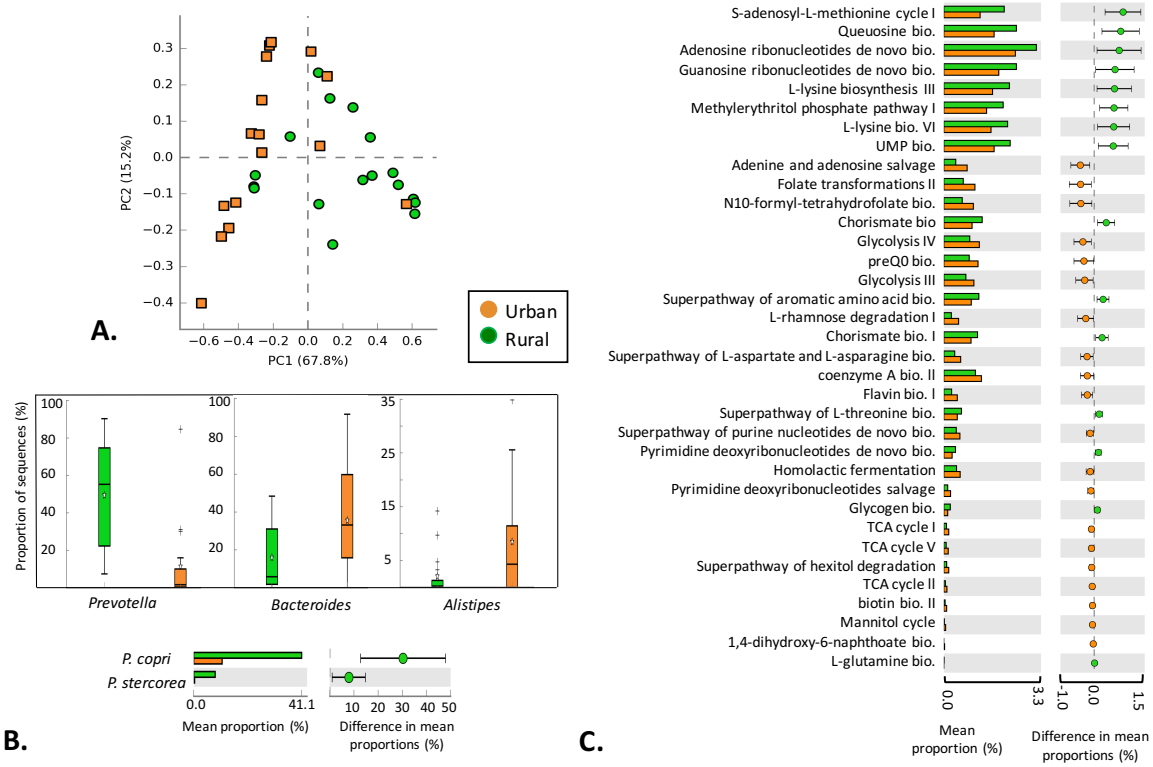


Figure 3.1 Differences in taxon and gene function abundances in non-ADD metagenomes from urban vs. rural subjects.

A. Principal component analysis of relative abundances of microbial members at the family level between the two groups of samples. B. Bar plots showing the proportion of sequences assigned to each differentially abundant taxon with the median (black central line) and the mean (star symbol). Small panels at the bottom indicate the mean proportion differences

and confidence intervals at 95% of *P. copri* and *P. prevotella* in rural and urban microbiomes (Tukey-Kramer post-hoc test, $P < 0.05$). Plots and statistical tests were performed using STAMP software v 2.1.3. C. Heatmap showing the relative abundance of 35 KEEG pathways that were significantly differentially abundant between the two groups (corrected-*p* value < 0.05 after multiple comparisons).

3.4.4 Diversity of Prevotella and Alistipes MAGs across the rural-to-urban gradient

To precisely characterize individual bacterial populations distinctive of each group, MAG analysis was performed. In total, 117 MAGs were recovered from non-ADD subjects from rural and urban settings (52 and 65, respectively). Taxonomic assignment using CheckM indicated that the majority of MAGs were assigned to the order of *Bacteroidales* (41%), followed by *Clostridiales* (25%), and *Enterobacteriales* (6%) (Fig. B.5). Among the recovered populations, a higher number of *Prevotella* MAGs were obtained from rural metagenomes than the urban ones (5 MAGs vs. 1 MAG, respectively), whereas *Alistipes* MAGs were only recovered from urban metagenomes, confirming the differentially abundant 16S rRNA gene-based taxa reported in non-ADD metagenomes between the two groups (Table B.6). Taxonomic assignment of *Prevotella* and *Alistipes* MAGs was confirmed by estimating the Average Nucleotide Identity (ANI) between the MAG and reference genomes from NCBI (NCBI_Prok) using MiGA. This analysis showed that most of the *Prevotella* MAGs (67%) represented uncharacterized species, defined at the 95% ANI level, with closely related species previously described (63) (Table B.7), providing higher resolution that the 16S rRNA gene-based results reported above.

We further examined the metabolic potential of *Prevotella* and *Alistipes* MAGs from the urban-to-rural gradient, revealing 50 biological processes (Fig. B.6.A) and 71 enzymatic reactions to be differentially abundant between these two population genomes (Fig. B.6.B) (Welch's t-test, $P < 0.05$ with Benjamini-Hochberg FDR correction). Functional annotation of protein-coding genes indicated that *Prevotella* MAGs harbored a higher number of genes related to amino acid (arginine, methionine), cofactors (pyridoxal phosphate), and nucleotide biosynthesis (adenine, NAD, AMP, purine salvage), and

metabolic processes (arginine and guanine catabolism) than *Alistipes* MAGs. Several of these pathways have been associated with beneficial effects in the host such as synthesis of essential amino acids, energy production, protein synthesis, and intestinal barrier function (62, 64), in addition to acting as precursors for several microbial metabolic pathways (65, 66). Moreover, some of the enzymes encoded in the *Prevotella* MAGs participate in polysaccharide metabolism (e.g., glycoside hydrolases, glycosyl transferases) as well as amino acid metabolism and thus, are likely associated with diet. *Prevotella* has been previously associated with a diet high in fiber, carbohydrate, vegetables, and egg food items (67–69).

On the other hand, *Alistipes* MAGs showed a higher number of predicted pathways and enzymes associated with carbohydrate (ribose, pentose-phosphate), phospholipid (cardiolipin), and secondary metabolism, cell response to starvation, and cellular protein modification process. *Alistipes* has previously been associated with a meat-based diet (70). Finally, when comparing functionally rural vs. urban whole-communities, several of the broad functional categories found to be differentially present between *Alistipes* and *Prevotella* (e.g., amino acid and nucleotide biosynthesis and carbohydrate metabolism) were also observed at the whole-community level, albeit the exact genes or pathways were not always the same.

3.4.5 Microbiome changes during diarrheal episodes

We profiled the gut microbiota during diarrheal episodes in subjects from the two groups in order to evaluate the existence of any significant differences in the microbiome during ADD and the role of the taxonomic and functional differences identified above from the non-ADD (control) state comparisons. Comparison of ADD samples between rural and urban populations based on 16S rRNA gene data indicated that samples from urban subjects presented lower number of observed OTUs than those from rural subjects (average number= 99 vs. 128, respectively; Welch's t-test, $P < 0.05$) despite the comparable number of OTUs at the non-ADD state (see above). A significant decrease in diversity during diarrheal episodes in urban subjects when compared to non-ADD samples was also

observed (Shannon index with the Chao Shen correction average 20 ± 13 vs. 27.1 ± 16; Welch's t-test, $P < 0.05$) (Fig. B.7).

OTU networks were compared between non-ADD and ADD states from rural and urban subjects and indicated a change in the network topology and connectivity patterns during the disease state (Fig. B.8, Table B.4). Specifically, the urban network showed a greater reduction, by 30% or more, relative to the non-ADD network in both the number of nodes (OTUs) and connections during diarrheal episodes. Further, many of the OTUs with the highest number of connections appeared to be lost during ADD (Fig. B.8). On the other hand, the rural network presented 6% fewer nodes and 21% fewer connections during ADD than the one from the non-ADD state. This network also presented a reduction of the most connected OTUs but the effect seems to be less pronounced than the one in the urban network.

At the metagenomic level, significant shifts in abundance during diarrheal episodes were observed in members of the *Desulfovibrionaceae* family in rural samples, while members of the *Bacteroidaceae*, *Porphyromonadaceae*, and *Ruminococcaceae* families changed more in abundance in the urban metagenomes (Tukey-Kramer post-hoc test, $P < 0.05$). A decreased abundance in these taxa has been previously associated with infectious diarrheal episodes (71, 72). In addition, five obligate anaerobes were significantly depleted in diarrheal metagenomes from urban samples (unclassified *Subdoligranulum*, *Desulfovibrio piger*, *Roseburia hominis*, *Parabacteroides distasonis*, *Ruminococcus obeum*; Tukey-Kramer post-hoc test, $P < 0.05$), which has also been observed previously (72–74).

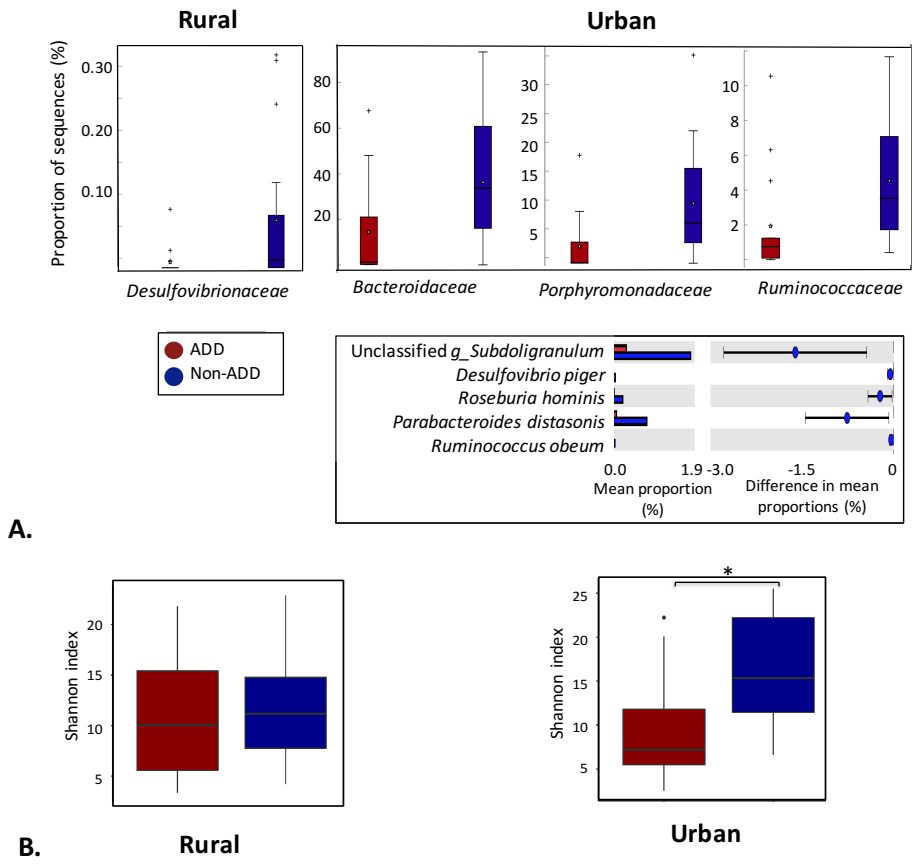


Figure 3.3 Comparison of taxon and gene function abundances between ADD and non-ADD samples in urban and rural metagenomes.

A. Bar plots of the relative abundance at the family level between the two groups of samples. Differences in mean proportion with the associated confidence intervals at 95% of individual species between cases and controls identified in rural populations are also shown (Tukey-Kramer post-hoc test, $P < 0.05$). Plots were produced using STAMP v 2.1.3 (47). B. Metabolic diversity (Shannon index) of KEGG pathways identified in the metagenomes. Colors represent the clinical status: ADD vs. non-ADD.

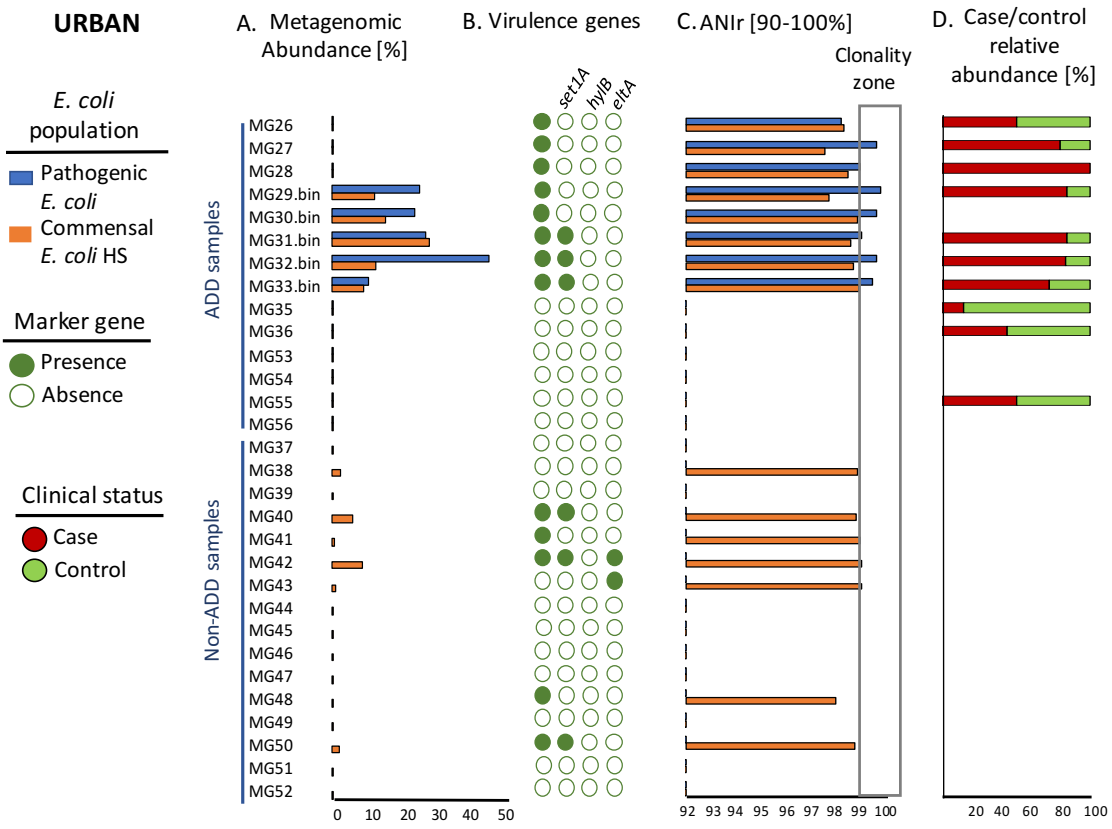
When comparing predicted functional pathways encoded by metagenomics reads between non-ADD and ADD samples, the metabolic diversity was significantly lower in urban vs. rural metagenomes during diarrhea (Welch's t-test, $P < 0.05$) (Fig. 3.3B) and a higher number of predicted metabolic pathways changed in abundance in urban samples (28 vs. 3 pathways, respectively; Fig. B.9.A), consistent with the taxonomic results

reported above. The bulk of the former pathways were involved in the biosynthesis of various co-factors (coenzyme A, pyridoxal phosphate, folate, N10-formyl-tetrahydrofolate) and amino acids (ornithine, arginine and polyamine) as well as carbohydrate metabolism (mannan degradation, gluconeogenesis, glycolysis). A decreased relative abundance of some of these pathways (e.g., amino acid biosynthesis, glycolysis and pentose phosphate pathway) has been previously reported in patients with *Clostridium difficile* infection (75). On the other hand, the metabolic pathways with a depleted abundance during diarrhea in rural populations were involved in sulfur oxidation, urea cycle, and L-isoleucine biosynthesis (Fig. B.9.B). These pathways have been associated with an elevated demand for energy production during mucosal inflammation and diarrhea affecting the urea cycle and amino acid levels in the colonic mucosal tissue (76, 77).

Metabolic pathways that showed increased abundance during diarrhea in urban samples were affiliated with pathogens including *E. coli*, *Shigella*, and *Haemophilus* (Fig. B.9.C). These pathways have been previously reported as key participants for maintaining pathogen viability and virulence. For instance, acyl-carrier protein biosynthesis is involved in primary and secondary metabolic pathways such as the formation of lipopolysaccharides (LPS), activation of exogenous fatty acids, and haemolysin synthesis (78–80). Heme biosynthesis is a vital mechanism of pathogens during infection (81). Lipid A (endotoxin), is the active component of lipopolysaccharide (LPS), which is an important pathogen-associated antigen that stimulates host immune responses (78, 82). These findings indicated that at least some of diarrheal cases were caused by the abovementioned pathogens, which warranted further investigation in order to isolate the effect of specific enteric pathogens from the taxonomic and gene function differences between rural and urban ADD samples revealed by our analysis.

3.4.6 Metagenomic comparison of ADD samples from rural and urban subjects after excluding cases of *E. coli infections*

In order to access the presence of pathogenic *E. coli* and its possible association with ADD, the abundance, clonality, and virulence profile of pathogenic *E. coli* in the metagenomes were estimated (see Materials and Methods). Results of these analyses indicated that five ADD samples (MG29, MG30, MG31, MG32, and MG33) from urban and two (MG57 and MG58) from rural subjects present evidence that most likely pathogenic *E. coli* was the causative agent of the infection (Fig. 3.4). Specifically, this set of samples exhibited the following metagenomics signatures: 1) higher abundance of pathogenic *E. coli* compared to the reference commensal, on average (10.3%±14 vs. 0.6%±0.8); 2) recovery of the diagnostic *E. coli* virulence factors for the isolate that was recovered from the same sample; 3) reduced intra-population diversity with ANIr values ≥ 99% for the pathogenic *E. coli* population and usually lower values for the reference commensal genome, and finally, 4) the recovered pathotype isolate(s) from the samples were generally grouped in phylogenetic clusters with isolates originated from cases of diarrhea than control (non-ADD) samples. To remove the signal of *E. coli* infection, these metagenomes were excluded and the analyses described above were repeated.


URBAN
E. coli population
Pathogenic E. coli
Commensal E. coli HS
Marker gene
Presence
Absence
Clinical status
Case
Control
A. Metagenomic Abundance [%]
B. Virulence genes
C. ANIr [90-100%]
Clonality zone
D. Case/control relative abundance [%]
set1A
hylB
eltA
ADD samples
Non-ADD samples
MG26
MG27
MG28
MG29.bin
MG30.bin
MG31.bin
MG32.bin
MG33.bin
MG35
MG36
MG53
MG54
MG55
MG56
MG37
MG38
MG39
MG40
MG41
MG42
MG43
MG44
MG45
MG46
MG47
MG48
MG49
MG50
MG51
MG52
0 10 20 30 40 50
92 93 94 95 96 97 98 99 100
20 40 60 80 100

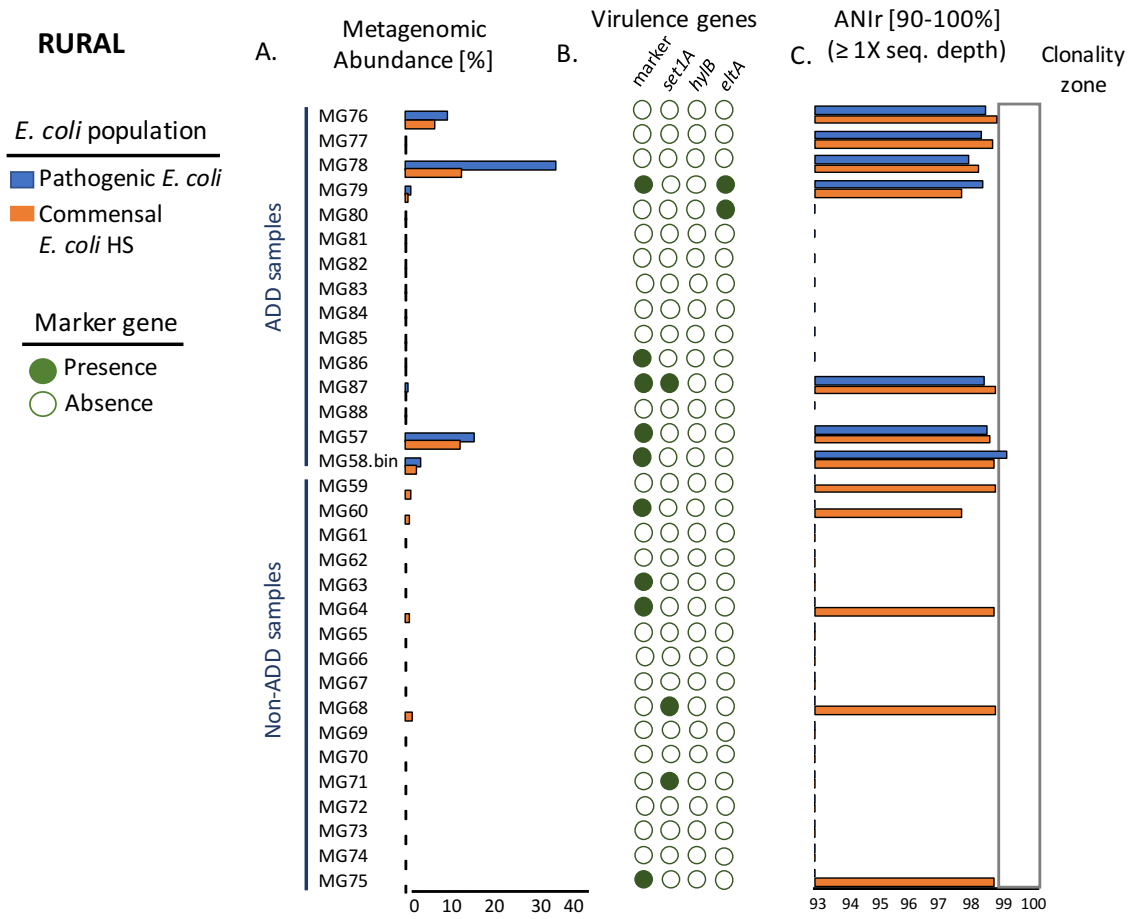


Figure 3.4. Identification of diarrheal cases (presumably) caused by pathogenic *E. coli*.

A. Comparison of the relative abundance of commensal (*E. coli* strain HS; in brown) and pathogenic *E. coli* (reference isolate genome or MAG obtained from the same sample; in blue) estimated based on the coverage of the reference genome by metagenomic reads. B. Presence of virulence genes in the metagenomes. The first column shows the marker gene specific for each *E. coli* pathotype represented by the recovered isolate (EAEC: *aggR*, EIEC: *ipaH*, DAEC: *afa*, EPEC atipica: *eaeA*, EPEC tipica: *bfp*) followed by genes encoding toxins (*hylB*, *set1A*, and *eltA*). C. *E. coli* intra-population diversity measured by ANIr calculated for both commensal (brown) and recovered pathogenic *E. coli* isolate (blue). To avoid potential biases by low *in-situ* abundance, only pathogenic *E. coli* with average sequence depth values ≥1X, were evaluated for ANIr analysis. D. Epidemiological data based on *E. coli* strains isolated from individuals living in urban and rural areas in

Ecuador. The percentage represents the frequency of cases and controls in the clonal complex that the *E. coli* strain isolated from the sample was assigned to (epidemiology).

Similar results were found in the taxonomic profile during ADD after excluding samples with pathogenic *E. coli* from both groups. Specifically, rural metagenomes showed a decreased abundance in microbial members affiliated to *Desulfovibrionaceae* (0.023% ± 0.07 vs. 0.1% ± 0.04) while urban metagenomes showed a reduction in *Ruminococcaceae* (0.49% ± 0.6 vs. 1.9% ± 2), *Porphyromonadaceae* (1.2% ± 1.5 vs. 4.8± 5.7), and *Coriobacteriaceae* (0 vs. 0.006% ± 0.01) abundances. In terms of functional profile, the metabolic diversity tended to be lower during diarrhea especially in urban samples, but the differences were not significant, similarly to the analysis with all samples included. In addition, the decreased abundance in urban samples of metabolic pathways involved in carbohydrate metabolism and in the biosynthesis of co-factors during ADD observed previously was also apparent. However, functional pathways with an increased abundance during ADD that were affiliated with *E. coli* were not significantly increased this time when compared to a non-ADD state. In the case of rural metagenomes, the same functional pathways (oxidation, urea cycle, and L-isoleucine biosynthesis) showed a decreased in abundance during ADD as before. Overall, the ADD metagenomic signal (shifts in abundance of microbial taxa and metabolic functions) was maintained after removing samples with pathogenic *E. coli*.

To evaluate whether metabolic changes during diarrhea were driven by the presence of a few taxa or represented instead a community-wide response to diarrheal episodes, the number of taxa that contribute to each pathway was calculated from the pathway abundance output file generated by HUMAnN2. This analysis indicated that the microbial response to ADD in the selected samples was influenced by taxon-specific shifts in both rural and urban samples but presenting different microbial members that are involved in the functional disturbance (Fig. B.10). For instance, *Ruminococcus bromii, Coprococcus sp* ART55.1, and *Treponema succinafaciens* were associated with the differential abundance observed in the three metabolic pathways in rural samples. On the other hand,

Faecalibacterium prausnitzii, Alistipes shahii, and *Lactococcus lactis* were among the taxa that participated in the reduction of metabolic pathways during ADD in urban samples.

3.5 DISCUSSION

In this study, we found distinctive taxonomic and metabolic features in non-ADD subjects across a rural-to-urban gradient in Northern Ecuador, most likely associated with local lifestyle factors (dietary habits, social status, economic development, antibiotic accessibility). We also attempted to provide insights into the value of these differences for the distinctive response or rural vs. urban microbiomes to ADD (Fig. B.9-B.10). A significant correlation between richness of functional and taxonomic profiles found in rural but not urban samples underlay, at least partly, these findings (Fig. B.4).

In particular, the intestinal microbiota from rural subjects showed a higher abundance of *Prevotella*, while *Bacteroides* and *Alistipes* presented a greater fraction of the total microbial community in the urban ones (Fig. 3.2B). Most *Prevotella* and *Bacteroides* MAGs recovered from the metagenomes and 16S rRNA gene sequences represented novel diversity since they could not be assigned to known bacterial species, suggesting that they may represent previously uncharacterized species (Fig. B.11). In the case of *Alistipes*, 16S rRNA gene amplicon sequences only identified one species (*A. indistinctus*) while taxonomic classification based MAGs or shotgun metagenomes identified 8 probably new species, allowing a higher resolution of the gut community for the metagenomes.

Interestingly, we also observed a differential response to diarrhea between rural and urban subjects. Specifically, the number of KEEG metabolic pathways and taxon abundance with significant shifts was higher in urban samples in comparison with the rural ones (Fig. B.9-B.10). However, the response to ADD seems to be variable among samples (large standard deviation) and the shifts to be taxon-specific as opposed to community-wide (Fig. 3.3 and B.9), which presumably reflected that specific samples were driving the differences in ADD. Therefore, a large cohort of samples is needed to further corroborate these preliminary findings and more firmly established that the diversity differences at non-

ADD state between rural and urban microbiomes play a significant role during ADD.

The differential response to diarrhea may be also associated with distinct extrinsic (causative agent of diarrhea, geographical factors) and intrinsic (microbial diversity patterns, functional capacities, community structure, host genetics) factors between the two groups. The causative agents of diarrhea include a broad spectrum of microbial, e.g., bacteria, viruses, and protozoa parasites, and non-infectious agents, e.g., food maldigestion, environmental exposures, endocrine diseases, among others (83). Consistent with these interpretations, pathogenic *E. coli* was identified as the probable etiological agent in five ADD samples from urban populations and two samples from rural ones (Fig. 3.4). Metagenomic comparison after removing the *E. coli* signal in ADD samples showed that the reduction in abundance in metabolic pathways involve in carbohydrate, vitamins, and amino acids metabolism were maintained; however, those affiliated with *E. coli* were not.

Analysis of OTU networks of the gut microbiota indicated that rural populations presented a higher number of OTUs and more connections among OTUs when compared to the urban one (Table B.4). A dense(r) network may be presumably associated with more connections among distinct bacterial species, modulating stability and community assembly in the gut during the response to perturbations. In this case, we observed that the urban network during ADD showed changes in more connections and nodes (e.g., loss of connections) than the rural one indicating possibly a more altered and unstable community with lower resilience to infection/diarrhea (Table B.4), which was also consistent with the taxonomic and gene function patterns revealed.

The observed differences in abundances of taxa between the two groups (locations) at the non-ADD (control) state might be attributable, at least in part, to multiple variables including local environmental conditions (temperature, elevation), socio-economic status, human contact with other communities and/or cities, and diet, in addition to unmeasured variables such as cultural factors and (human) genetic differences. Although dietary preferences were not recorded in this study, previous studies and literature have reported a diet based on fish, shrimps, plantain, rice, and coconut in the communities of San Lorenzo

(villages) (87). Villages are located closed to the coast and have access to three different rivers thus, freshwater and saltwater fish is one of the main components of the local diet. Subjects from the villages also use medicinal plants (88). On the other hand, a typical daily dish in Quito is composed by a high percentage of carbohydrates, proteins, and fats, and less than 15% of vegetables. Pork is one of the main ingredients in Ecuadorian Andean dishes (89). Dietary preferences most likely accounted, at least partly, for the differences observed between rural and urban non-ADD microbiomes (e.g., *Rikenellaceae* OTU associated with a higher fat diet in the urban population (11, 90). Moreover, it has previously been reported that housing density, road access, social connectedness, food-sharing, among others factors influence microbiome, including pathogen, prevalence and transmission in a community (25, 84).

The microbial community structure in the sampled subjects from the villages resembled those previously found in rural populations including Yanomami (Venezuela), Malawian (Amazon), Hadza (Tanzaia), and Matses (Peru) populations (4, 5, 8–10, 12, 67, 91). Moreover, Stagaman and collaborators (18) reported a similar taxonomic profile (abundance of *Prevotella* negatively correlated with house modernity while abundance of *Bacteroides* positively correlated) along a gradient of economic development in five villages close to the Cordillera de Cucutú in Southeastern Ecuador, indicating that there might be some universal patterns accompanying the phenomenon of urbanization in Ecuador and elsewhere. Nonetheless, despite the presence of western-type taxa in subjects from Quito, comparisons to the gut microbiota from US subjects revealed a clear segregation between these two locations (Fig. 3.1C), suggesting that still the gut microbiota of subjects living in Ecuador's Capital have not totally acquired a westernized microbial profile.

Collectively, our data revealed compositional differences across a rural-to-urban gradient in Northern Ecuador. Rural populations appeared to present smaller changes during ADD in comparison with urban ones driven by a (more) stable microbiome in terms of microbial composition, microbe-microbe interactions and stability, and functional diversity. However, further studies with larger and with ADD samples with known etiological agents are needed to further corroborate these results and conclusions on the

extent to which urbanization/lifestyle contribute on modulating the microbial response during ADD.

3.6 CONCLUSIONS AND PERSPECTIVES

Shotgun metagenomics had led to a remarkable growth of collective knowledge and information of the human microbiome. The reduction of sequencing cost had allowed to perform culture-independent analyses in research groups around the word and opened a new door to study microbial communities from human populations across the globe. This collection has allowed us to compare the diversity and composition of microbial communities from human populations with different lifestyles and geographical areas.

Given the complex host-microbiome-environment interplay, current research efforts have been focused on studying how ecological processes driven by local geographical factors influence the composition of the microbial communities. Understanding the interactions among microbial members at the community level requires the application of high-throughput technologies, both computational and laboratory, in order to characterize populations *in situ* and estimate/quantify their response to altered systems.

In this Chapter, we studied the role of geographical factors on the gut microbiome composition from a rural-to-urban gradient and compared the microbial response during diarrheal disease across the gradient. To the best of our knowledge, both, lifestyle and infectious diarrhea, have independently contributed to important insights in the gut microbial ecology, but to date have not been overlapped. However, the lack of metadata regarding dietary habits in the sampled populations limits our conclusions in any connection between diet and the metabolic and taxonomic potential of the assessed gut communities across the gradient. We also observed high inter-individual variation of the gut microbial communities among subjects from the same location. Therefore, extending this analysis to a larger sample size and with extensive metadata might help us to define the microbial signatures of each host population and their response to the urbanization phenomenon.

The investigation of human populations from multiple age groups, different locations (countries), and for longer periods, will be crucial to gain a more comprehensive view of the ecological and evolutionary processes *in situ* that are associated with community composition and patterns of diversity. The impact of a westernized behavior on the gut microbiota seems to have produced a global/worldwide convergent effect towards a reduction in diversity in the gut with traditional bacteria tending to extinct, an altered gut state (dysbiosis), and an increase of metabolic diseases.

Moreover, the elucidation of key microbial traits that compensate the impact/disturbance of the gut homeostasis and their associated mechanisms will allow us to identify treatment candidates or disease biomarkers to develop genomic approaches and establish ideal strategies for preventing microbiome-associated modern diseases.

3.7 ACKNOWLEDGEMENTS

This study was supported by the National Institute for Allergy and Infectious Diseases [grant number K01AI103544] at the US National Institutes of Health. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

3.8 REFERENCES

1. World Bank. 2018. World Bank. Urban population.
2. Broussard JL, Devkota S. 2016. The changing microbial landscape of Western society: Diet, dwellings and discordance. Molecular Metabolism 5:737–742.
3. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL. 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357:802–806.

4. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

5. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villarroel O, Foster M, Guija-Poma E, Troncoso-Corzo L, Warinner C, Ozga AT, Lewis CM. 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun 6.

6. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, Magris M, Hidalgo G, Contreras M, Noya-Alarcón Ó, Lander O, McDonald J, Cox M, Walter J, Oh PL, Ruiz JF, Rodriguez S, Shen N, Song SJ, Metcalf J, Knight R, Dantas G, Dominguez-Bello MG. 2015. The microbiome of uncontacted Amerindians. Sci Adv 1:1–12.

7. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489:220–230.

8. Tyakht AV, Kostryukova ES, Popenko AS, Belenikin MS, Pavlenko AV, Larin AK, Karpova IY, Selezneva OV, Semashko TA, Ospanova EA, Babenko VV, Maev IV, Cheremushkin SV, Kucheryavyy YA, Shcherbakov PL, Grinevich VB, Efimov OI, Sas EI, Abdulkhakov RA, Abdulkhakov SR, Lyalyukova EA, Livzan MA, Vlassov VV, Sagdeev RZ, Tsukanov VV, Osipenko MF, Kozlova IV, Tkachev AV, Sergienko VI, Alexeev DG, Govorun VM. 2013. Human gut microbiota community structures in urban and rural populations in Russia. Nat Commun 4:2469.

9. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, Fiori J, Gotti R, De Bellis G, Luiselli D, Brigidi P, Mabulla A, Marlowe F, Henry AG, Crittenden AN. 2014. Gut microbiome of the Hadza hunter-gatherers. Nature Communications 5:3654.

10. Martínez I, Stegen JC, Maldonado-Gómez MX, Eren AM, Siba PM, Greenhill AR, Walter J. 2015. The Gut Microbiota of Rural Papua New Guineans: Composition, Diversity Patterns, and Ecological Processes. Cell Reports 11:527–538.

11. De Filippo C, Di Paola M, Ramazzotti M, Albanese D, Pieraccini G, Banci E, Miglietta F, Cavalieri D, Lionetti P. 2017. Diet, Environments, and Gut Microbiota. A Preliminary Investigation in Children Living in Rural and Urban Burkina Faso and Italy. Front Microbiol 8.

12. Mancabelli L, Milani C, Lugli GA, Turroni F, Ferrario C, van Sinderen D, Ventura M. 2017. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. Environ Microbiol 19:1379–1390.

13. Fragiadakis GK, Smits SA, Sonnenburg ED, Van Treuren W, Reid G, Knight R, Manjurano A, Changalucha J, Dominguez-Bello MG, Leach J, Sonnenburg JL. 2018.

Links between environment, diet, and the hunter-gatherer microbiome. Gut Microbes 1–12.

14. Ruggles KV, Wang J, Volkova A, Contreras M, Noya-Alarcon O, Lander O, Caballero H, Dominguez-Bello MG. 2018. Changes in the Gut Microbiota of Urban Subjects during an Immersion in the Traditional Diet and Lifestyle of a Rainforest Village. mSphere 3.

15. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. 2016. Diet-induced extinctions in the gut microbiota compound over generations. Nature 529:212.

16. Winglee K, Howard AG, Sha W, Gharaibeh RZ, Liu J, Jin D, Fodor AA, Gordon-Larsen P. 2017. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. Microbiome 5.

17. Blaser MJ, Falkow S. 2009. What are the consequences of the disappearing human microbiota? Nat Rev Microbiol 7:887–894.

18. Stagaman K, Cepon-Robins TJ, Liebert MA, Gildner TE, Urlacher SS, Madimenos FC, Guillemin K, Snodgrass JJ, Sugiyama LS, Bohannan BJM. 2018. Market Integration Predicts Human Gut Microbiome Attributes across a Gradient of Economic Development. mSystems 3:e00122-17.

19. Kamada N, Chen GY, Inohara N, Núñez G. 2013. Control of Pathogens and Pathobionts by the Gut Microbiota. Nat Immunol 14:685–690.

20. Blumstein DT, Levy K, Mayer E, Harte J. 2014. Gastrointestinal Dysbiosis. Evol Med Public Health 2014:163–163.

21. McGrady-Steed J, Harris PM, Morin PJ. 1997. Biodiversity regulates ecosystem predictability. Nature 390:162–165.

22. Kennedy TA, Naeem S, Howe KM, Knops JMH, Tilman D, Reich P. 2002. Biodiversity as a barrier to ecological invasion. Nature 417:636–638.

23. Smith S, Montero L, Paez M, Ortega E, Hall E, Bohnert K, Sanchez X, Puebla E, Endara P, Cevallos W, Trueba G, Levy K. 2018. Locals Get Travelers' Diarrhea Too: Risk factors for diarrheal illness and pathogenic E. coli infection across an urban-rural gradient in Ecuador. Tropical Medicine & International Health In review.

24. Rival L. 2003. The meanings of forest governance in Esmeraldas, Ecuador. Oxford Development Studies 31:479–501.

25. Eisenberg JNS, Cevallos W, Ponce K, Levy K, Bates SJ, Scott JC, Hubbard A, Vieira N, Endara P, Espinel M, Trueba G, Riley LW, Trostle J. 2006. Environmental change

and infectious disease: How new roads affect the transmission of diarrheal pathogens in rural Ecuador. PNAS 103:19460–19465.

26. Zambrano-Barragán C, Zevallos O, Villacís M, Enríquez D. 2011. Quito's Climate Change Strategy: A Response to Climate Change in the Metropolitan District of Quito, Ecuador, p. 515–529. *In* Resilient Cities. Springer, Dordrecht.

27. Galvez H, Regalado J. 2007. Caracteristicas de las precipitaciones, la temperatura del aire y los vientos en la costa Ecuatoriana. Acta Oceanografica del Pacifico 14:201–205.

28. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl Environ Microbiol 79:5112–5120.

29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583.

30. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335–336.

31. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618.

32. Chao A, Shen T-J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10:429–443.

33. Hausser J, Strimmer K. 2009. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. J Mach Learn Res 10:1469–1484.

34. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16.

35. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2018. vegan: Community Ecology Package.

36. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13:R79.

37. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology 11:e1004226.

38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

39. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. 2012. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc 7:670–685.

40. Csardi C, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems:1695.

41. Rodriguez-R LM, Gunturu S, Harvey W, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. 2018. The Microbial Genomes Atlas (MiGA) webserver: Taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Research 46.

42. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. mSystems 3.

43. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17:132.

44. Goslee S, Urban D. 2007. The ecodist Package for Dissimilarity-based Analysis of Ecological Data. Journal of Statistical Software, Articles 22:1–19.

45. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12:902–903.

46. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8:e1002358.

47. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30:3123–3124.

48. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607.

49. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

50. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. e1900v1. PeerJ Preprints.

51. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

52. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

53. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

54. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191.

55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

57. The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 45:D331–D338.

58. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, Goldasich LD, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, Tripathi A, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, Consortium TAG, Knight R.

2018. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 3:e00031-18.

59. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60.

60. Gophna U, Konikoff T, Nielsen HB. 2016. Oscillospira and related bacteria – From metagenomic species to metabolic features. Environmental Microbiology 19:835–841.

61. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. 2013. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Curr Opin Biotechnol 24:160–168.

62. Lin R, Liu W, Piao M, Zhu H. 2017. A review of the relationship between the gut microbiota and amino acid metabolism. Amino Acids 49:2083–2090.

63. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102:2567–2572.

64. De Santis S, Cavalcanti E, Mastronardi M, Jirillo E, Chieppa M. 2015. Nutritional Keys for Intestinal Barrier Modulation. Front Immunol 6:612.

65. Kibe R, Kurihara S, Sakai Y, Suzuki H, Ooga T, Sawaki E, Muramatsu K, Nakamura A, Yamashita A, Kitada Y, Kakeyama M, Benno Y, Matsumoto M. 2014. Upregulation of colonic luminal polyamines produced by intestinal microbiota delays senescence in mice. Scientific Reports 4:4548.

66. Gulati K, Anand R, Ray A. 2016. Chapter 16 - Nutraceuticals as Adaptogens: Their Role in Health and Disease, p. 193–205. *In* Gupta, RC (ed.), Nutraceuticals. Academic Press, Boston.

67. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci USA 107:14691–14696.

68. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science 334:105–108.

69. Lim MY, Rho M, Song Y-M, Lee K, Sung J, Ko G. 2014. Stability of Gut Enterotypes in Korean Monozygotic Twins and Their Association with Biomarkers and Diet. Scientific Reports 4:7348.

70. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ.

2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

71. Antharam VC, Li EC, Ishmael A, Sharma A, Mai V, Rand KH, Wang GP. 2013. Intestinal dysbiosis and depletion of butyrogenic bacteria in *Clostridium difficile* infection and nosocomial diarrhea. J Clin Microbiol 51:2884–2892.

72. Chen S-Y, Tsai C-N, Lee Y-S, Lin C-Y, Huang K-Y, Chao H-C, Lai M-W, Chiu C-H. 2017. Intestinal microbiome in children with severe and complicated acute viral gastroenteritis. Scientific Reports 7:46130.

73. Pérez-Cobas AE, Artacho A, Ott SJ, Moya A, Gosalbes MJ, Latorre A. 2014. Structural and functional changes in the gut microbiota associated to *Clostridium difficile* infection. Front Microbiol 5:335.

74. The HC, Florez de Sessions P, Jie S, Pham Thanh D, Thompson CN, Nguyen Ngoc Minh C, Chu CW, Tran T-A, Thomson NR, Thwaites GE, Rabaa MA, Hibberd M, Baker S. 2018. Assessing gut microbiota perturbations during the early phase of infectious diarrhea in Vietnamese children. Gut Microbes 9:38–54.

75. Seekatz AM, Aas J, Gessert CE, Rubin TA, Saman DM, Bakken JS, Young VB. 2014. Recovery of the gut microbiome following fecal microbiota transplantation. MBio 5:e00893-00814.

76. Gomez DE, Arroyo LG, Costa MC, Viel L, Weese JS. 2017. Characterization of the Fecal Bacterial Microbiota of Healthy and Diarrheic Dairy Calves. J Vet Intern Med 31:928–939.

77. Jia H, Hanate M, Aw W, Itoh H, Saito K, Kobayashi S, Hachimura S, Fukuda S, Tomita M, Hasebe Y, Kato H. 2017. Eggshell membrane powder ameliorates intestinal inflammation by facilitating the restitution of epithelial injury and alleviating microbial dysbiosis. Scientific Reports 7:43993.

78. Raetz CRH, Whitfield C. 2002. Lipopolysaccharide endotoxins. Annu Rev Biochem 71:635–700.

79. Byers DM, Gong H. 2007. Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. Biochem Cell Biol 85:649–662.

80. Ma J-C, Wu Y-Q, Cao D, Zhang W-B, Wang H-H. 2017. Only Acyl Carrier Protein 1 (AcpP1) Functions in Pseudomonas aeruginosa Fatty Acid Synthesis. Frontiers in Microbiology 8:2186.

81. Choby JE, Skaar EP. 2016. Heme Synthesis and Acquisition in Bacterial Pathogens. J Mol Biol 428:3408–3428.

82. Steimle A, Autenrieth IB, Frick J-S. 2016. Structure and function: Lipid A modifications in commensals and pathogens. Int J Med Microbiol 306:290–301.

83. Hammer HF, Hammer J. 2012. Diarrhea Caused By Carbohydrate Malabsorption. Gastroenterology Clinics 41:611–627.

84. Bates SJ, Trostle J, Cevallos WT, Hubbard A, Eisenberg JNS. 2007. Relating Diarrheal Disease to Social Networks and the Geographic Configuration of Communities in Rural Ecuador. Am J Epidemiol 166:1088–1095.

85. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. 2014. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 15:382–392.

86. Minamoto Y, Otoni CC, Steelman SM, Büyükleblebici O, Steiner JM, Jergens AE, Suchodolski JS. 2015. Alteration of the fecal microbiota and serum metabolite profiles in dogs with idiopathic inflammatory bowel disease. Gut Microbes 6:33–47.

87. Gualotuna-Norona D. 2013. Investigación de la cocina ancestral ecuatoriana en la Comunidad Tolita y la Comunidad Cayapas de los cantones San Lorenzo y Eloy Alfaro de la provincia de Esmeraldas, y caracterización de un establecimiento gastronómico como medida de salvaguarda. Bachelor thesis, Universidad Technologica Equinoccial, Ecuador.

88. Trostle JA, Yépez-Montufar JA, Corozo-Angulo B, Rodríguez M. 2010. Diarrheal illnesses on the Ecuadorian coast: socio-environmental changes and health concepts. Cadernos de Saúde Pública 26:1334–1344.

89. Pazos-Carrillo S. 2010. Permanencias culturales y culinarias del Manual de Cocina de Juan Pablo Sanz en Quito (Ecuador): protocolos, cocina tradicional y formas de preparación. Master Thesis, Universidad Andina Simon Bolivar, Ecuador.

90. Daniel H, Gholami AM, Berry D, Desmarchelier C, Hahne H, Loh G, Mondot S, Lepage P, Rothballer M, Walker A, Böhm C, Wenning M, Wagner M, Blaut M, Schmitt-Kopplin P, Kuster B, Haller D, Clavel T. 2014. High-fat diet alters gut microbiota physiology in mice. ISME J 8:295–308.

91. Ayeni FA, Biagi E, Rampelli S, Fiori J, Soverini M, Audu HJ, Cristino S, Caporali L, Schnorr SL, Carelli V, Brigidi P, Candela M, Turroni S. 2018. Infant and Adult Gut Microbiome and Metabolome in Rural Bassa and Urban Settlers from Nigeria. Cell Reports 23:3056–3067.

CHAPTER 4. QUANTIFYING RECENT GENE EXCHANGE AMONG CLOSELY RELATED BACTERIAL GENOMES AND IMPLICATIONS FOR THE BACTERIAL SPECIES CONCEPT

4.1 ABSTRACT

High-throughput sequencing has revealed that bacterial genomes are highly dynamic, driven mostly by horizontal gene transfer (HGT). Quantifying HGT and its role in bacterial genome evolution and speciation has been challenging, especially between genomes of the same species, due to the high sequence identity of core genes at this level (e.g., low signal-to-noise ratio). Here, we devised a new approach to estimate the fraction of recent HGT among closely related genomes based on the frequency of identical genes (observed F_{100}) shared between two genomes relative to the number of such genes expected by chance according to their genome-aggregate average nucleotide identity (ANI) (expected F_{100}). Results from comparisons of hundreds of available genomes showed that our approach can reliably estimate the genomic fraction under recent exchange between closely related genomes (ANI 95.00% up to 99.97%). In particular, higher F_{100} than the average expected frequency denoted recombinogenic species as exemplified by ecologically versatile organisms, including opportunistic pathogens, while lower F_{100} values denoted clonal species as exemplified by obligate endosymbionts. Highly recombinogenic species showed non-random spatial and functional distribution of the recently exchanged genes across the genome indicating selection-driven HGT. Nonetheless, comparison of the effect of recombination and mutation on ANI for bacterial genomes close to the average expected (i.e., observed F_{100} = expected F_{100}) showed that recent exchanges were sufficient to counter the effect of random mutations, and thus, could lead to sexual speciation for the genome pairs analyzed.

4.2 INTRODUCTION

One important mechanism that accounts, at least in part, for the immense genetic diversity of bacteria is horizontal gene transfer (HGT). HGT can mediate the replacement of highly similar genetic segments at the nucleotide level through the process of homologous recombination (for integration of the horizontally transferred DNA into the genome of the recipient cell) or the transfer of non-shared DNA segments through illegitimate or non-homologous recombination (1, 2). It has even been argued that HGT might be so frequent and show no spatial biases across the genome (i.e., be random) that at least some bacterial lineages may be evolving sexually, similarly to several eukaryotes (3–5). However, the number of species and genomes analyzed to date remain limited, primarily due to the lack of high-throughput methods for robust HGT detection, while the effect of positive selection in driving the HGT events in these previous studies was not typically assessed (6). Further, most -if not all- methods employed to date for this purpose are based on assumptions that are frequently violated by the data analyzed such as that genes evolve under the molecular clock and lack of selection, limiting the broad applicability of the derived conclusions.

Several methods have been developed to identify genomic segments acquired through HGT. These methods include BLAST best-match analysis (7, 8), atypical G+C% or codon usage of exchanged genes compared to the average composition of the genome (9), networks of gene sharing (10, 11), and incongruent gene phylogeny in comparison with the species phylogeny (12, 13). While these methods are typically high-throughput, they have their own limitations. Phylogenetic approaches, while among the most robust for HGT detection, do not scale up well with an increasing number of sequences for analysis. As a result, alternative approaches including Bayesian methods have been proposed for the analysis of large genome datasets by offering robust estimations of the uncertainty in complex systems and high accuracy in comparison with traditional statistical tests (14). Bayesian statistics have been applied, among others, to detect evolutionary relationships among genomes within a phylogenetic framework (15, 16), gene transfer among bacteria (17, 18), and DNA rearrangements across the genome (19, 20). However, most of the available tools to detect recombination, including Bayesian methods, use the core genes

(i.e., genes shared by a group of genomes) as the input data (e.g., in order to build a robust phylogeny or training dataset), which does not take into consideration HGT events involving the accessory (variable) genes, and have assumptions that may (or may not) be violated by the data such as that genes evolve under the molecular clock. For instance, if accessory genes are not subjected to frequent intra-specific HGT but instead grow as a fraction of the total genome over evolutionary time, then sexual speciation will be more unlikely to occur.

With recent advances in high-throughput sequencing technologies, the systematic comparison of hundreds of bacterial genomes from different species in terms of their frequency of gene exchange becomes a highly interesting task. Quantifying genetic exchange across species, especially recent exchange events, is important not only for the bacterial species concept (e.g., sexual vs. asexual speciation) but also for understanding how adaptable different bacterial species are to the environment. For instance, quantifying recombination rates among vs. within sub-populations of a species might offer new insight into how these sub-populations may be responding differently to environmental fluctuations, leading to speciation (21–23). Moreover, recent gene transfers in pathogens could underlie rapid (new) host colonization, vaccine ineffectiveness, and resistance to antibiotics (24–26).

In this study, we introduced an alternative mathematical model to identify recent genetic exchange events present in both the accessory and the core genome in a pair of closely related bacterial genomes based on the genome-aggregate Average Nucleotide Identity (ANI) concept (27). ANI represents the average identity of all genes shared between any two genomes and has been shown to be a reliable measure of genetic relatedness that correlates tightly with DNA–DNA hybridization (DDH) experiments, i.e., the golden standard of prokaryotic taxonomy. In particular, two genomes showing higher than 95% ANI, which is equivalent to 70% DDH, could be assigned to the same species, assuming they also share the same key phenotype (27, 28); and this threshold is >97% of the times consistent with presently named species (29). Our model is based on the concept that at a given value of ANI, the frequency of identical nucleotide genes, F_{100}, follows a beta distribution (expected F_{100}) and that newly exchanged genes among genomes available

in our collection or their immediate ancestors will show an increased frequency of identical genes (observed F_{100}), which will represent outliers of the distribution. We used this concept to quantify differences in the fraction of recent HGT within and across species with different lifestyles (e.g., symbiotic vs. free-living) and ecological niches (e.g., fluctuating or more stable environments) and provide new insights into the species issue.

4.3 MATERIALS AND METHODS

4.3.1 Model overview

Let us consider a pair of genomes, descendants of the same ancestor and thus, members of the same species, which accumulate differences (e.g., point mutations, horizontal gene transfer events) over time at a variable (not constant) rate. Initially, the genomes will have the maximum level of sequence relatedness (100% ANI), and so, all of their genes will also show 100% nucleotide identity. Through time, the genomes will accumulate nucleotide mutations and thus, will show reduced relatedness and a decreased fraction of genes with 100% identity (Fig. C.1A). Nucleotide mutations are fixed at different rates due to differential selection pressures. Nonetheless, a decreasing fraction of genes will still show 100% identity (F_{100}) as divergence of genomes increases (within the same species). F_{100} is therefore a function of genome relatedness, measured in our model by ANI distance, defined as:

$$D = 100\% - \text{ANI}$$

If a gene undergoes homologous recombination between the two genomes in the pair or their immediate ancestors (i.e., not enough evolutionary time elapsed for the genes to have acquired additional mutations), the differences it may have accumulated will be reduced due to the introgression event (30) and the gene in the recipient genome will be identical to that of the donor genome. After the event, the gene is subjected to mutations. Multiple recombination events and/or non-homologous gene transfer events between two genomes or their immediate ancestors will result in an increased (observed) F_{100} relative to the expected fraction ($E[F_{100}]$). The difference between the expected and the observed

fraction is proportional to the rate of gene exchange between the genomes considered. The set of genomes with higher F_{100} than the $E[F_{100}]$ would be considered as recombinogenic while those with F_{100} less than the $E[F_{100}]$ value would be classified as clonal or low recombinogenic populations.

For each given value of *D*, F_{100} follows a probability density function given by the gene exchange rate. The expected value would be the sum of the values of F_{100} (at that value of *D*) multiplied by their corresponding probabilities such as:

$$E[F_{100}] = \sum F_{100} P(F_{100})$$

The probability distribution of F_{100} is however unknown (given that F_{100} is related with the gene exchange rate, which is unknown for most of bacterial species). Additionally, given the relationship between *D* and F_{100} (explained above and in Fig. C.1A), it is assumed that for every value of *D*, the distribution of F_{100} should follow a similar shape and can be described by the same type of parameters.

In a $\log_{10}$ space, *D* and F_{100} should follow a linear model. Thus, F_{100} is a linear combination of a series of vectors δ_m and unknown parameters θ_m:

$$E[F_{100}] = \theta_0\delta_0 + \theta_1\delta_1 \ldots \theta_m\delta_m$$

δ_m is a transformation of log(*D*). The predictor matrix (Δ) was obtained by a combination of transformations that best fit the model (without over-fitting) (Table C.1). Therefore, by estimating the regression parameters between these two variables we can calculate the expected value of F_{100} using the equation expressed in matrix notation:

$$E[F_{100}] = E[\theta]^T \Delta$$

Parameters were estimated using a Bayesian inference model. The model assumes that although for each value of *D* the distribution of F_{100} is unknown, in bacterial genomes from the same species (95%-100% ANI (28)) the sum of these distributions converges to a normal distribution following the central limit theorem. Because F_{100} varies between 0

and 1, we assumed that the conjugate non-informative prior of F_{100} (F_{100_0}) to be given by a beta distribution:

$$\mathrm{F100\_0} \sim \beta\left(\frac{1}{\sigma^2}, \frac{1}{\sigma^2}\right)$$

$$\mathrm{F100\_0} \propto \frac{1}{\sigma^2}$$

Where σ^2 is the variance of the whole dataset. The posterior distribution of the parameter θ is given by the product of the maximum likelihood estimation (MLE) of the regression parameters (based on θ and F_{100}) and the conditional probability of the values of F_{100} given their variance. The MLE corresponds to the joint probability distribution of F_{100} and the conditional probability of the parameter θ given the variance:

$$p\left(\theta \middle| \left(\frac{1}{\sigma^2}\right), F_{100}\right) = p\left(\left(\frac{1}{\sigma^2}\right) \middle| \theta, F_{100}\right) p(\sigma^2 | F_{100})$$

4.3.2 Parameter estimation based on empirical data

The training dataset to estimate the parameters of the above mentioned model consisted of 11,244 genomes belonging to 691 bacterial species, which were obtained from the Integrated Microbial Genomes (IMG) database (31). In order to capture most of the diversity, we sampled genomes with distinct lifestyles including symbiotic (i.e., host-associated) and free-living, ecologically versatile (i.e., marine and soil microbes with varied genome size) species. Due to overrepresentation of some lineages in the IMG database and in order to not bias the results by these few lineages (e.g., *Escherichia coli*, *Staphylococcus aureus*, and *Mycobacterium tuberculosis*), the genomes were first assigned to 95%-ANI groups, i.e., ANI among members of the group being >95% vs. <95% between groups, and three genomes were selected at most, at random, to represent each group. ANI values were calculated using the application of the enveomics script collection (32), essentially as previously described (28).

Reciprocal best matches (RBMs) between two genomes were identified by Blastn searches of their protein-coding genes using a cutoff of at least 70% nucleotide identity

and 70% coverage of the query length. The observed F_{100} was calculated as the number of RBMs with 100% nucleotide identity over the total number of RBMs in a genome pair. Pairwise genome comparisons from the same species (ANI values ranging from 95% to 100% (28)) were grouped every 0.2% ANI value brackets.

ANI values from all genome pairs were converted to distance (D) and expressed in a $\log_{10}$ scale ($\delta = \log_{10}(D)$). This dataset was subjected to a series of transformations (see also Results), and different combinations were used to build different predictor matrices (Δ). These matrices were ranked based on their Akaike Information Criterion (AIC) and the best Δ was then selected to perform all posterior prediction analyses (Table C.1). The best Δ was composed by the raw value of δ, a squared transformation (δ^2), and a cubic transformation (δ^3) plus a bias term, such as:

$$\Delta = [1, \delta, \delta^2, \delta^3]$$

In order to select the best predictor matrix, parameter estimation was carried out according to the Bayesian model (above). The posterior probability was calculated using Monte Carlo Markov Chain (MCMC) simulation. We ran a total of 1,000,000 simulations with a Burn-in of 300,000 and thinning of 1,000. With these simulations, we estimated the empirical values of $E[\theta]$ and established 95% credible intervals.

The values of Δ essentially represented the average gene transfer rate among all 691 bacterial species analyzed, and it was used to identify recombinogenic (larger values) and more clonal or less recombinogenic (lower values) genomes compared to the bacterial average $E[F_{100}]$. The mean parameter estimates and the 95% credible interval were as follows:

Parameter	Mean	Lower CI	Upper CI
θ_0	1.7958	1.6267	1.9640
θ_1	2.8698	2.6453	3.0955
θ_2	1.4218	1.3273	1.5157

θ_3	0.1907	0.19072	0.19073

The regression equation that describes F_{100} as a function of Δ is:

$$E[F_{100}] = 1.7958 + 2.8698\delta + 1.42180\delta^2 + 0.1907\delta^3$$

Based on the parameter values, we used this equation to estimate the expected F_{100} for any pair of genomes within a given (short) range of *D*. We selected pairs of genomes with ANI values ranging from 95% to 99.97%. These were mapped to the space of δ, ranging from –Inf to -1.3 [$\delta = \log_{10}(100\%\text{-ANI})$; Fig. 1B]. Pairs with δ higher than -1.3 (below 95% ANI) present low values of F_{100} and the signature of recent genetic exchange events is diluted within the effect of whole genomic diversification for our approach, which targets the within-species level. Conversely, when the value of δ is less than -3.65 (above 99.97% ANI), F_{100} tends to 1 and genetic exchange events cannot be reliably estimated based on sequence identity alone. Therefore, the estimated parameters have predictive power in the 95% to 99.97% ANI range (note: using the model outside this range can introduce errors that would not represent the behavior of the recombination process).

4.3.3 Detection of candidate genes under recent exchange

The above model provided gross estimates of the total number of genes recently exchanged between any two genomes. In order to identify the specific genes that underwent recent exchange within a genome pair (showing 100% nucleotide identity) and distinguish them from genes showing 100% nucleotide identity due to high sequence conservation (strong negative selection constraints), the following procedure was followed:

1. ANI values from pairwise comparisons of genomes assigned to the same species (95% ANI cluster) were used to produce an ANI matrix, which was subsequently subjected to the Partition Around Medoids (PAM) clustering algorithm in order to identify sub-clusters of genomes. The Silhouette algorithm was used to select the number of sub-clusters (32).

2. Genome pairs with ANI values similar to the genome pair query (ANI ± 0.2) were selected from as many sub-clusters as were available within the same species as the query pair (typically 100s of pairs analyzed for each case).
3. Genomes belonging to the same sub-cluster than the query pair were excluded from further analysis in order to not bias the result if the HGT event occurred at the immediate ancestor of the sub-cluster (and thus, all members of the sub-cluster possess the transferred gene).
4. Genes with 100% nucleotide identity present in the variable and core genome were identified. Core genes were defined as clusters of orthologous groups (COGs) present in 90% or more of the genomes of the species analyzed. COGs were defined by Markov Clustering (MCL) on the sets of RBMs for all pairs of genomes using the script ogs.mcl.rb (33).
5. For each gene present in the core genome, the frequency of its nucleotide identity greater or equal to 99.8% was calculated among its orthologs in each COG.
6. Genes with such high identities in 80% of more of the pairs analyzed were considered highly conserved genes in terms of sequence identity and were excluded from the list of potential transferred candidates. Genes encoding for proteins with less than 50 amino acids were also excluded to avoid truncated proteins or artifacts related to the short protein sequence. The remaining core genes as well as the non-core genes from step #4 were considered candidates of HGT.
7. Candidate transferred genes were annotated based on BLASTp searches against proteins sequence from complete genomes from the Reference Sequence (RefSeq) database at NCBI.

4.3.4 Estimation of the effect of recent mutations and recombination on ANI

In this study, recent gene exchanges were targeted, which represent genes that have not acquired any mutations since the recombination event (i.e., they still show 100% nucleotide identity). Thus, the recombination events that were identified were as old as, at maximum, about the time required to obtain 1 mutation/gene.

Sexual maintenance of a species requires that intra-species recombination affects the entire genome, not just a few foci, otherwise the non-affected loci will continue to diverge in sequence. Further, the uniformly distributed gene exchanges must occur at high-enough frequency to counteract the effect of point mutation, even if some non-uniform gene exchanges could occur at higher frequency, on top of the uniform ones, and are concentrated to one or a few locations of the genome. Thus, the distance between recombined genes (100% nucleotide identity in our case) should be inversely proportional to the frequency of uniformly exchanged genes, i.e., the more frequent the gene exchanges are the less the distance between exchanged genes will be, and can be modeled to provide insights into the frequency of recombination. Specifically, the distance between the most spatially distant recombined genes in the genome (percentile 99^{th}) will represent the lower bound of (recent) recombination frequency. Further, if dense clusters (i.e., short distance among recombined genes) dominate the distribution of distances among recombined genes, because -for instance- several genes are exchanged together as part of whole operons in single HGT events, there should be an over-inflation at or near zero in the distribution of distance values. Therefore, the distribution of distances among recombined genes outlined above was also modeled by removing distances of zero initially (i.e., merging contiguous recently exchanged genes), and then instances of ≤ 1, etc. More generally, we masked distances $\leq k$, where k is a parameter. The fitdristrplus package (34) in R was used to identify the parameters of a lognormal distribution with best fit to the distribution of the distance values, and the smallest k with a qualitative change in the shape of the distribution (k=3), was selected.

The expected number of exchanged genes (E) across the genome (i.e., excluding clusters) was calculated as the total number of genes in the genome (N) divided by the mean of the lognormal distribution above:

$$\mathrm{E}[\mathrm{Rec.\,genes}] = N / \exp(\mu + \delta^2/2)$$

Next, the effect of recombination on the ANI of a genome pair relative to point mutation was estimated based on the change in the ANI value when the number of genes calculates from the equation above (E [Rec. genes]) were allowed to recombine between the two

genomes (i.e., become 100% identical at the nucleotide level), *in-silico* (by introducing nucleotide changes in orthologous genes), for evolutionary time that was, at maximum, equal to the time required to acquire 1 mutation per gene (in order for the recombined genes to remain 100% identical and not have enough time to accumulate any mutations; see also above). Accordingly, the effect of recombination within this time was calculated by subtracting the initial ANI and the new ANI, and multiplying that value by E [Rec. genes]:

$$\text{Rec. effect} = (ANInew - ANI) * E[Rec. genes]$$

Analogously, in order to estimate the maximum effect of recently introduced mutations on ANI of a genome pair within the same evolutionary time interval (i.e., time to acquire up to 1 mutation/gene), the following expression was used:

$$\text{Mut. effect} = 1 * N/Genome\ size$$

The effect of recombination and mutation on ANI was subsequently calculated as the subtraction of the two estimated values, which means that positive values result in a greater effect of recombination, hence sexual maintenance with tendency to increase ANI over time. In contrast, negative values reflect a greater effect of mutation, hence a tendency to decrease ANI over time.

The effect of recombination and mutation on ANI was subsequently estimated under four different, increasingly more conservative scenarios for the effect of recombination:

1. Including all the possible genes under recent exchange.
2. Collapsing clusters of gene exchange to be represented by a single gene/event.
3. Correcting by functional bias that may reflect strong selection bias for exchanging specific functions: functional annotation of exchanged genes was performed and the resulting distribution was compared to that of all the genes in the genome. Categories that were enriched in exchanged genes were corrected by manual inspection of the gene annotation in the category and removed from consecutive analysis.

4. Using the lower bound of (recent) recombination represented as the distance between the most spatially distant recombined genes in the genome (percentile 99th).

A subset of genome pairs that showed F_{100} values equal or close to $E[F_{100}]$ was selected for the analysis described above in order to represent genomes that were not outliers (Fig. C.1) and thus, were presumably under no extreme selection pressure for gene exchange. From the distribution plot on Fig. C.1, the value of the lower credible interval was used as a proxy of the fraction that corresponded to highly conserved 100% identity genes and that number was subtracted from F_{100} to provide the new F_{100} that, most likely, represented only recently exchanged 100% identity genes. Genes at 100% nucleotide identity that most likely represent highly conserved genes were identified and excluded using the approach mentioned above. The remaining (most likely recently exchanged) genes were mapped to the genome sequence, and the distance between exchanged genes and their spatial distribution across the genome were assessed as described above.

4.4 **RESULTS**

4.4.1 Application of the model to species with different ecologies

Bacterial genomes from the IMG database assigned to species with different lifestyles including obligate intracellular symbionts, host-associated, and free-living species were selected for further analysis (Table C.2). For each genome pair within a species, the expected F_{100} was calculated based on all genome pairs from all species in the database showing the same ANI bracket, i.e., ANI +/- 0.1 The observed and expected F_{100} values were combined in a single measurement (sigma) that represents the number of standard deviations that the observed F_{100} differs from the expected F_{100} for genome pairs with same ANI value (Figs. 4.1A and C.2A).

Notably, we observed two distinct clusters based on the sigma values (or amount of genetic exchange; observed F_{100}). One cluster included bacterial species associated with restricted habitats (Fig. 4.1A, lower right part of the plot). In this cluster, the observed that

the fraction of identical genes is lower than the expected average (negative sigma value), meaning that fewer recent exchanges were predicted to have occurred compared to the bacterial average. The cluster included *Buchnera aphidicola,* an intracellular symbiont that showed the lowest exchanged fraction followed by *Yersinia pestis, Mycobacterium tuberculosis,* and *Rickettsia rickettsia*, which represent obligate pathogens of humans.

Conversely, the other cluster corresponded to ecologically versatile species, exhibiting higher rates of recent genetic exchange (Fig. 4.1A upper part, above the dashed line; positive sigma values). In this cluster, the observed fraction of identical genes was greater than expected, indicating that multiple genetic exchange events have recently occurred between the genomes compared. These events have also been introduced very recently, i.e., in a time frame shorter than that required for point mutations to occur in the same time (otherwise the genes would not have been 100% identical to be included in the F_{100} metric). This cluster included commensal and environmental bacteria, including -but not limited to- opportunist pathogens such as *Campylobacter coli* (35)*, Enterococcus faecalis* (36), and *Neisseria meningitidis* (37).

Our model also provided quantitative estimates of the differences in genetic exchanges among bacterial species. In the cluster of recombinogenic bacteria, the fraction of genetic exchange was, in general, higher in opportunistic pathogens such as *Campylobacter jejuni* (average sigma: 2.9) and *Campylobacter coli* (average sigma: 1.8) than environmental organisms associated with terrestrial or marine environments such as *Alteromonas* (average sigma: 0.8) and *Synechococcus* (average sigma: 0.3)), respectively (Fig. C.2B). Interestingly, the average of genes detected to be recently exchanged (139) was similar among opportunistic pathogens (e.g., *C. coli*, *N, meningitidis*, and *E. faecalis*) with the exception of *Vibrio cholerae* (lower, at 56 genes) and *C. jejuni* (higher, at 254 genes). The elevated number of exchanged genes in *C. jejuni* might be related to its diverse ecology and ability to colonize multiple host species. For instance, *C. jejuni* is considered commensal in chicken, but pathogenic when colonizing the intestinal tract of mammals (38), and it can also be found in water sources through contamination with feces, (39). This ecological versatility could represent varied environmental selections pressures, including

antibiotic treatment, and thus, a higher demand for adaptation through genetic exchange than other, less ecologically versatile species.

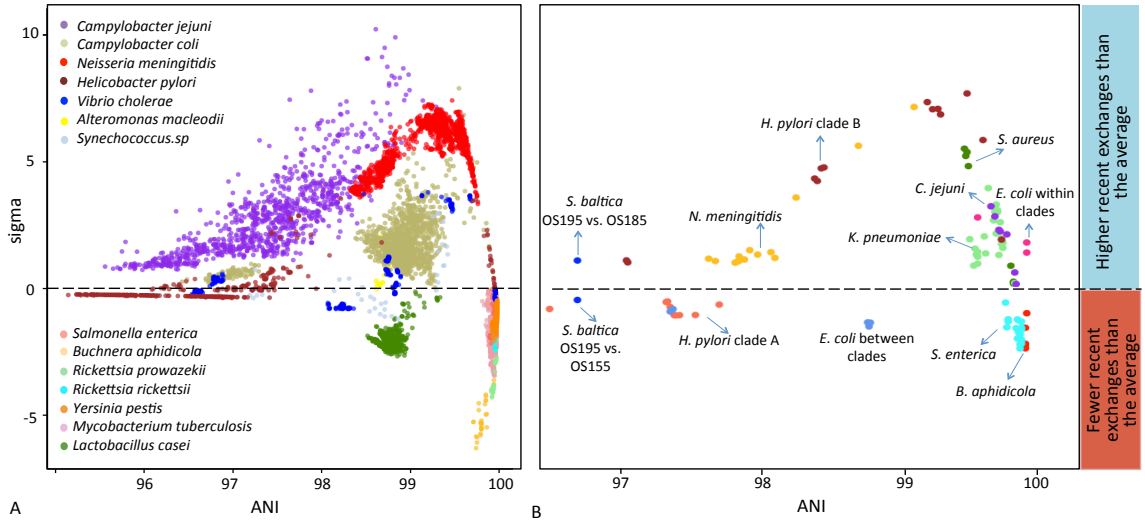


Figure 4.1. Variation in the fraction of recent gene exchange among bacterial species.

Sigma (y-axis) represents the number of standard deviations that the observed F_{100} differs from the expected F_{100} for genome pairs with same ANI value (x-axis). Data points represent genome pairs colored by their species assigned to, using 95% ANI as the threshold for species. Panel A shows all genome pairs of selected species with different lifestyles from the IMG database and panel B shows selected genome pairs form panel A (for details, see table C.2).

4.4.2 Quantifying recent genetic exchange within bacterial species

To further test the predictions of our model, we analyzed genomic data from previous studies (Table C.2, Fig. 4.1B). The *Helicobacter pylori* genome dataset previously published (40) was analyzed to assess recent genetic exchange among strains isolated from the same and different human individuals. In agreement with previous results, our model indicated that most of the strains living in individuals from the same family presented

positive sigma values and a high number of exchanged genes (up to 400 genes of the total genes in the genome). In contrast, *H. pylori* strains isolated from different families exhibited negative sigma values indicating null or low recent exchange (Table C.3). In addition, the estimated fraction of the genome under recent exchange varied among co-occurring *H. pylori* strains from the same individual. For instance, 316 exchanged genes were estimated in the genome pair SA161A-SA161C (approximately, 19% of the total genes in the genome), 198 genes for SA227A-SA227C (11% of the total), 77 for SA163C-SA163A (5% of the total). Cases without recent gene exchange signatures were also detected, e.g., SA160A-SA160C and SA300A-SA300C. Results consistent with previously published reports (41) were also obtained for *C. jejuni* genomes belonging to specialist and generalist lineages as well as originating from different hosts (Fig. C.4), further corroborating that ecological differentiation, driven by distinct host niche preferences, might restrict gene flow among closely related genotypes.

Four *Shewanella baltica* genomes from different depths of the Baltic Sea were previously studied. At least two of them were shown to have undergone extensive recent exchange, based on manual inspection of whole genome alignments, that could lead to sexual speciation (5). Consistent with the results of the original study, our model revealed positive sigma values equal to 1.11 and an estimated number of exchanged genes of 242 among strains OS195 and OS185, the two most recombinogenic genotypes previously identified (Fig. C.3A), and negative sigma values (-0.018) among the remaining genome pairs, which also represented the most different depths sampled (e.g., strains OS195-OS155). Manual inspection of the 100% nucleotide identity genes in the genome pair OS195-OS185 identified as exchanges by Caro-Quintero and colleagues showed that our model identified mostly (160/163) the same genes under recent exchange, with enrichment in proteins belonging to metabolism and mobile functional categories. Conversely, for the genome pair OS195-OS155, which did not show elevated gene transfer, the 100% nucleotide identity genes were enriched in functions associated with essential cellular processes and cell viability (Fig. C.3B). The latter genes are more likely to show 100% identity due to their high sequence conservation (e.g., greater selective constraints) rather than recent gene exchange (see also next section).

We also ran our model on a set of *C. jejuni* genomes previously reported (41) in order to compare the signatures of gene flow in isolates belonging to specialist and generalist lineages as well as originating from different hosts (Fig. C.4). Overall, higher rates of recent exchange were detected in the generalist clonal complex ST-45 presenting sigma values greater than 1 and an elevated estimated number of exchanges genes (n=366) in comparison with the other generalist lineage, ST-21. Genome pairs from the same clonal complex, either generalist (e.g., ST-21) or specialist (e.g., ST-61, a cattle specialist), showed higher rates compared to pairs belonging to distinct clonal complexes. This is presumably attributable, at least in part, to the fact that strains that colonize the same host or source have more opportunities for cell-cell contact, and are under the same environmental conditions and selection pressures compared to strains living in different hosts, in addition to their typically higher relatedness at the sequence level and mechanisms related to gene transfer and recombination, which could facilitate genetic exchange. On the other hand, pairwise comparisons of genomes from different clonal complexes including ST-353 (host source: chicken) vs. ST-42 (host source: cattle) and ST-353 vs. ST-61 (host source: cattle) showed lower exchange values than the bacterial average. Similarly, comparisons between genomes from generalist vs. specialist linages, e.g., ST-45 vs. ST-353 and ST-45 vs. ST-42, presented sigma values below one. These results confirm previous hypothesis that ecological differentiation, driven by distinct host niche preferences, might restrict gene flow among closely related genotypes.

4.4.3 Candidate genes that undergo recent exchange

Among the 100% identity genes shared between a pair of genomes, the ones that were likely under recent exchange, as opposed to showing high identity due to high sequence conservation (evolutionary constraints), were identified based on their sequence identity patterns among genomes of the same species as the pair but from different sub-species clades than the clades represented by the genomes of the pair. Briefly, highly conserved genes were identified as those showing 99.8-100% nucleotide identity in 80% or more randomly drawn genomes from different sub-clades of the species and were removed from the list of recently exchanged genes. Functional annotation indicated that

essential genes including ribosomal structural genes, RNA operons, and DNA/RNA polymerases were overrepresented among the removed genes, as expected.

To confirm recent exchange signatures in the identified candidate genes, we randomly selected a subset of genome pairs and performed a phylogenetic assessment by comparing the gene tree topology with that obtained using the ANI distance matrix of the two query genomes of the pair and their close relatives (genomes from the same species). Examples of topological incongruence between the two trees due to gene exchange events are shown in Figures C.5 and C.6. Manual inspection of tree topologies indicated that most of the transferred genes, e.g., 97.2% of genes tested in a *C. jejuni* genome pair and 76% in a *N. meningitidis* pair, clustered in the same sub-clade and presented shorter branch lengths in comparison with the ANI tree, which was consistent with recent exchange of the genes. The remaining genes corresponded to cases where the phylogeny was not conclusively resolved because of the high nucleotide sequence identity among genes or might represent false positive calls by our approach.

In addition, candidate exchanged genes among genome pairs from *N. meningintidis* and *C. jejuni* were functionally annotated. Our model identified exchanged genes in both the core and variable genes and their frequency in these two gene sets was variable. For instance, the *C. jejuni* pair 30318-LMG_9879 presented 84% of its exchanged genes in the core while pair 63-117 showed 68% of its exchanges in the variable genes. We found also differences between the predicted functions of exchanged genes located in the core and variable genome (Fisher's exact test, $P < 0.05$). Particularly, the majority of exchanged genes present in the variable genome were associated with unknown and metabolic processes, up to 50% and 80% respectively, as expected since the variable gene set is typically enriched in accessory genes compared to the core (42, 43).

For each pair of genomes, the functional composition of putatively exchanged genes was compared with that of the genes in the genome excluding highly conserved genes (100% nucleotide identity) from the comparison (Fig. C.7). Overall, we found differences in the frequency of the functional categories between these two groups (Hypergeometric test, $P < 0.05$). Most of the exchanged genes were involved in metabolic

processes, but also a high percentage was of unknown function (ranging between 40% and 60% of the total). A small fraction was assigned to the cell motility category including flagellar proteins (max 5% in *C. jejuni* pairs), and to mobile elements including phage-related sequences, integrative elements, and transposases (max 3% in *N. meningintidis* pairs).

Specifically, exchanged genes in *C. jejuni* were enriched in functions associated with interaction with host cells and the environment such as efflux proteins, ABC transport systems, flagellar, secreted, and membrane proteins. Antibiotic resistant factors, including β-lactamase OXA-61, and the efflux pumps SMR and CmeABC (detected in the query pairs 63-117 and 30318-LMG.9879, respectively), were also enriched in the recently exchanged gene pool. In the case of *N. meningitidis*, we found genes encoding transcriptional regulators as well as membrane and secretion proteins that allow efficient interaction with host cells. However, most of the exchanged genes were annotated as hypothetical proteins (Fig. C.7). This is not surprising since a high percentage of *N. meningitidis* variable genes linked to gene transfer events have been poorly characterized (44, 45). A second set of genes was comprised by sequences related to DNA rearrangements, insertion sequences, and putative phage genes including IS1106, IS360, and DDE transposases as well as phage tail proteins. These genes may influence the activation or inactivation of virulence genes and also contribute to variation in the envelope structure (44, 46). Among others, ABC transporters, hemagglutinin, and MafB proteins also presented signals of recent exchange but were not enriched in comparison to their frequency in the genome. These findings suggested that genetic exchange was not random across the genome but driven by selection for specific functions, which we evaluated in more detail.

4.4.4 Spatial biases of recently exchanged genes across the genome

To evaluate whether the spatial distribution of exchanged genes across the genome of highly recombinogenic bacteria was random, multiple non-parametric tests were applied including Moran's (47), Cramer-von Mises (48), Watson (49), and Kolmogorov-Smirnov (50) tests. This battery of tests was used to account for different deviations from the

uniformity of the locations of the exchanged genes. Overall, in most of the cases the distribution of exchanged genes in the tested genomes deviated from random (Table C.4) and clusters of exchanged genes were observed even when highly conserved genes with 100% nucleotide identity, which represent blind spots of our approach, were removed (Figs 4.2 and C.8).

For instance, 33% of recent exchanges were located in two consecutive regions in the *C. jejuni* pair 110-117 and more than half (58%) were concentrated in three consecutive regions in the *C. jejuni* pair 63-117, which were among the most recombinogenic pairs found by our analysis with large numbers of genes exchanged. This spatial clustering was consistent with the functional bias, mentioned above, that revealed exchanged genes to be significantly enriched in metabolic, mobile, and unknown functions (Hypergeometric test, $P < 0.05$) (Fig. C.7).

Similarly, *N. menintigitis* genome pairs showed exchanged genes in specific, non-random genomic regions, including pathogenicity islands (Fig. C.8). However, the genome pair 2531839670-2537562110, which presented a relatively small number of recent imports (n=47) did not present gene clustering across the genome (Table C.4). Further inspection revealed an enrichment of transposases and phage proteins flanking the exchanged genes, indicating that, most likely, the recent HGT events in this pair are being driven by mobile elements with no strong preference for integrating at specific sites of the genome (Fig. C.8). Moreover, 80% of its exchanged genes were annotated as hypothetical proteins, revealing a strong functional preference in the genes carried by the mobile elements (Fig. C.7).

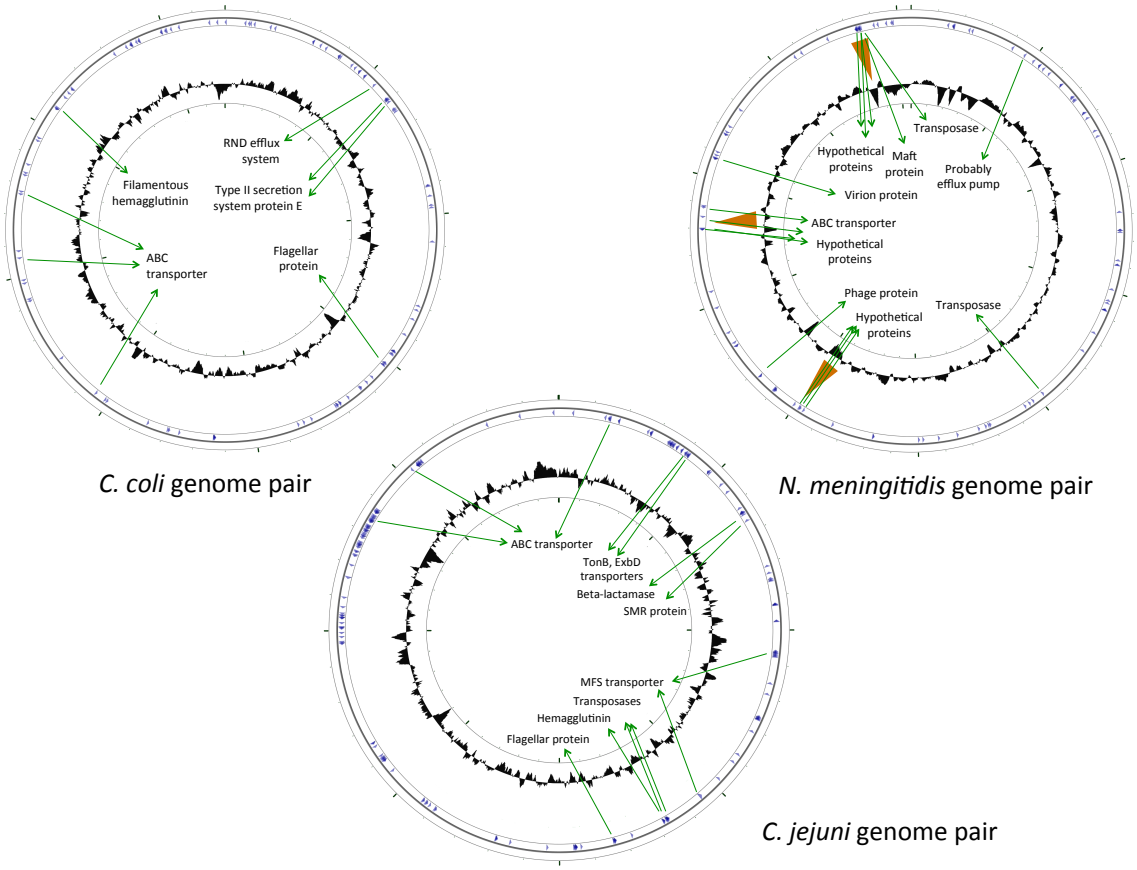


Figure 4.2. Spatial distribution of exchanged genes across the genome.

Circular plots of query genome pairs with genes identified as recently exchanged denoted by blue arrowheads. Conserved regions (genes with 100% nucleotide identity without signatures of gene exchange) were removed from the genome sequence before assessing the spatial distribution of recent exchanges. The location of genes of interest are highlighted with green arrows, and pathogenicity islands are represented by orange triangles, identified using PAIDB v2 (51) and IslandViewer (52). Plots were drawn using CGview (53)

4.4.5 Relative importance of recombination to mutation indicates sexual speciation

Since the recently exchanged genes were mostly found in clusters, likely indicating selection-driven recombination for their corresponding functions, we were interested in evaluating next whether the identified recent exchanged genes in a genome pair were sufficient to counteract the effect of random mutations (sexual speciation), when the effect

of selection is removed. For this, the effect of recombination and mutation on ANI was evaluated in a subset of genome pairs from different bacterial species that presented F_{100} values equal or close to $E[F_{100}]$, in order to avoid extreme cases of selection for the transferred genes.

The analysis showed that, after correcting for functional enrichment and/or spatial bias of the exchanged genes driven presumably by selection for the corresponding gene functions, the effect of recombination is larger than the mutation one on ANI for most of the analyzed genome pairs (Table 4.1). For instance, 101 recently exchanged genes were estimated in the pair *N. meningitids* 2523533512-2534681689 and when clusters of these genes where collapsed in one event (Fig C.9), the remaining set of exchanged genes (n=81) were distributed randomly across the genome and appeared to be sufficient to compensate the effect of mutation. Moreover, unknown functions were enriched in exchanged genes when compared to the total genes in the genome. After excluding unknown function genes, which were typically located transposases and integrases, the remaining set of recent exchanges (n=75) showed a greater effect on ANI in comparison with the mutation. Similar results were observed for the genome pairs *C. jejuni* 111-64, *N. meningitids* 2523533512-2534681689, *N. meningitids* 2523533512- 2537562111, and *S. pneumoniae* 2519899810-2528311139. In the case of the genome pairs *C. jejuni* 172-210 and *C. coli* 2516143039-2516143061, the effect of mutation on ANI was larger than recombination when the clusters of exchanged genes were removed but not greatly different from zero (e.g., no large difference between the effects of recombination and mutation) (-0.0001 and -0.0003, respectively). For an even more conservative estimation of the effect of recombination, we examined the distance in the genome between the two most distant exchanged genes across the genome (percentile 99^{th}) as a proxy for recombination frequency (see Materials and Methods for details). The effect of recombination based on this lower bound was lower than the mutation one in all genome pairs examined but not greatly different from zero, i.e., in the 0.0008-0.0005 ANI range.

Table 4.1 Effect of recombination and mutation on ANI. The column named "Actual ANI" (2nd column) corresponds to the initial ANI value of the genome pair (1st column), "Genes sexual" is the number of estimated recently exchanged genes after correcting for clusters of exchanged genes (i.e., clustered genes are represented by only one gene), "New ANI" refers to the new ANI value after increasing the nucleotide identity of one RBM to 100%, "1 Rec. ANI" to the subtraction of the two values of ANI, i.e., New ANI vs. Actual ANI, "Rec. effect" to the estimated effect of recombination on ANI, i.e., number of Genes sexual multiplied by 1 Rec. ANI, "Mut. effect" to the effect of one mutation event on ANI in the same evolutionary time period, and "rec-mut" to the subtraction of the recombination and mutation effects.

Genome pair	Actual ANI	No of genes	Genes sexual	New ANI	1 rec. ANI	Rec effect	Mut effect	rec-mut
***C. jejuni* 111-64**								
case1	0.97849	1695	190	0.97851	2E-05	0.0038	0.00103	0.0028
case2	0.97849	1695	63.39	0.97851	2E-05	0.0013	0.00103	0.0002
case3	0.97849	1695	63.39	0.97851	2E-05	0.0013	0.00103	0.0002
case 4	0.97849	1695	13.22	0.97851	2E-05	0.0003	0.00103	-0.0008
***C. jejuni* 172-210**								
case1	0.98370	1703	190	0.98372	1.2E-05	0.0023	0.001042	0.0012
case2	0.98370	1703	80.72	0.98372	1.2E-05	0.0010	0.001042	-0.0001
case 4	0.98370	1703	19.5	0.98372	1.2E-05	0.0002	0.00104192	-0.0008
***C. coli* 2516143039-2516143061**								
case1	0.96726	1790	87	0.96727	1.5E-05	0.0013	0.001064	0.0002
case2	0.96726	1790	48.06	0.96727	1.5E-05	0.0007	0.001064	-0.0003
case 4	0.96726	1790	11.89	0.96727	1.5E-05	0.0002	0.001064	-0.0009
***N. meningitidis* 252353351-2537562111**								
case1	0.97355	1990	94	0.97357	2.2E-05	0.0021	0.001003	0.0011
case2	0.97355	1990	80	0.97357	2.2E-05	0.0012	0.001003	0.0002
case3	0.97355	1990	55	0.97357	2.2E-05	0.0018	0.001003	0.0008
case 4	0.97355	1990	16.7	0.97357	2.2E-05	0.0004	0.001003	-0.0006

Table 4.1. continued

***N. meningitidis* 252353351-2534681689**								
case1	0.97364	1989	101	0.97366	1.4E-05	0.0014	0.001002	0.0004
case2	0.97364	1989	81	0.97366	1.4E-05	0.0011	0.001002	0.0001
case3	0.97364	1989	75	0.97366	1.4E-05	0.0011	0.001002	0.0001
case 4	0.97364	1989	22.1	0.97366	1.4E-05	0.0003	0.001002	-0.0007
***S. pneumoniae* 251989981-2528311139**								
case1	0.98942	2207	365	0.98942	0.000009	0.0032	0.00105	0.0022
case2	0.98942	2207	174.5	0.98942	0.000009	0.0015	0.00105	0.0005
case 4	0.98942	2207	61.6	0.98942	9E-06	0.0005	0.00105	-0.0005

4.4.6 Comparison to other high-throughput HGT detection methods

Exchanged genes identified by our approach were also tested using a newly described recombination detection tool, fastGear (54). This tool identifies ancestral recombination events present in all strains of a lineage and recent events in a subset of strains within the same lineage based on multiple-sequence alignments. We analyzed a set of 64 *C. jejuni* genomes previously published (41). Clusters of orthologous genes (COGs) were defined for this analysis using the script ogs.mcl.rb (33), lineages were identified using the BASP software (55), and fastGear was executed with default parameters.

Comparative analysis of the *C. jejuni* genome pair 110-117 (Fig. C.10) showed that 65% of the exchanged genes identified by our tool also presented recombination signals by fastGear. The remaining set of genes (35%) included strain-specific (non-core) genes in 20% of the genomes and/or genes with few informative sites, which can affect the detection ability of fastGear. In addition, fastGear identified recombination events in 60% of the total COGs, which showed no signatures of recent exchange by our tool. These genes were not detected by our method because they were not 100% nucleotide identical (fastGear

identifies both recent and older events that might have occurred in previous generations in the lineages of the genomes compared).

Similar results were obtained with other bacterial genome pairs, e.g., the *C. jejuni* genome pair 40-117 in which 50% of the exchanged genes identified by our approach were also identified by fastGear. Tree topology analysis of the recently exchanged candidates by our approach showed that 82% of them (225/275) presented a different branching pattern in comparison with the genome tree, indicating gene transfer. The remaining 18% corresponded to genomic regions with few informative sites for robust assessment, and so, some of these might represent false positive calls by our approach. In addition, among the highly conserved COGs detected by our approach (groups of orthologs with frequency values >80%), 92% (180/196) presented no signal of recombination by fastGear. In the remaining 16 COGs, 12 genes presented only one ancestral exchange event and the rest between two and five events by fastGear.

In summary, the two approaches largely agreed on exchanged genes with 100% identity, with our approach uniquely identifying non-core exchanged genes, and their differences in the remaining genes being attributable to the definition and assumptions for calling a gene transfer event used by each method. Thus, the application of both methods appears to be complementary in the identification of recombination signals in a group of genomes.

4.5 **DISCUSSION**

Quantifying recent gene flow between and within bacterial species can provide new insights into how populations are responding to the environment and how flexible their genomes are to acquire gene functions from their closely related genotypes. In this study, we quantified and compared the rates of recent gene exchange within bacterial species with similar or different lifestyles, and efficiently identified the transferred genes and functions. For instance, comparisons among free-living bacteria, which are able to colonize multiple hosts and changing environments, indicated that the fraction of recently exchanged genes

compared to the average expected varied from ~14%, on average, of the total genes in the genome in *C. jejuni* to ~7.7% in *C. coli*, ~6% in *N. meningitidis*, ~3% in *Alteromonas* (2.95%), and ~1% in *Synechococcus* (Fig. C.2B). The elevated fraction of gene exchange estimated by our approach in populations of the opportunistic pathogens mentioned above is presumably associated with the high genetic diversity observed within these species as well as their ability to colonize multiple environments and selection pressures.

Close examination of the exchanged genes showed that, overall, recent HGT events appeared to target specific functions associated with adaptation and survival to the local environmental pressures. Genes with unknown function, mostly found in the variable gene set, were enriched in the exchanged genes relative to the total genome (up to 2-fold in *N. meningitidis* genome pairs), indicating that additional functions with likely important roles in the adaptation process remain to be elucidated. In addition to the functional bias revealed, the newly transferred genes also appeared to be located in clusters across the genome (e.g., genomic islands), even in the genome pairs with elevated fraction of exchanged genes (Fig. 4.2), indicating a strong spatial clustering for genetic exchange and (positive) selection for the corresponding functions.

While several theories have been advanced in order to explain how bacterial populations evolve (sexual vs. asexual), our understanding of how clusters of closely related genotypes emerge and are maintained under high rates of gene transfer is far from complete. In this study, we attempted to obtain data that test the predicted signatures of these theories on the genome. Overall, our results indicated that in most of the genome pairs studied, exchanged genes were randomly distributed around the genome when (most) genes under selection were removed from the analysis, and these genes appeared to be enough for sexual evolution of genomes within the 95-100% ANI genome clusters (Table 4.1). Even when the most conservative estimation was used based on the spatial distance of the recently exchanged genes in the genome, the effect of recombination was very close to, albeit a bit smaller in general, than that of mutation. If one also considers that this analysis was applied to genomes pairs with observed F_{100} close to the expect one (i.e., not extremely recombining pairs), and our approach likely underestimates recombination due to the effect of recently exchanged genes on ANI and thus, observed F_{100} (see also below), these

findings indicated that strong clonality or asexual speciation is likely not occurring, at least for the genomes examined here.

While our methodology purposely targeted recent exchange events, it is important to point out that it still encompassed timescales in the order of several thousand years. For instance, *E. coli* has an estimated mutation rate of $4.5x10^{-9}$ per nucleotide per generation (56) and between 100 to 300 generations per year (57). For a 1000bp gene, the average gene length of bacterial genomes, this mutation rate and number of generations per year translate to about 1,000 years to observe a single fixed synonymous nucleotide mutation in the gene. In the case of nonsynonymous mutations, it would take about 20,000 years to observe a single, fixed nonsynonymous substitution by chance alone, based on an average estimate of 1:20 ratio of synonymous to nonsynonymous substitutions in the *E. coli* genomes (30) [assuming no strong selection for the mutation, as is the case for neutral or nearly neutral mutations]. When we restricted our analysis of recently exchanged genes to those with nonsynonymous substitutions only, by not considering synonymous nucleotide changes, we also observed the strong functional and spatial biases described above. Thus, it appears that the genomes analyzed here have been exchanging (and getting fixed in the genome) genes non-randomly for at least a few thousand years.

Our model offers a robust estimation of recent gene exchange among pairs of genomes because its parameters are based on empirical data derived from a wide range of bacterial species with different lifestyles to represent the average gene exchange rate within species as a reference point. This is advantageous compared to alternative approaches because, in most methods, recombination is detected based on a neutral model assuming no recombination or on a coalescent model with no selection and no population structure (58). Our approach most likely underestimates the rate of gene flow (relative to the reference database average) in highly recombinogenic bacteria since an elevated frequency of F_{100} increases the ANI value between two genomes, i.e., recent gene exchange and ANI are not totally independent from each other. However, our model is not based on a recombination constant but rather relies on an expected F_{100} value given an ANI range. Further, our simulations showed that when we artificially increased the number of 100% identity genes in a genome pair by randomly introducing nucleotide changes in genes that

were not originally 100% identical (which increased ANI of the pair), the genome pairs that were detected as outliers based on our original analysis (Fig. 4.1), were still outliers up to when the ANI reached ~99.97% (Fig. C.11). Therefore, even if closely related genomes might have higher chances of gene exchange, the effect of 100% identical genes on the ANI value and our ability to detect and quantify outliers is problematic only for very identical genomes (>99.97% ANI) or extremely high frequency of genetic exchange (>80-90% of the total genes affected), and did not affect our conclusions substantially.

Our model can be easily implemented and is generally applicable to any set of genomes of the same species. The input data, which consist of the fraction of shared genes with 100% nucleotide identity and the genome relatedness (ANI), can be estimated using the scripts ani.rb and rbm.rb available as part of the enveomics collection (33) or directly uploading FastA sequences of a genome pair to the online ANI calculator at http://enve-omics.ce.gatech.edu/. The core and accessory genome can be estimated using the script ogs.mcl.rb. Table C.5 indicates the running time that the model and scripts take to estimate the fraction of recent exchange as well as the COGs in a set of genomes. The fraction of recent exchange in a group of genomes can be efficiently calculated in seconds and the identification of candidate genes requires more time, and varies with the number of genomes since the prediction of COGs is required. The procedure and set of scripts used in this step are described in the Materials and Methods section. This last step is needed only when it is desirable to extract the list of recently transferred genes in a genome pair.

Our methodology and associated model offer an important addition to the toolbox for studying recombination and gene content adaptation, especially during relative short timescales. Interestingly, the estimates of recent genetic exchange by this model allowed us to evaluate the mode of bacterial evolution under HGT such as that sexual maintenance may be possible and recent recombination may act as a cohesive force that counteracts mutational divergence, at least for the genomes analyzed here.

4.6 CONCLUSIONS AND PERSPECTIVES

The increasing availability of NGS data in public repositories has opened the possibility to study bacterial population at the genome level revealing important properties of their evolution mode. Advances of comparative genomics have allowed us to advance our knowledge of bacterial population dynamics and the effects of gene exchange in shaping microbial community structure. However, due to the volume and complexity of genomic datasets raise computational challenges, currently available tools applied to large-scale data are limited. Thus, new and efficient computational/bioinformatic approaches are urgent required.

In this chapter, we applied Bayesian inference to estimate and quantify recent events of genetic exchange in genome pairs from a collection of hundreds bacterial genomes. Bayes models have been applied to analyze NGS data including genome-wide association studies, protein-protein interactions, and bacterial evolution. Bayesian approaches allow us to integrate diverse data types, unravel high dimensional problems, and analyze large-scale data sets. However, it can be computationally intensive when they are applied to infer phylogenetic relationships using NGS data.

Currently, multiple bioinformatics tools and algorithms have been developed to minimize computational cost and scale analysis over parallel computer. Among the routine tasks to tackle omics data include data reduction, feature selection, and data selection. However, as NGS technologies continue to improve, much future work in improving the efficacy and accuracy of computational tools will be required in order to extract biological insights from omics data, answer biological questions, and understand the complexity of biological systems.

4.7 ACKNOWLEDGEMENTS

We would like to thank Brani Vidakovic for his help with statistical analysis. This work was supported by US National Science Foundation (award numbers 1356288 and 1241046) to KTK.

4.8 **REFERENCES**

1. Thomas CM, Nielsen KM. 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nature Reviews Microbiology 3:711–721.

2. Vos M. 2009. Why do bacteria engage in homologous recombination? Trends in Microbiology 17:226–232.

3. Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. Science 320:237–239.

4. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336:48–51.

5. Caro-Quintero A, Auchtung J, Deng J, Brettar I, Hofle M, Tiedje JM, Konstantinidis KT. 2012. Genome Sequencing of Five *Shewanella baltica* Strains Recovered from the Oxic-Anoxic Interface of the Baltic Sea. Journal of Bacteriology 194:1236–1236.

6. Caro-Quintero A, Rodriguez-Castano GP, Konstantinidis KT. 2009. Genomic Insights into the Convergence and Pathogenicity Factors of *Campylobacter jejuni* and *Campylobacter coli* Species. Journal of Bacteriology 191:5824–5831.

7. Eisen JA. 2000. Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol 3:475–480.

8. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742.

9. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62.

10. Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. Mol Biol Evol 28:1057–1074.

11. Corel E, Méheust R, Watson AK, McInerney JO, Lopez P, Bapteste E. 2018. Bipartite Network Analysis of Gene Sharings in the Microbial World. Mol Biol Evol 35:899–913.

12. Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10:1719–1725.

13. Ragan MA. 2001. Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev 11:620–626.

14. Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. Nature Reviews Genetics 5:251–261.

15. Xu Y. 2008. Computational Methods for Understanding Bacterial and Archaeal Genomes. Imperial College Press.

16. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria. Microbial Genomics 2.

17. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. 2009. Hyper-Recombination, Diversity, and Antibiotic Resistance in Pneumococcus. Science 324:1454–1457.

18. Marttinen P, Baldwin A, Hanage WP, Dowson C, Mahenthiralingam E, Corander J. 2008. Bayesian modeling of recombination events in bacterial populations. BMC Bioinformatics 9:421.

19. Darling AE, Miklós I, Ragan MA. 2008. Dynamics of Genome Rearrangement in Bacterial Populations. PLoS Genetics 4:e1000128.

20. Sandberg R. 2001. Capturing Whole-Genome Characteristics in Short Sequences Using a Naive Bayesian Classifier. Genome Research 11:1404–1409.

21. Driebe EM, Sahl JW, Roe C, Bowers JR, Schupp JM, Gillece JD, Kelley E, Price LB, Pearson TR, Hepp CM, Brzoska PM, Cummings CA, Furtado MR, Andersen PS, Stegger M, Engelthaler DM, Keim PS. 2015. Using Whole Genome Analysis to Examine Recombination across Diverse Sequence Types of *Staphylococcus aureus*. PLOS ONE 10:e0130955.

22. den Bakker HC, Didelot X, Fortes ED, Nightingale K, Wiedmann M. 2008. Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. BMC Evolutionary Biology 8:277.

23. Sikorski J, Teschner N, Wackernagel W. 2002. Highly different levels of natural transformation are associated with genomic subgroups within a local population of *Pseudomonas stutzeri* from soil. Appl Environ Microbiol 68:865–873.

24. Perron GG, Lee AEG, Wang Y, Huang WE, Barraclough TG. 2012. Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations. Proceedings of the Royal Society B: Biological Sciences 279:1477–1484.

25. Mostowy R, Croucher NJ, Hanage WP, Harris SR, Bentley S, Fraser C. 2014. Heterogeneity in the Frequency and Characteristics of Homologous Recombination in Pneumococcal Evolution. PLoS Genetics 10:e1004300.

26. Hacker J, Carniel E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. EMBO reports 2:376–381.

27. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102:2567–2572.

28. Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology 57:81–91.

29. Rodriguez-R LM, Castro JC, Kyrpides NC, Cole JR, Tiedje JM, Konstantinidis KT. 2018. How much do rRNA gene surveys underestimate extant bacterial diversity? Appl Environ Microbiol AEM.00014-18.

30. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397.

31. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research 40:D115–D122.

32. Rodriguez-R LM. MIGA: Microbial Genome Atlas. MIGA.

33. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. e1900v1. PeerJ Preprints.

34. Delignette-Muller M, Dutang C. 2015. fitdistrplus : An R Package for Fitting Distributions.

35. Graham JR. 1995. *Helicobacter pylori*: human pathogen or simply an opportunist? Lancet 345:1095–1097.

36. Christoffersen TE, Jensen H, Kleiveland CR, Dorum G, Jacobsen M, Lea T. 2012. In vitro comparison of commensal, probiotic and pathogenic strains of *Enterococcus faecalis*. Br J Nutr 108:2043–2053.

37. Bernardini G, Renzone G, Comanducci M, Mini R, Arena S, D'Ambrosio C, Bambini S, Trabalzini L, Grandi G, Martelli P, Achtman M, Scaloni A, Ratti G, Santucci A. 2004. Proteome analysis of *Neisseria meningitidis* serogroup A. Proteomics 4:2893–2926.

38. Conlan AJ., Coward C, Grant AJ, Maskell DJ, Gog JR. 2007. *Campylobacter jejuni* colonization and transmission in broiler chickens: a modelling perspective. Journal of The Royal Society Interface 4:819–829.

39. Ketley JM, Konkel ME. 2005. Campylobacter: Molecular and Cellular Biology. Horizon Scientific Press.

40. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. 2013. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. Proceedings of the National Academy of Sciences 110:13880–13885.

41. Sheppard SK, Cheng L, Méric G, de Haan CPA, Llarena A-K, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, Strachan NJC, Forbes K, Colles FM, Jolley KA, Bentley SD, Maiden MCJ, Hänninen M-L, Parkhill J, Hanage WP, Corander J. 2014. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. Molecular Ecology 23:2442–2451.

42. Lapierre P, Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. Trends in Genetics 25:107–110.

43. Nogueira T, Touchon M, Rocha EPC. 2012. Rapid Evolution of the Sequences and Gene Repertoires of Secreted Proteins in Bacteria. PLoS ONE 7:e49403.

44. Hotopp JCD, Grifantini R, Kumar N, Tzeng YL, Fouts D, Frigimelica E, Draghi M, Giuliani MM, Rappuoli R, Stephens DS, Grandi G, Tettelin H. 2006. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. Microbiology 152:3733–3749.

45. Kong Y, Ma JH, Warren K, Tsang RSW, Low DE, Jamieson FB, Alexander DC, Hao W. 2013. Homologous Recombination Drives Both Sequence Diversity and Gene Content Variation in *Neisseria meningitidis*. Genome Biology and Evolution 5:1611–1627.

46. Schoen C, Tettelin H, Parkhill J, Frosch M. 2009. Genome flexibility in *Neisseria meningitidis*. Vaccine 27:B103–B111.

47. Moran P a. P. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17–23.

48. Anderson TW. 1962. On the Distribution of the Two-Sample Cramér-von Mises Criterion. The Annals of Mathematical Statistics 33:1148–1159.

49. Watson GS. 1962. Goodness-of-Fit Tests on a Circle. II. Biometrika 49:57–63.

50. Kolmogoroff A. 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Mathematische Annalen 104:415–458.

51. Yoon SH, Park Y-K, Kim JF. 2015. PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic Acids Research 43:D624–D630.

52. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, Brinkman FS. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research 45:W30–W35.

53. Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21:537–539.

54. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. 2017. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Molecular Biology and Evolution 34:1167–1182.

55. Corander J, Waldmann P, Sillanpää MJ. 2003. Bayesian Analysis of Genetic Differentiation Between Populations. Genetics 163:367.

56. Guttman DS, Dykhuizen DE. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380–1383.

57. Gibbons RJ, Kapsimalis B. 1967. Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. J Bacteriol 93:510–512.

58. Wall JD. 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. By Jotun Hein, Mikkel H Schierup, and Carsten Wiuf. The Quarterly Review of Biology 80:473–474.

APPENDIX A. SUPPLEMENTARY MATERIAL CHAPTER 2

SUPPLEMENTARY TABLES AND FIGURES

Table A. 1. General statistics of five biofilm metagenomes from the shower hoses

	Sample				
	M1	**M3**	**M4**	**M5**	**M6**
Assembly					
No. of reads (millions)	18.01	14.96	29.48	37.03	23.43
No. of contigs*	15,323	2,878	9,394	18,311	21,910
N50 of contigs (b)	3,242	19,478	6,469	7,483	6,226
Average size (Mb)	34.86	15.95	25.23	58.99	51.11
No. of predicted genes	45,366	17,676	31,656	70,196	64,737
Taxonomy classification (%)**					
Mycobacterium	45.3	58.5	52.5	26.3	30
Erythrobacter	13.8	6.71	9.94	9.6	5.81
Sphingomonas	3.87	6.34	10.7	6.41	6.24
Novosphingobium	3.83	5.99	5.4	3.65	2.06
Bradyrhizobium	2.66	0.22	0.18	15.45	8.24
Citromicrobium	2.09	2.02	1.73	1.47	1.02
Sphingobium	1.41	6	8.98	2.28	2.21

*Contigs longer than 5000 bases were counted.

** Relative abundance of taxa at genus level based on annotated proteins recovered in each sample and classified by MyTaxa.

Table A. 2. Statistics of the 94 isolate genomes recovered from the shower hose biofilms.

Isolate ID	OTU designation	Ident (%)	GG content (%)	Assembly size (Mb)	No of contigs	No of protein-coding genes
CCH3-A3	*Blastomonas*	100	64.24	4.34	162	4,238
CCH9-F3	*Blastomonas*	100	64.26	4.18	154	4,078
CCH6-A6	*Blastomonas*	100	64.16	4.39	169	4,301
CCH8-E1	*Blastomonas*	100	64.33	4.27	145	4,167
CCH10-E1	*Blastomonas*	100	64.18	4.35	230	4,292
CCH9-A1	*Blastomonas*	100	64.19	4.2	95	4,083
CCH13-E1	*Blastomonas*	97.4	64.2	4.36	159	4,265
CCH2-E1	*Blastomonas*	100	64.24	4.4	164	4,344
CCH4-A2	*Acidovorax delafieldii*	100	64.29	4.28	192	4,155
CCH5-A5	*Sphingomonas*	98.2	66.88	3.55	81	3,410
CCH3-E3	*Blastomonas*	100	64.27	4.35	153	4,256
CCH8-A3	*Blastomonas*	100	64.11	4.57	217	4,567
CCH18-B1	*Sphingomonas*	99.7	64.23	4.32	173	4,198
CCH7-E1	*Blastomonas*	100	64.18	4.19	111	4,067
CCH15-G10	*Blastomonas*	100	64.22	4.35	192	4,300
CCH6-E2	*Blastomonas*	100	64.07	4.49	1088	5,127
CCH13-A3	*Blastomonas*	100	64.23	4.33	148	4,197

Table A.2. continued

CCH21-G11	*Sphingomonas*	99.8	66.73	3.65	59	3,483
CCH5-E3	*Blastomonas*	100	64.31	4.3	189	4,254
CCH4-D12	*Blastomonas*	100	64.29	4.28	173	4,148
CCH12-A3	*Novosphingobium*	100	63.27	5.48	328	5,404
CCH3-A5	*Rhizobiales*	99.7	66.5	5.1	208	5,042
CCH5-F6	*Bradyrhizobium*	99.7	64.12	8.15	206	7,766
CCH12-B7	*Dermacoccus*	99.8	67.66	4.75	279	4,559
CCH16-B10	*Sphingomonas*	99.8	66.85	3.56	74	3,386
CCH9-G4	*Pseudoxanthomo-nas mexicana*	99.8	66.55	3.98	105	3,718
CCH10-E5	*Rhizobiales*	99.6	67.66	5.32	252	5,333
CCH3-G3	*Acidovorax delafieldii*	100	64.78	5.79	634	5,698
CCH6-A11	*Sphingomonas*	100	68.67	3.87	261	3,839
CCH6-A12	*Neisseria perflava*	100	41.73	2.18	105	2,121
CCH10-B3	*Sphingomonas*	99.8	65.15	3.83	98	3,779
CCH9-A3	*Rhizobiales*	99.7	67.34	5.69	281	5,658
CCH4-E10	*Chryseobacterium*	100	36.66	4.41	195	4,120
CCH4-A6	*Bradyrhizobiaceae*	99.8	60.89	5.28	159	5,103
CCH10-C7	*Bradyrhizobiaceae*	99.8	60.72	5.6	201	5,427
CCH5-A3	*Blastomonas*	100	64.26	4.43	132	4,334

Table A.2. continued

CCH19-E1	*Porphyrobacter donghaensis*	100	64.32	3.96	237	3,914
CCH12-A10	*Comamonadaceae*	100	63.34	4.87	470	4,791
CCH7-A10	*Porphyrobacter donghaensis*	100	51.11	2.52	564	2,521
CCH1-A6	*Blastomonas*	100	64.37	4.8	168	4,740
CCH10-H12	*Neisseria perflava*	100	51.12	2.52	561	2,495
CCH4-C5	*Comamonadaceae*	100	63.19	4.61	456	4,509
CCH4-E1	*Caulobacter*	99.8	66.39	4.12	216	4,128
CCH5-D3	*Streptococcus*	99.8	39.94	2.23	102	2,099
CCH5-D2	*Methylobacterium*	99.8	71.09	6.09	420	6,144
CCH1-B1	*Bradyrhizobiacea*	99.7	67.36	5.81	302	5,764
CCH8-H5	*Streptococcus*	100	39.8	2.2	105	2,105
CCH9-E1	*Caulobacter*	99.8	67.66	4.73	260	4,521
CCH3-A4	*Porphyrobacter donghaensis*	100	66.58	4.16	236	4,162
CCH20-B6	*Sphingomonas*	99.8	66.62	4.08	100	3,963
CCH5-E12	*Caulobacter*	99.5	66.74	4.95	189	4,741
CCH8-G7	*Streptococcus*	99.3	40.43	2.24	81	2,215

Table A.2. continued

CCH9-H8	*Sphingomonas*	99.2	66.06	4.57	220	4,529
CCH7-A1	*Porphyrobacter*	100	64.23	4.41	182	4,351
CCH5-B3	*Xylophilus ampelinus*	100	68.27	6.21	320	6,056
CCH9-H3	*Phenylobacterium*	99.2	69.48	5.62	149	5,528
CCH1-A1	*Porphyrobacter donghaensis*	99.8	66.77	4.21	215	4,178
CCH6-E1	*Porphyrobacter donghaensis*	100	66.45	4.28	275	5,127
CCH5-A9	*Bosea*	99.2	67.39	5.65	278	5,628
CCH9-E2	*Sphingomonas*	99.2	66.09	4.31	203	4,260
CCH8-C6	*Streptococcus*	100	43.37	2.21	107	2,205
CCH9-F2	*Sphingomonas*	100	68.51	4.08	261	4,082
CCH6-A4	*Rhizobiales*	99.5	66.82	6.25	380	6,339
CCH8-A2	*Porphyrobacter donghaensis*	100	66.47	4.33	274	4,449
CCH8-D1	*Rhizobiales*	99.7	66.78	6.16	295	6,160
CCH17-B8	*Porphyrobacter donghaensis*	100	66.58	4.17	223	4,136
CCH2-D9	*Dermacoccus*	99.7	69.19	3.01	58	2,733

Table A.2. continued

CCH12-G6	*Porphyrobacter donghaensis*	100	66.92	6.13	454	6,291
CCH11-D2	*Rhizobiales*	99.7	66.91	6.12	340	6,181
CCH5-D1	*Microbacterium*	100	68.25	4.01	110	3,918
CCH2-A4	*Rhizobiales*	99.7	66.86	6.17	350	6,246
CCH11-B1	*Sphingobium*	99.5	64	4.77	212	4,713
CCH6-A1	*Porphyrobacter donghaensis*	100	66.64	4.39	246	4,384
CCH10-A2	*Mycobacterium mucogenicum*	100	66.8	6.72	531	6,883
CCH11-A4	*Blastomonas*	100	64.18	4.23	152	4,122
CCH3-E2	*Micrococcus luteus*	100	73.18	2.51	237	2,369
CCH19-C6	*Sphingomonas*	99.8	66.7	3.93	109	3,819
CCH12-B4	*Phenylobacterium*	99.7	69.56	5.55	199	5,487
CCH12-C2	*Erythrobacteraceae*	99.5	63.86	4.18	124	4,053
CCH15-F11	*Sphingomonas*	99.8	66.77	4.35	109	4,236
CCH5-D11	*Sphingomonas*	99.1	65.55	4.44	116	4,235
CCH18-H6	*Sphingomonas*	99.8	66.66	4.22	107	4,108
CCH5-H10	*Rhodospirillaceae*	99.7	66.24	6.05	240	5,941
CCH2-A2	*Blastomonas*	100	64.26	4.41	180	4,357

Table A.2. continued

CCH12-A2	*Mycobacterium mucogenicum*	100	66.84	6.65	370	6,653
CCH7-E3	*Porphyrobacter donghaensis*	100	66.85	4.1	212	4,064
CCH7-A2	*Blastomonas*	100	67.36	6.02	279	5,959
CCH15-A1	*Sphingomonas*	99.2	67.87	5.11	3796	7,882
CCH13-B11	*Sphingomonas*	99.7	66.76	3.89	64	3,766
CCH5-A1	*Porphyrobacter donghaensis*	99.7	66.4	4.33	241	4,309
CCH6-D9	*Corynebacterium durum*	99.3	49.23	4.69	3475	6,946
CCH7-B2	*Sphingomonas*	99.8	66.77	9.82	5483	10,000
CCH15-E2	*Porphyrobacter*	100	57.23	6.21	251	6,054
CCH12-A4	*Blastomonas*	100	63.77	9.12	592	5,487

Table A. 3. General statistics of the binned populations recovered from the shower hose biofilm metagenomes.

	Binned population genomes	
	***Blastomonas* sp.**	***Mycobacterium* sp.**
Average sequencing depth	41.40	115.9
Size of the genome (Mb)	7.36	6.4
Completeness (%)	94.9	100
Contamination (%)	39.3	2.0
GC content (%)	7.36	6.4
Average nucleotide identity (%)	84.2	85.9
Virulence-associated genes*	0	15
Antibiotic resistance profile	Aminoglycoside, polymyxin, kanamycin, macrolide, bacitracin	Fluoroquinolone, penicillin, cephalosporin, gentamicin, Netilmicin

* BLASTp searches against the Virulence Factors of Pathogenic Bacteria and PATRIC databases.

Table A. 4. Functional categories and relative abundance of protein sequences recovered from the biofilm metagenomes.

Relative abundance was calculated as the number of metagenomic proteins annotated to a specific function in the UniProt database divided by the total number of annotated proteins in the respective sample.

UniProt id	Name	Sample				
		M1	M3	M4	M5	M6
Q50615	PE-PGRS family protein PE_PGRS33	0.23	0.01	0.48	0.21	0.18
P50360	Protein y4hP	0.16	0.42	0.1	0.12	0.13
Q79FV4	pyridoxal phosphate-dependent protein	0.12	0.17	0.15	0.09	0.09
P0A690	PE-PGRS family protein PE_PGRS46	0.12	0.01	0.22	0.15	0.07
Q10637	PE-PGRS family protein PE_PGRS24	0.12	0.27	0.24	0.12	0.11
P35883	Transposase for insertion element IS6120	0.12	0.21	0.15	0.01	0.09
O86034	D-beta-hydroxybutyrate dehydrogenase	0.07	0.14	0.05	0.09	0.11
P55501	Uncharacterized protein y4jA/y4nE/y4sE	0.06	0.14	0.03	0.07	0.06
P08080	Transposase for insertion element	0.04	0.14	0.01	0.12	0.14
A5TY80	Insertion element IS6110 protein	0.06	0.14	0.06	0.02	0.04
Q2G6U3	Protein translocase subunit SecA	0.03	0.14	0.05	0.09	0.07
P17985	Insertion element ISR1 protein A3	0.04	0.07	0.01	0.13	0.06
P38054	Cation efflux system protein CusA	0.06	0.13	0.12	0.09	0.08
A1KQG0	Phthioceranic/hydroxyphthioceranic acid synthase	0.2	0.01	0.12	0.1	0.07
P72003	Serine/threonine-protein kinase PknF	0.16	0.01	0.2	0.01	0.09
O65934	ABC transporter ATP-binding/permease protein Rv1747	0.16	0.01	0.14	0.01	0.21
P9WPS6.1	Probable cation-transporting ATPase G	0.14	0.01	0.01	0.01	0.01
P96218	Glutamate synthase [NADPH]	0.14	0.1	0.06	0.01	0.06
P60230	Transposase for insertion element IS1081	0.13	0.13	0.13	0.11	0.11
Q02251	Mycocerosic acid synthase	0.13	0.01	0.05	0.01	0.05

Table A.4. continued

P9WPS2.1	Probable copper-exporting P-type ATPase V	0.13	0.01	0.01	0.01	0.01
O53735	Putative membrane protein mmpL4	0.13	0.07	0.12	0.01	0.01
O53303	Putative alcohol dehydrogenase D	0.12	0.2	0.13	0.01	0.01
Q57307	Cholesterol oxidase	0.12	0.07	0.09	0.01	0.01
Q2G480	Phosphoenolpyruvate carboxykinase	0.043	0.1	0.12	0.01	0.01
P55390	Probable cold shock protein y4cH	0.01	0.01	0.01	0.08	0.11
Q5NRH4	Glutamine-fructose-6-phosphate aminotransferase	0.01	0.001	0.09	0.08	0.07
	Disinfectant mechanisms					
Q9KU26	Extracellular polymeric substance	0.086	0.03	0.03	0.14	0.084
P37578	60 kDa chaperonin GroEL1	0.302	0.17	0.33	0.32	0.253
L8F435	GDP-mannose 6-dehydrogenase AlgD	0.014	0.42	0.01	0.01	0.01
R4R145	Alginate biosynthesis protein AlgA	0.000	0.01	0.02	0.01	0.011
P04425	Glutathione synthetase gshB	0.043	0.07	0.1	0.1	0.095
Q73VT8	Putative glutathione reductase gorA	0.014	0.03	0.02	0.01	0.011
A0R692	DNA protection during starvation protein	0.331	0.001	0.38	0.42	0.242
P52214	Thioredoxin reductase trxB	0.475	0.56	0.58	0.5	0.442
Q51506	Redox-sensitive transcriptional activator SoxR	0.043	0.07	0.03	0.04	0.032
P9WGE7	Superoxide dismutase sodA	0.029	0.01	0.01	0.07	0.053
Q1I657	RNA polymerase sigma factor RpoS	0.000	0.07	0.17	0.12	0.179
P9WGE9	Superoxide dismutase SodC	0.058	0.07	0.09	0.06	0.084
P96273	Exodeoxyribonuclease III protein XthA	0.000	0.14	0.12	0.18	0.168
A0PSD2	Hydrogen peroxide-inducible genes activator, OxyR	0.173	0.1	0.15	0.21	0.116
A0QYP1	Catalase-peroxidase katG	0.274	0.21	0.19	0.28	0.2

Table A.4. continued

Q9I6Z2	Alkyl hydroperoxide reductase ahpF	0.086	0.07	0.09	0.05	0.042
P31224	Multidrug efflux pump subunit AcrB	0.070	0.01 4	0.03 4	0.00 2	0.011
P52599	Probable multidrug resistance protein emrK	0.070	0.02 9	0.03 4	0.00 1	0.011
Q98D15	Probable multidrug resistance protein NorM	0.070	0.01 4	0.01 7	0.00 1	0.021
P52002	Multidrug resistance protein MexB	0.035	0.01 4	0.01 7	0.00 1	0.021
P78966	Mating factor M secretion protein mam1	0.035	0.00 0	0.00	0.00 3	0.00
Q6D2B1	Multidrug resistance protein MdtB	0.011	0.01	0.00	0.00 3	0.01
Q73V87	Multidrug resistance protein mmr	0.035	0.00	0.01 7	0.00 1	0.01
P34713	Multidrug resistance protein PGP3	0.035	0.01 4	0.00 0	0.00	0.00
Q1GR76	UPP (Bacitracin resistance protein) (Undecaprenyl pyrophosphate phosphatase) UPPP	0.070	0.02 9	0.05 1	0.00 3	0.032

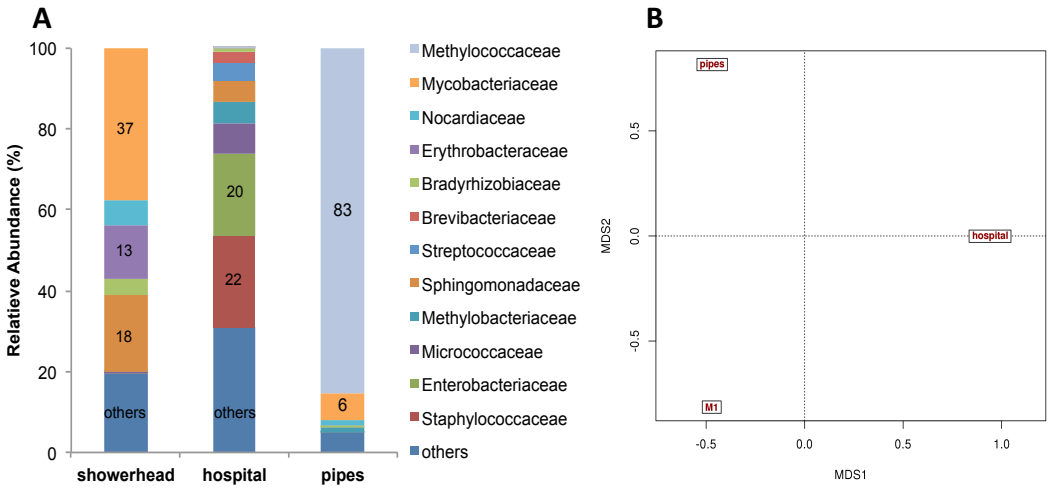


Figure A. 1 Comparison of the taxonomic profile of the shower hose metagenome with other microbial communities at the family level.

(A) The bacterial community structures were assessed using 16S rRNA gene-encoding metagenomic reads. The dataset called "Hospital" was collected from an ICU ward surface of the University Hospital A Coruña, Spain, (SRA ID: SRX099356). The dataset called "Pipes" was collected from drinking water pipes in Florida, USA (SRA ID: SRX472092). (B) Multidimensional scaling (MDS) plot based on the relative abundance of taxa on each dataset using the Bray–Curtis distance.

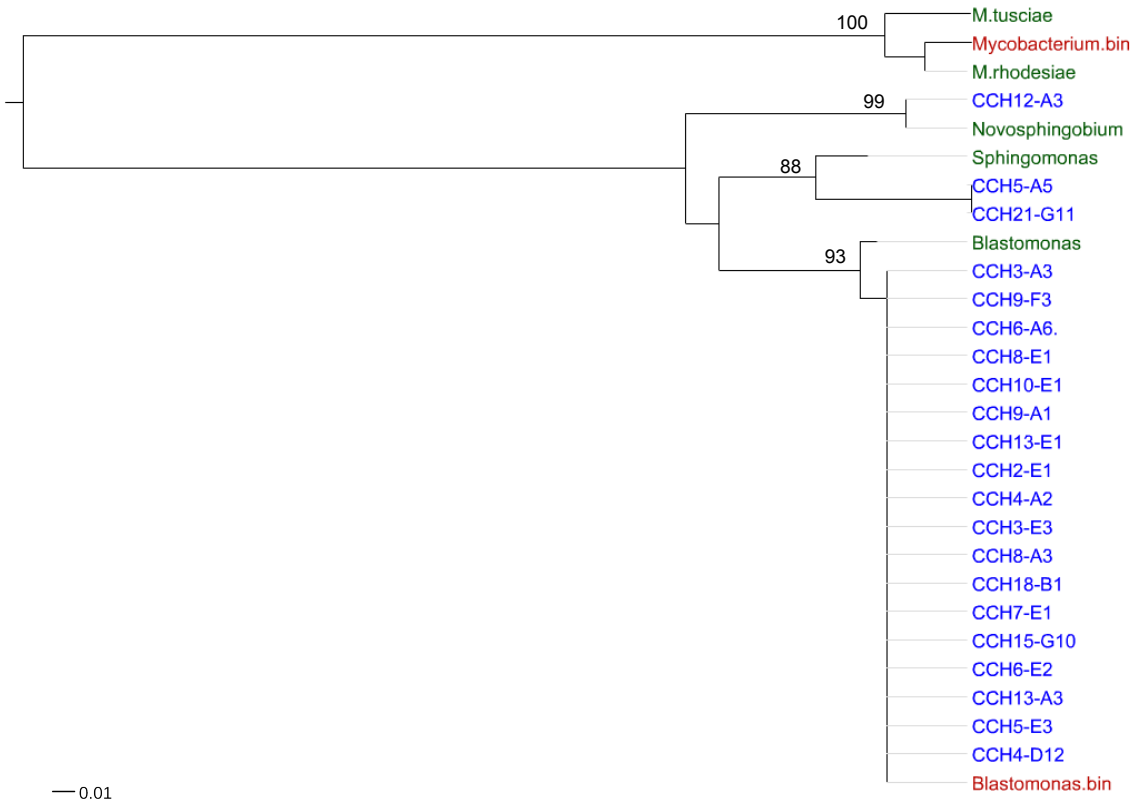


Figure A. 2. Phylogenetic relationships among population bins and isolate genomes based on 30S ribosomal protein S20 sequences.

The isolates from the shower hoses are colored in blue; the recovered bins in red, and the reference genomes in green. The phylogeny was generated using the Neighbor-joining algorithm with 1000 bootstrap replicates using MEGA V.5. The number at nodes indicates the bootstrap support. Scale bar represents substitutions per site.

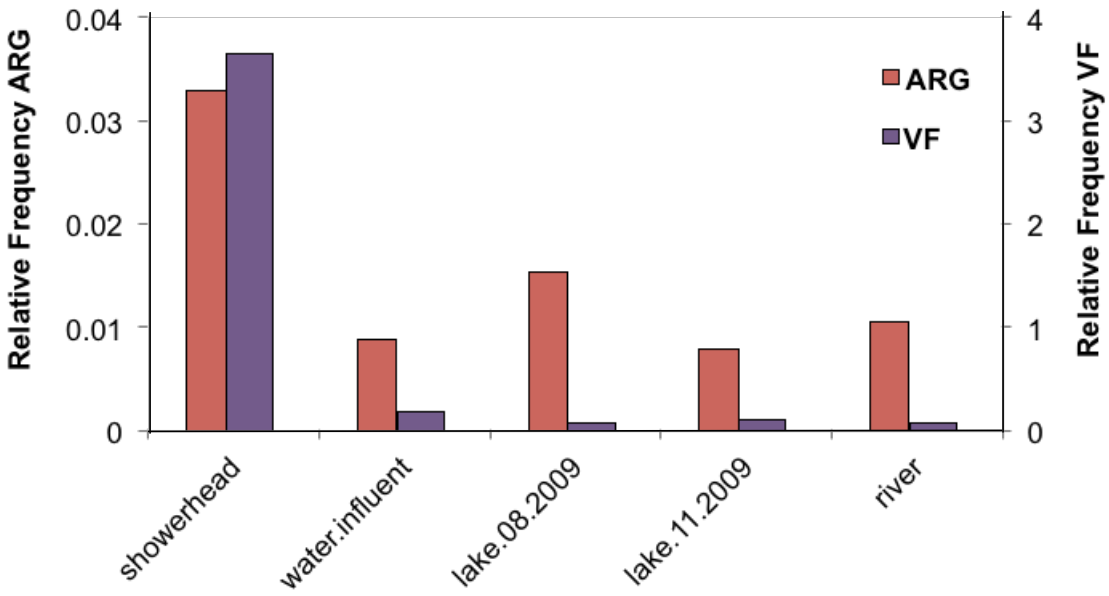


Figure A. 3. Metagenomic comparison of the genes involved in antibiotic resistance (ARG) and virulence (VF) mechanisms between the shower hose and other water-associated microbial communities.

The "river" dataset corresponds to samples from river water (SRA ID: SRR1022353), the "drinking.water" dataset from a drinking water treatment plant in China (SRA ID: SRR835363), the "lake.08.2009" dataset collected during summertime (SRA ID: SRR096386), and the "lake.11.2009" dataset collected during falltime (SRA ID: SRR096389) from Lake Lanier in Georgia, USA. Pearson's Chi-squared test values between the shower hose metagenome and each water metagenome were significant (p-value < 0.05).

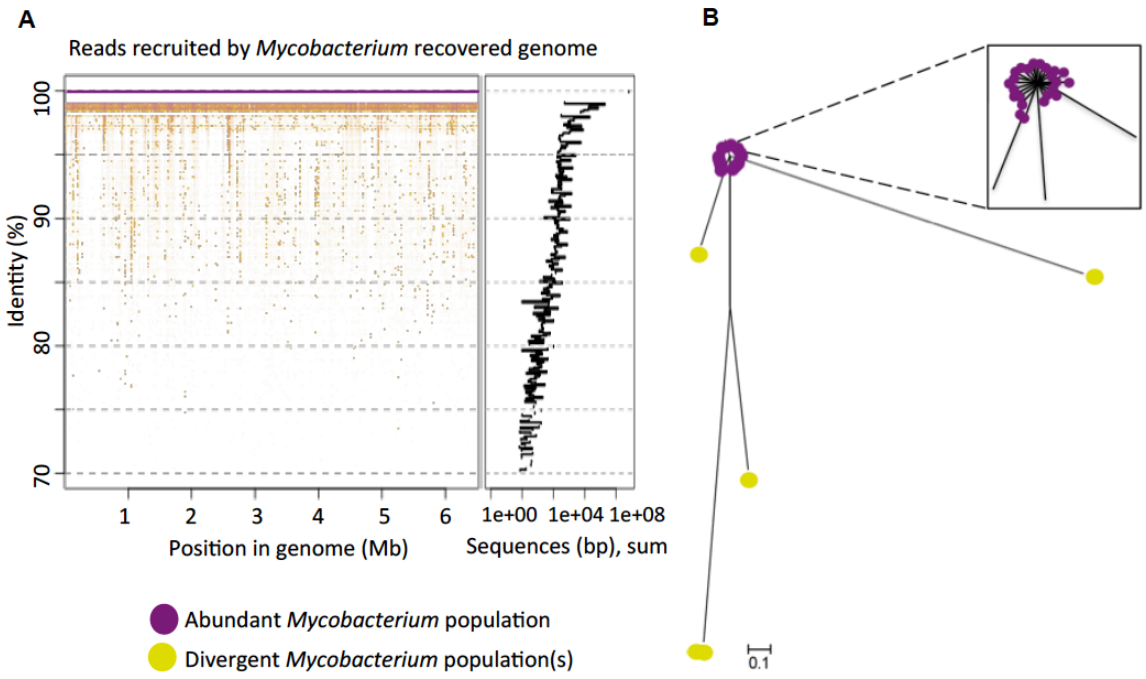


Figure A. 4. Intra-population diversity of the abundant *Mycobacterium* population recovered from the shower hose metagenomes.

(A) Fragment recruitment plot of the recovered *Mycobacterium* population versus the M1 metagenome. Metagenomic reads were searched against the recovered genome sequence using a cut-off of at least 70% nucleotide identity and complete alignment to the genome reference. The y-axis corresponds to the identity of each read and the x-axis to the position of the read mapped on the genome. The histogram on the right represents the sum of the total base pairs of the reads recruited per unit of nucleotide identity. (B) Neighbor-joining phylogenetic tree of metagenomic reads that mapped on the single-copy 30S ribosomal protein S9 encoded on the recovered genome. Inset represents a zoomed in view of the tree where the reads representing the abundant *Mycobacterium* sp. population clustered together (in purple color). Note the star-like phylogeny formed by the latter reads. Scale bar represents substitutions per site.

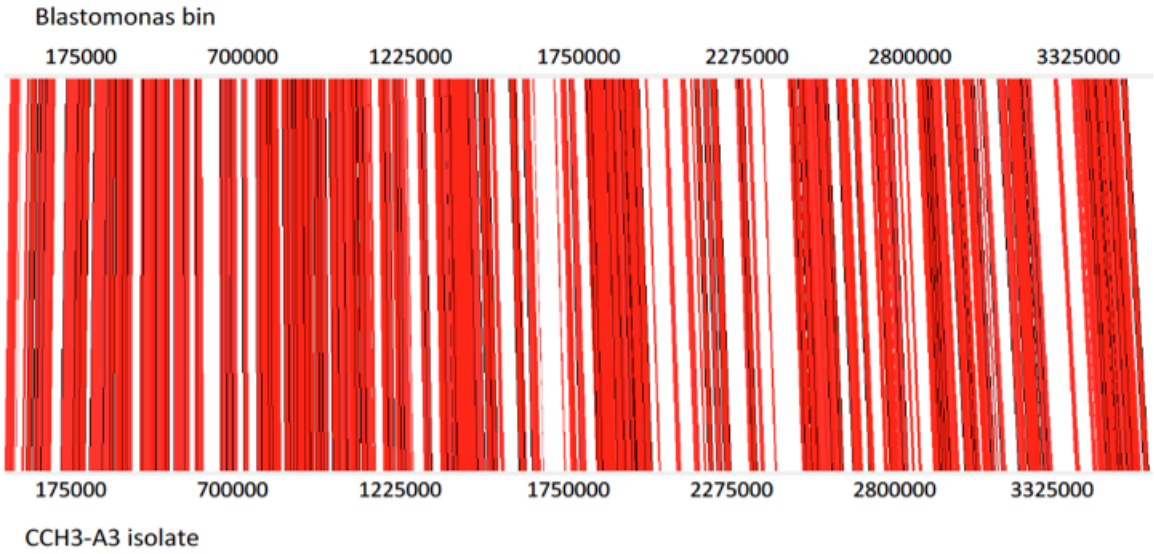


Figure A. 5. Genome alignment of the (binned) *Blastomonas* population genome against the *Blastomonas* isolate genome.

The Artemis Comparison Tool (ACT) was used to visualize the alignment of the two genomes. Contigs of the *Blastomonas* bin (top) were ordered based on homology searches and (assumed) synteny with a *Blastomonas* isolate genome available in GenBank (Accession number: GCA_000331245.1). Red bars indicate regions of similarity with the same orientation; empty/white bars indicate regions of gene content differences. Note, however, that most of the gene-content differences probably represent sequencing gaps (e.g., neither genome was complete) as opposed to real gene-content differences.

APPENDIX B. SUPPLEMENTARY MATERIAL CHAPTER 3

SUPPLEMENTARY TABLES AND FIGURES

Table B. 1. Metadata for all samples analyzed in this study

Sample ID	Location	Age	Age category	Clinical status	Sex	Race
Q101	Quito	4	Children	control	female	mestizo
Q104	Quito	4	Children	control	female	mestizo
Q105	Quito	1	YoungChildren	control	male	mestizo
Q106	Quito	7	Children	control	male	mestizo
Q107	Quito	0	New.born	control	female	mestizo
Q108	Quito	10	PreAdolescent	case	female	mestizo
Q116	Quito	2	YoungChildren	control	male	mestizo
Q117	Quito	10	PreAdolescent	control	male	mestizo
Q127	Quito	1	YoungChildren	control	male	mestizo
Q128	Quito	19	Adult	case	female	mestizo
Q130	Quito	45	Adult	control	female	mestizo
Q131	Quito	5	Children	control	male	mestizo
Q132	Quito	3	YoungChildren	control	male	mestizo
Q133	Quito	31	Adult	control	female	mestizo
Q139	Quito	4	Children	case	female	mestizo
Q142	Quito	33	Adult	case	female	mestizo
Q143	Quito	6	Children	control	male	mestizo
Q144	Quito	6	Children	control	female	mestizo
Q145	Quito	30	Adult	control	female	mestizo
Q146	Quito	46	Adult	control	male	mestizo
Q147	Quito	31	Adult	case	female	mestizo
Q148	Quito	9	PreAdolescent	control	female	mestizo
Q157	Quito	2	YoungChildren	control	female	mestizo
Q158	Quito	1	YoungChildren	control	male	mestizo
Q159	Quito	17	PreAdolescent	control	female	mestizo
Q160	Quito	0	New.born	control	male	mestizo
Q168	Quito	5	Children	control	female	mestizo

Table B. 1. continued

Q169	Quito	21	Adult	control	female	mestizo
Q170	Quito	12	PreAdolescent	control	female	mestizo
Q174	Quito	0	Babies	control	male	mestizo
Q178	Quito	11	PreAdolescent	control	male	mestizo
Q186	Quito	8	PreAdolescent	control	male	mestizo
Q188	Quito	0	Babies	control	female	mestizo
Q189	Quito	0	Babies	control	male	mestizo
Q192	Quito	25	Adult	control	female	mestizo
Q196	Quito	4	Children	case	male	mestizo
Q199	Quito	0	Babies	control	female	mestizo
Q203	Quito	0	New.born	control	male	mestizo
Q207	Quito	13	PreAdolescent	case	male	blanco
Q212	Quito	12	PreAdolescent	control	male	mestizo
Q215	Quito	0	New.born	control	male	mestizo
Q223	Quito	40	Adult	case	female	mestizo
Q227	Quito	9	PreAdolescent	case	female	mestizo
Q233	Quito	2	YoungChildren	case	male	mestizo
Q239	Quito	1	YoungChildren	case	male	mestizo
Q240	Quito	44	Adult	case	male	mestizo
Q243	Quito	43	Adult	case	female	mestizo
Q245	Quito	3	YoungChildren	control	male	mestizo
Q249	Quito	4	Children	control	male	mestizo
Q253	Quito	2	YoungChildren	control	male	mestizo
Q259	Quito	26	Adult	control	female	Amerindian
Q270	Quito	20	Adult	case	female	mestizo
Q275	Quito	49	Adult	control	female	mestizo
Q282	Quito	2	YoungChildren	control	female	mestizo
Q284	Quito	1	YoungChildren	control	female	mestizo
Q288	Quito	8	PreAdolescent	control	male	mestizo
Q289	Quito	30	Adult	control	female	mestizo
Q291	Quito	0	Babies	case	male	mestizo
Q294	Quito	1	YoungChildren	case	female	mestizo
Q295	Quito	78	Adult	case	female	mestizo
Q300	Quito	4	Children	case	male	mestizo
Q304	Quito	3	YoungChildren	case	female	mestizo
Q308	Quito	1	YoungChildren	case	female	manaba
Q310	Quito	1	YoungChildren	case	female	mestizo
Q312	Quito	0	Babies	control	male	mestizo

Table B. 1. continued

Q40	Quito	38	Adult	case	male	mestizo
Q49	Quito	1	YoungChildren	case	male	mestizo
Q51	Quito	4	Children	case	male	manaba
Q53	Quito	5	Children	case	female	Amerindian
Q56	Quito	1	YoungChildren	case	female	mestizo
Q57	Quito	1	YoungChildren	case	male	mestizo
Q61	Quito	1	YoungChildren	case	male	mestizo
Q65	Quito	2	YoungChildren	case	female	mestizo
Q69	Quito	49	Adult	control	male	mestizo
Q70	Quito	32	Adult	control	female	NA
Q71	Quito	1	YoungChildren	case	male	mestizo
Q74	Quito	64	Adult	case	female	mestizo
Q83	Quito	40	Adult	control	female	mestizo
Q86	Quito	1	YoungChildren	control	male	mestizo
Q87	Quito	1	YoungChildren	control	male	mestizo
Q89	Quito	31	Adult	case	male	mestizo
Q90	Quito	27	Adult	case	female	mestizo
Q91	Quito	26	Adult	case	male	mestizo
Q92	Quito	18	Adult	case	male	mestizo
Q97	Quito	41	Adult	control	female	mestizo
Q98	Quito	61	Adult	control	female	mestizo
Q99	Quito	47	Adult	control	female	mestizo
R0001	Villages	1	YoungChildren	case	female	Amerindian
R0003	Villages	3	YoungChildren	case	male	Amerindian
R0006	Villages	2	YoungChildren	case	male	Amerindian
R0007	Villages	1	YoungChildren	case	male	mestizo
R0008	Villages	1	YoungChildren	control	male	African American
R0009	Villages	2	YoungChildren	case	male	mestizo
R0010	Villages	4	Children	case	male	Amerindian
R0011	Villages	57	Adult	case	female	Amerindian
R0012	Villages	1	YoungChildren	control	female	Amerindian
R0013	Villages	54	Adult	case	female	Amerindian
R0014	Villages	1	YoungChildren	case	female	Amerindian
R0015	Villages	3	YoungChildren	control	male	Amerindian
R0017	Villages	37	Adult	case	male	Amerindian
R0021	Villages	5	Children	case	male	Amerindian
R0022	Villages	1	YoungChildren	control	female	Amerindian

Table B. 1. continued

R0024	Villages	39	Adult	control	female	African American
R0025	Villages	4	Children	control	male	African American
R0026	Villages	2	YoungChildren	control	female	mestizo
R0029	Villages	5	Children	control	female	mestizo
R0030	Villages	8	PreAdolescent	control	female	mestizo
R0031	Villages	33	Adult	case	female	Amerindian
R0032	Villages	5	Children	case	male	Amerindian
R0039	Villages	32	Adult	case	female	African American
R0040	Villages	4	Children	control	male	mestizo
R0041	Villages	6	Children	control	male	mestizo
R0042	Villages	30	Adult	case	male	African American
R0043	Villages	3	YoungChildren	control	male	African American
R0044	Villages	10	PreAdolescent	case	male	African American
R0045	Villages	5	Children	case	female	mestizo
R0046	Villages	5	Children	control	male	mestizo
R0050	Villages	3	YoungChildren	case	male	Amerindian
R0051	Villages	5	Children	case	male	Amerindian
R0052	Villages	2	YoungChildren	control	male	Amerindian
R0053	Villages	15	PreAdolescent	control	female	mestizo
R0054	Villages	15	PreAdolescent	control	male	African American
R0055	Villages	15	PreAdolescent	control	female	African American
R0056	Villages	16	PreAdolescent	control	male	Amerindian
R0057	Villages	4	Children	case	female	Amerindian
R0058	Villages	13	PreAdolescent	control	male	Amerindian
R0059	Villages	6	Children	case	male	African American
R0060	Villages	38	Adult	case	female	African American
R0061	Villages	37	Adult	control	female	African American
R0062	Villages	4	Children	case	male	African American

Table B. 1. continued

R0063	Villages	8	PreAdolescent	case	male	African American
R0064	Villages	16	PreAdolescent	control	male	mestizo
R0065	Villages	51	Adult	case	female	African American
R0066	Villages	8	PreAdolescent	case	female	African American
R0067	Villages	10	PreAdolescent	control	male	African American
R0068	Villages	12	PreAdolescent	control	female	African American
R0071	Villages	5	Children	case	male	mestizo
R0074	Villages	9	PreAdolescent	control	female	African American
R0076	Villages	0	Babies	case	male	African American
R0077	Villages	1	YoungChildren	case	male	African American
R0078	Villages	2	YoungChildren	case	female	mestizo
R0079	Villages	2	YoungChildren	case	male	African American
R0080	Villages	1	YoungChildren	control	female	African American
R0081	Villages	1	YoungChildren	control	male	African American
R0083	Villages	2	YoungChildren	control	male	African American
R0084	Villages	1	YoungChildren	control	male	African American
R0085	Villages	2	YoungChildren	control	female	African American
R0088	Villages	1	YoungChildren	case	male	African American
R0090	Villages	1	YoungChildren	control	female	African American
R0091	Villages	4	Children	control	male	African American
R0093	Villages	2	YoungChildren	case	female	African American
R0097	Villages	2	YoungChildren	control	female	African American

Table B. 1. continued

R0098	Villages	8	PreAdolescent	control	male	African American
R0101	Villages	1	YoungChildren	case	female	African American
R0102	Villages	0	Babies	case	female	African American
R0104	Villages	9	PreAdolescent	control	male	African American
R0105	Villages	3	YoungChildren	control	female	African American
R0109	Villages	1	YoungChildren	control	female	African American
R0110	Villages	0	Babies	control	male	African American
R0111	Villages	2	YoungChildren	case	female	Amerindian
R0113	Villages	0	New.born	case	male	manaba
R0114	Villages	2	YoungChildren	case	male	mestizo
R0116	Villages	17	PreAdolescent	control	male	African American
R0118	Villages	57	Adult	case	female	mestizo
R0119	Villages	7	Children	control	male	African American
R0120	Villages	45	Adult	case	female	African American
R0122	Villages	11	PreAdolescent	case	male	African American
R0123	Villages	0	New.born	control	male	African American
R0124	Villages	2	YoungChildren	control	male	African American
R0125	Villages	7	Children	case	male	African American
R0126	Villages	1	YoungChildren	case	female	African American
R0127	Villages	7	Children	case	male	African American
R0128	Villages	6	Children	control	male	African American
R0129	Villages	2	YoungChildren	control	male	African American
R0130	Villages	5	Children	control	male	African American
R0131	Villages	4	Children	control	male	African American

Table B. 1. continued

R0132	Villages	7	Children	control	female	African American
R0134	Villages	1	YoungChildren	control	male	African American
R0135	Villages	2	YoungChildren	case	female	mestizo
R0136	Villages	59	Adult	case	male	African American
R0137	Villages	2	YoungChildren	control	male	African American
R0138	Villages	14	PreAdolescent	case	female	African American

Table B. 2. Permutational multivariate analysis of variance (PERMANOVA) at the OTU level.

R^2 values indicate the amount of variation attributed to each categorical factor. PERMANOVA was conducted on the Bray-Curtis distance matrix using the ADONIS function in the vegan R package with 999 permutations of residuals. Levels of significance: *P < 0.05, **P< 0.01, ***P < 0.001.

Variable	Df	Sus ofSqs	MeanSqs	F model	R^2	Pr(>F)	
Location	1	1.82	1.82	4.98	0.02	0.001	***
Age	5	4.12	0.82	2.25	0.06	0.001	***
Race	4	1.82	0.45	1.24	0.02	0.027	*
Delivery mode	3	1.31	0.43	1.2	0.012	0.071	
Education	3	0.85	0.28	0.77	0.01	0.977	
House sanitation	2	0.74	0.37	1.01	0.013	0.424	
Water treatment type	2	0.94	0.47	1.29	0.008	0.059	
Water treatment	1	626	0.62	1.71	0.004	0.012	*
Gender	1	0.29	0.29	0.81	0.82	0.789	
Residuals	158	57.7	0.36		1		
Total	180	70.32					

Table B. 3 Significant associations between location and microbial abundances at all taxonomic levels controlling for the effects of race and Age using MaAsLin (Multivariate microbial Association by Linear models) with a Q-value < 0.1, as a cut-off for significant associations.

Variable	Feature	Value	Coefficient	P-value	Q-value
Location	k__Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Prevotellaceae|Prevotella|copri	Rural	0.185	0.0032	0.055
Location	k__Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Comamonadaceae|Comamonas	Rural	0.006	0.0013	0.027
Location	k__Bacteria|Elusimicrobia|Elusimicrobia|Elusimicrobiales|Elusimicrobiaceae	Rural	0.01	2.4E-05	0.001
Location	k__Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Rikenellaceae	Rural	-0.069	5.2E-06	0.0007
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|Oscillospira	Babies	-0.103	0.0011	0.079
Age	k__Archaea|Euryarchaeota|Methanobacteria|Methanobacteriales|Methanobacteriaceae|Methanobrevibacter|	Babies	-0.025	0.0011	0.079
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Christensenellaceae	Babies	-0.019	0.0010	0.075
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Clostridiaceae|Clostridium	Babies	-0.025	0.0007	0.058
Age	k__Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|_Barnesiellaceae	Babies	-0.068	0.0006	0.055
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|_Ruminococcus	Babies	-0.047	0.0001	0.018
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|Ruminococcus	Babies	-0.081	0.0001	0.018
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae	Babies	-0.155	2.2E-05	0.004
Age	k__Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Rikenellaceae	Babies	-0.143	1.0E-05	0.003
Age	k__Bacteria|Actinobacteria|Actinobacteria|Bifidobacteriales|Bifidobacteriaceae|Bifidobacterium_Other	Babies	0.133	1.1E-05	0.003
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae_Other_Other	Babies	-0.12	1.15E-06	0.007
Age	k__Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Bacteroidaceae|Bacteroides|eggerthii	Children	-0.051	0.0001953	0.022
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|Faecalibacterium|prausnitzii	New born	-0.156	1.30E-05	0.003
Age	k__Bacteria|Firmicutes|Clostridia|Clostridiales|Clostridiaceae|Clostridium|perfringens	New born	0.046	6.97E-07	0.0007
Age	k__Bacteria|Firmicutes|Bacilli|Lactobacillales|Enterococcaceae|Enterococcus	New born	0.05	1.65E-07	0.0003

Table B. 4. Network topological parameters calculated by NetworkAnalyzer (Cytoscape v3.6.1) and SPIEC-EASI (SParse InversE Covariance Estimation for Ecological ASsociation Inference) of the rural and urban OTU co-occurrence networks during a healthy and a disease state.

Parameter	Urban		Rural	
	Control	Case	Control	Case
Clustering coefficient	0.064	0.099	0.082	0.086
Connected components	1	1	1	6
Network diameter	10	10	12	13
Network radius	6	6	7	1
Network centralization	0.031	0.031	0.02	0.015
Shortest path	104652	51756	155630	130692
Avg. number of neighbors	4.025	4.009	4.05	3.38
Network density	0.012	0.018	0.01	0.009
Network heterogeneity	0.497	0.45	0.514	0.517
Number of nodes	324	228	395	372
Number of edges	652	457	801	630
Number of positive edges	541	367	684	585
Number of negative edges	111	90	117	45

Table B. 5. Topological features of the OTUs with the highest number of edges (Top 10) identified in healthy rural and urban OTU co-occurrence network analysis

RURAL

OTU	Number of egdes	Average Shortest path length	Clustering Coefficient	Betweenness centrality
f_Bacteroidaceae|*Bacteroides uniformis*	12	3.961	0.136	0.038
f_S24-7|	10	3.703	0.066	0.062
f_Ruminococcaceae| *Oscillospira*	10	3.791	0.088	0.063
f_Methanobacteriaceae| *Methanobrevibacter*	10	3.667	0.088	0.059
o_Clostridiales	9	4.22	0.277	0.046
f_Paraprevotellaceae|CF231	9	3.878	0.083	0.036
f_Ruminococcaceae| *Anaerofilum*	9	3.931	0.083	0.027
f_Ruminococcaceae| *Faecalibacterium prausnitzii*	9	4.116	0.055	0.025
f_Ruminococcaceae	9	3.964	0.083	0.034
f_Prevotellaceae| *Prevotella stercorea*	9	4.162	0.055	0.033

URBAN

OTU	Number of egdes	Average Shortest path length	Clustering Coefficient	Betweenness centrality
f_Ruminococcaceae| *Oscillospira*	14	3.65	0.087	0.072
f_Bifidobacteriaceae| *Bifidobacterium*	10	3.823	0.011	0.036
f_Erysipelotrichaceae	10	4.068	0.017	0.019
f_Ruminococcaceae| *Oscillospira*	10	3.681	0.044	0.057
f_Ruminococcaceae| *Faecalibacterium prausnitzii*	10	3.486	0.066	0.071
f_Veilonellaceae| *Anaerovibrio*	10	3.743	0.013	0.032
f_Paraprevotellaceae|CF231	10	3.901	0.015	0.032
f_Enterobacteriaceae| *Escherichia coli*	9	4.191	0.194	0.013
f_Veilonellaceae| *Dialister*	9	3.938	0.055	0.031
f_Lachnospiraceae| *Blautia*	9	3.891	0.055	0.045

Table B. 6. General statistics of recovered genome populations (MAGs) from urban and rural metagenomes

Urban								
Bin name	Abundanc	Completeness	Genome size	Mul.copy gene	No. of contigs	N50	No. genes	Taxonomy
MG37.007	18.28	94.40%	3733296	0	132	65315	3327	f__Lachnospiraceae
MG37.008	15.07	98.10%	4970725	0	158	72329	4228	s__Parabacteroides_merdae
MG37.009	14.62	94.40%	2531943	7	214	86138	2257	g__Alistipes
MG37.012	8.97	93.50%	4513646	1	321	25650	3736	f__Porphyromonadaceae
MG37.013	5.75	86.90%	1894597	1	486	5751	1829	f__Ruminococcaceae
MG38.011	11.17	100.00%	4556472	4	255	32183	4231	f__Enterobacteriaceae
MG38.017	6.84	86.00%	1823641	0	367	7089	1689	g__Lactobacillus
MG38.018	6.27	93.50%	1753329	1	479	4739	1621	s__Lactobacillus_ruminis
MG39.001	64.88	97.20%	2046595	0	42	92351	2080	f__Ruminococcaceae
MG39.002	58	96.30%	3075878	0	136	81106	2643	f__Porphyromonadaceae
MG39.005	28.9	95.30%	3516048	0	232	31856	2948	g__Prevotella
MG39.006	17.79	96.30%	4080469	4	612	88405	3299	g__Alistipes
MG39.008	14.43	90.70%	2841400	0	437	15660	2457	f__Ruminococcaceae
MG41.005	17.86	99.10%	3199525	1	346	39692	3003	g__Faecalibacterium
MG42.001	103.46	86.90%	2847111	0	427	13734	2349	g__Prevotella
MG42.002	70.51	98.10%	2658054	0	81	59598	2369	f__Coriobacteriaceae
MG42.004	61.84	96.30%	2276505	0	46	116041	1935	s__Bifidobacterium_bifidum
MG42.005	44.06	100.00%	2098131	0	85	81805	1935	g__Lactobacillus
MG42.007	31.95	97.20%	2440673	0	207	29005	2035	g__Dialister
MG42.015	8.23	86.00%	1807730	1	458	5421	1620	s__Lactobacillus_ruminis
MG43.002	93.08	98.10%	2189335	0	77	138660	2067	g__Dialister
MG43.008	15.62	94.40%	2620289	0	173	28119	2406	g__Ruminococcus
MG43.010	12.4	98.10%	2964917	0	265	32224	2756	g__Faecalibacterium
MG43.011	12.37	97.20%	4886734	7	769	49598	3924	g__Alistipes
MG43.014	10.17	99.10%	4398925	0	263	31720	4165	f__Enterobacteriaceae
MG45.003	107.69	93.50%	3042737	0	85	72663	2441	g__Alistipes
MG45.004	64.87	97.20%	4589847	2	147	92570	3838	g__Bacteroides
MG45.006	51.39	96.30%	2991689	4	116	102579	2475	s__Sutterella_wadsworthensis
MG45.007	50.64	96.30%	3107789	0	102	81179	2859	g__Lachnobacterium
MG45.009	40.5	98.10%	3179948	0	88	58192	2551	g__Barnesiella
MG45.013	25.15	97.20%	2701135	5	189	31484	2503	g__Lachnospira
MG45.015	22.08	93.50%	2976026	0	152	53996	2508	g__Alistipes
MG45.018	17.94	94.40%	3021949	0	269	17854	2512	o__Clostridiales
MG45.022	9.6	87.90%	2478407	4	347	13033	2198	o__Clostridiales
MG46.001	177.26	96.30%	3890214	0	148	70866	3233	g__Prevotella
MG46.004	32.32	96.30%	3514909	0	164	36965	2912	g__Barnesiella
MG47.003	44.32	97.20%	4660751	0	137	132120	3843	s__Bacteroides_uniformis
MG47.005	13.32	98.10%	2716083	3	239	24476	2438	g__Ruminococcus
MG48.001	81.84	98.10%	3735897	0	78	254596	3105	k__Bacteria;pg__Barnesiella
MG48.002	74.73	97.20%	2537968	0	78	62919	2182	f__Succinivibrionaceae
MG48.003	66.02	97.20%	1835262	0	39	198237	1724	g__Dialister
MG48.004	48.87	92.50%	3102035	0	74	129418	2533	g__Alistipes
MG48.005	44.76	97.20%	2708096	0	67	68833	2301	f__Verrucomicrobiaceae
MG48.006	40.65	96.30%	5056758	0	231	70196	4305	s__Bacteroides_uniformis
MG48.007	32.5	99.10%	4379787	0	132	62863	3682	g__Odoribacter
MG48.008	25.22	96.30%	1473005	2	206	11542	1403	o__Clostridiales
MG48.022	9.6	91.60%	2085142	0	198	16951	1938	f__Ruminococcaceae
MG48.025	8.88	88.80%	2186862	1	217	20945	1771	g__Alistipes
MG49.002	201.43	94.40%	3710456	0	139	55536	3250	g__Prevotella
MG49.003	48.42	91.60%	4858283	0	255	39570	4259	s__Bacteroides_vulgatus
MG49.004	36.64	96.30%	2762346	6	111	109459	2614	g__Alistipes
MG49.014	8.14	87.90%	6011913	2	1499	6118	4771	g__Odoribacter
MG50.001	121.41	91.60%	1132980	0	51	98547	954	k__Bacteria
MG50.002	97.29	87.90%	1833833	4	64	135304	1565	o__Clostridiales
MG50.005	35.96	97.20%	2975474	0	136	36994	2386	o__Bacteroidales
MG50.006	33.82	89.70%	2160932	0	67	95372	1870	o__Clostridiales
MG50.010	17.38	95.30%	2582834	0	87	64943	1995	o__Bacteroidales
MG50.013	14.56	90.70%	2360458	2	117	37849	2138	o__Clostridiales
MG50.014	14.25	92.50%	2245686	0	168	23813	1852	g__Dialister
MG50.019	10.64	92.50%	4453327	0	374	20243	3612	f__Porphyromonadaceae
MG50.020	10.51	95.30%	3496156	5	323	23306	2784	o__Bacteroidales
MG50.021	10.06	96.30%	4448887	0	859	8356	4256	f__Enterobacteriaceae
MG52.001	88.24	97.20%	2600836	0	74	67155	2259	g__Treponema
MG52.002	76.32	96.30%	3346884	0	111	104456	2636	o__Bacteroidales
MG52.006	14.05	98.10%	3003113	2	201	43913	2556	g__Alistipes

Table B.6. continued

Rural								
Bin.id	**Abundance**	**Completeness**	**Genome size**	**multi genes**	**No. of contigs**	**N50**	**No. genes**	**Taxonomy**
MG59.002	27.02	97.20%	2503888	0	110	73574	2714	g__Treponema
MG59.005	7.81	95.30%	4540301	1	104	54764	2453	g__Shigella
MG59.014	68.35	90.70%	2869740	0	497	14975	4702	g__Lachnobacterium
MG60.006	25.78	97.20%	2256584	0	124	75584	2002	f__Coriobacteriaceae
MG60.007	22.21	93.50%	2580265	0	61	77177	2285	g__Acidaminococcus
MG60.008	21.07	88.80%	2390140	1	758	4064	2396	_Sutterella_wadsworthens
MG60.012	17.41	94.40%	2926481	3	319	46138	2609	g__Bifidobacterium
MG60.014	16.49	95.30%	2409097	0	88	40785	2228	g__Megasphaera
MG60.015	15.63	97.20%	1981287	0	219	12864	1842	g__Campylobacter
MG61.012	9.67	85.00%	4281703	7	8637	877	3290	g__Prevotella
MG61.013	9.43	97.20%	1960965	0	265	12390	1801	f__Ruminococcaceae
MG62.003	50.58	98.10%	3241538	0	75286	157	2671	g__Prevotella
MG63.001	162.73	97.20%	4228134	0	47469	182	3531	g__Prevotella
MG63.006	15.85	94.40%	2411838	0	112	34878	2123	_Bifidobacterium_bifidun
MG63.007	11.21	93.50%	3164260	0	175	34830	2666	f__Verrucomicrobiaceae
MG63.009	9.93	92.50%	2468181	0	159	28684	2068	_Sutterella_wadsworthens
MG64.002	65.76	95.30%	2649889	0	53160	103	2076	g__Prevotella
MG64.009	16.49	94.40%	5013766	1	339	36081	4733	f__Enterobacteriaceae
MG64.011	6.54	89.70%	4276051	7	1617	3286	3951	g__Faecalibacterium
MG64.013	5.57	88.80%	2370191	5	826	3895	2354	g__Sutterella
MG65.004	43.9	97.20%	3097527	0	96	85110	2708	f__Verrucomicrobiaceae
MG65.006	21.15	96.30%	4563325	0	116	58233	3732	f__Porphyromonadaceae
MG65.007	14.71	92.50%	4017810	0	236	29251	3218	g__Odoribacter
MG65.013	10.5	96.30%	2915403	7	353	23222	2745	g__Faecalibacterium
MG66.002	160.73	98.10%	3066042	0	52856	137	2425	g__Prevotella
MG66.003	128.33	97.20%	2869881	0	107052	76	2338	g__Prevotella
MG66.004	57.6	97.20%	2961713	0	52	192067	2740	f__Lachnospiraceae
MG66.006	34.74	99.10%	5070905	0	209	45180	4199	s__Bacteroides_vulgatus
MG66.013	8.59	90.70%	1821887	7	599	3840	1638	f__Ruminococcaceae
MG66.016	7.49	90.70%	2347245	0	701	4521	2024	g__Ruminococcus
MG68.003	39.4	95.30%	2927141	0	61	104975	2457	o__Clostridiales
MG68.007	19.5	92.50%	1528520	0	58	48420	1418	c__Spirochaetia
MG68.011	12.38	89.70%	2403570	3	149	40080	2084	_Sutterella_wadsworthens
MG68.014	9.72	86.00%	1691932	1	230	11538	1538	f__Ruminococcaceae
MG68.018	7.76	97.20%	2006733	0	247	18570	1865	f__Ruminococcaceae
MG68.019	7.7	95.30%	4311421	0	451	17177	4041	f__Enterobacteriaceae
MG69.002	69.62	90.70%	2633556	4	170	59107	2284	f__Succinivibrionaceae
MG69.008	15.87	95.30%	2077596	3	487	6862	1931	Bacteria
MG71.004	51.91	95.30%	2637785	3	22454	209	2026	g__Prevotella
MG71.009	21.81	86.90%	3320628	4	13640	433	2535	g__Prevotella
MG71.012	16.8	98.10%	2592073	0	113	49844	2503	o__Clostridiales
MG71.021	12.4	95.30%	3575576	6	461	25495	2729	o__Bacteroidales
MG71.040	6.02	89.70%	3657102	5	1578	2648	3397	o__Clostridiales
MG72.002	43.66	97.20%	3715711	3	109	113458	2937	o__Bacteroidales
MG72.006	20.77	86.90%	895264	0	12	107820	858	o__Rickettsiales
MG73.003	73.11	92.50%	2939039	3	48889	116	2299	g__Prevotella
MG74.001	55.31	99.10%	4970481	0	211	78836	4164	s__Bacteroides_vulgatus
MG74.002	25.37	95.30%	4022962	0	118	68391	3741	f__Lachnospiraceae
MG74.006	21.19	95.30%	2791285	0	93	68123	2335	f__Verrucomicrobiaceae
MG75.005	45.27	97.20%	2770360	0	48309	135	2161	g__Prevotella
MG75.010	13.73	95.30%	2834905	0	183	33592	2231	o__Bacteroidales
MG75.012	12.33	98.10%	1907444	1	254	12309	1775	g__Helicobacter

Table B. 7. Genomic characteristics of recovered population genomes (MAGs) classified as *Prevotella* and *Alistipes* from rural and urban metagenomes.

Urban

MAGS	Completeness	*A. finegoldii* (ANI)
MG37_009	94.40%	81.10%
MG39_006	96.30%	91.65%
MG43_011	97.20%	81.97%
MG45_003	93.50%	85.29%
MG45_015	93.50%	99.23%
MG48_004	92.50%	98.70%
MG49_004	96.30%	85.75%
MG52_006	98.10%	87.84%

Rural

MAGS	Completeness	*P. stercorea* (ANI)
MG62_003	98.10%	96.81%
MG63_001	97.20%	80.56%
MG66_002	95.30%	85.47%
MG66_003	97.20%	97.08%
MG71_009	86.90%	81.**37**%
MG73_003	92.50%	81.07%

Table B. 8. Epidemiology of the clonal complex of the *E. coli* isolates (in bold) from ADD samples. ST profile was evaluated using the Warwick MLST database (http://enterobase.warwick.ac.uk/species/ecoli) and the Clermont phylogroup membership was determined based on the correspondence between Warwick sequence type number and triplex PCR genotype as described in Clermont *et al*. (2012)

Strain ID	clinical	*ST*	*Clermont phylogroup*
B45_2	case	4	A
B46_1	control	4	A
C46_4	control	4	A
Q294	case	4	A
R17_2	case	4	A
R66_4	case	4	A
R66_5	case	4	A
B88_3	case	6	A
B118_2	control	10	A
B119_1	case	10	A
B145_4	control	10	A
B188_1	control	10	A
B201_3	case	10	A
B201_5	case	10	A
B66_1	case	10	A
B66_4	case	10	A
C21_2	control	10	A
C21_4	control	10	A
E135_2	case	10	A
E135_5	case	10	A
E57	case	10	A
Q145	control	10	A
Q147	case	10	A
Q212	control	10	A
Q240	case	10	A
Q289	control	10	A
Q300	case	10	A
R36_1	case	10	A

Table B.8 continued

R42_2	case	10	A
R67_3	control	10	A
Q308	case	93	A
B100_5	control	131	B2
B12_1	case	131	B2
E13	case	131	B2
E170	case	131	B2
E26	case	131	B2
Q295	case	131	B2
Q51	case	131	B2
Q56	case	131	B2
R86_1	case	131	B2
SE15	NA	131	B2
LF82	NA	135	B2
E173	case	394	D
E205	case	394	D
Q196	case	394	D
Q243	case	394	D
B259_1	control	517	B1
E33_4	control	517	B1
Q233	case	517	B1
Q249	control	517	B1
Q275	control	517	B1
Q142	case	636	B2
Q223	case	636	B2
Q310	case	636	B2
Q49	case	4407	B1
SE11	NA	4407	B1

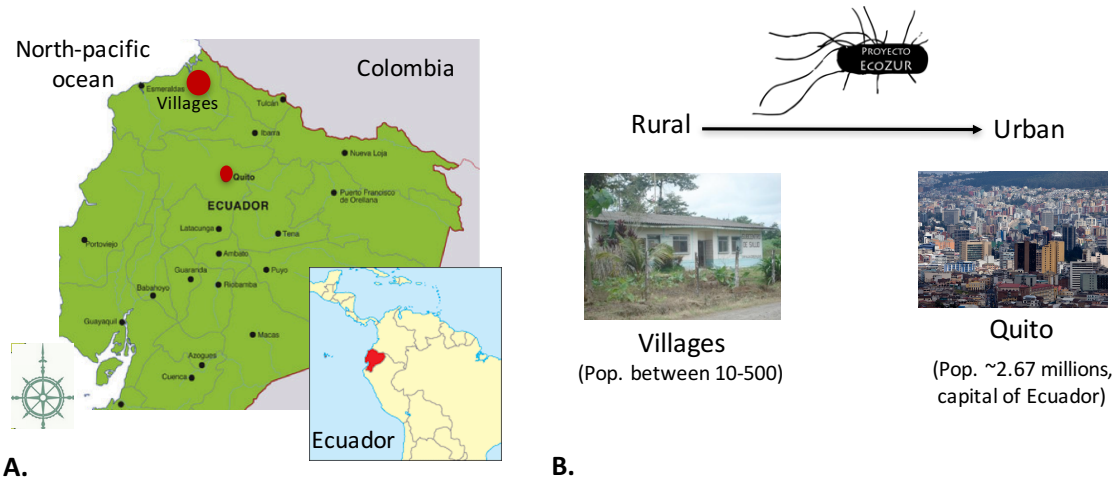


Figure B. 1. A. Map of the study site in Ecuador. Red points indicate the location of the sampling sites, including Quito (Ecuador's Capital) and where all the communities reside. B. Representative photographs of the rural and urban sites studied in Northern Ecuador.

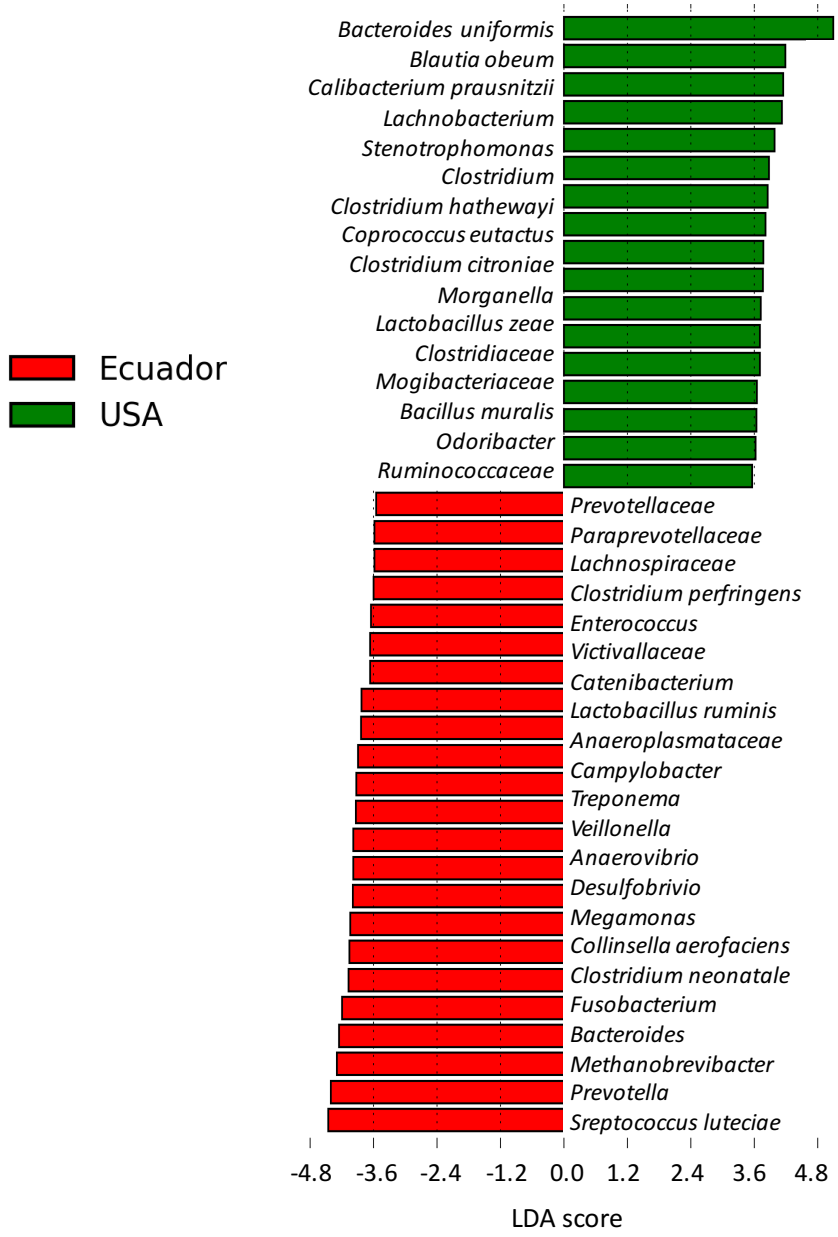


Figure B. 2. Top discriminative bacteria (LDA score > 3.5) in the gut microbiota between Ecuadorian and US populations identified by linear discriminant analysis (LDA) effect size (LEfSe) (Segata *et al.*, 2011).

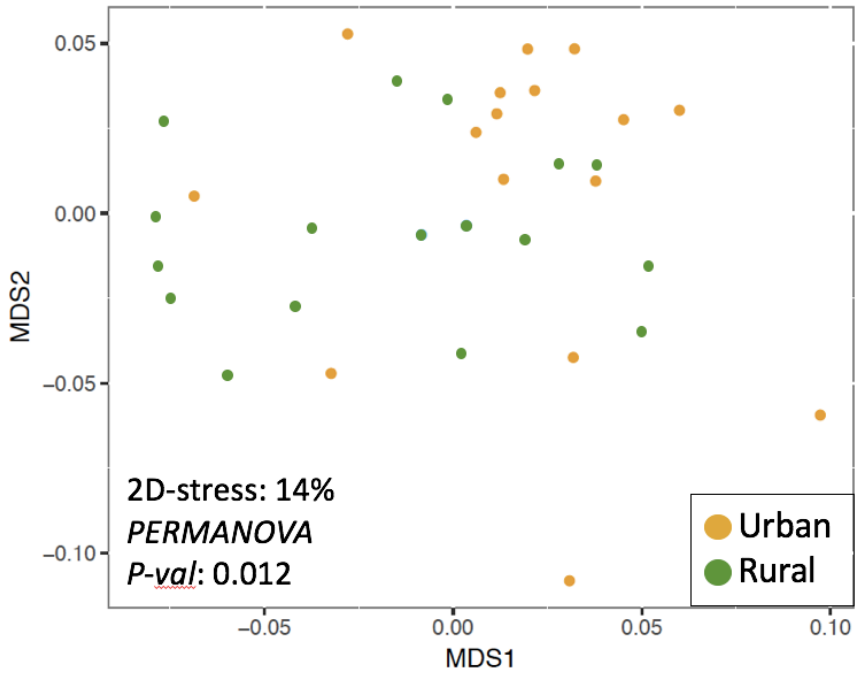


Figure B. 3. Nonmetric Multidimensional Scaling (MDS) analysis based on Mash distances with k= 21.

Each dot corresponds to a metagenomic sample and is colored according to its location (see Figure key). Mash similarities distances were calculated using Mash with default parameters between whole metagenomic datasets (Ondov *et al.*, 2016).

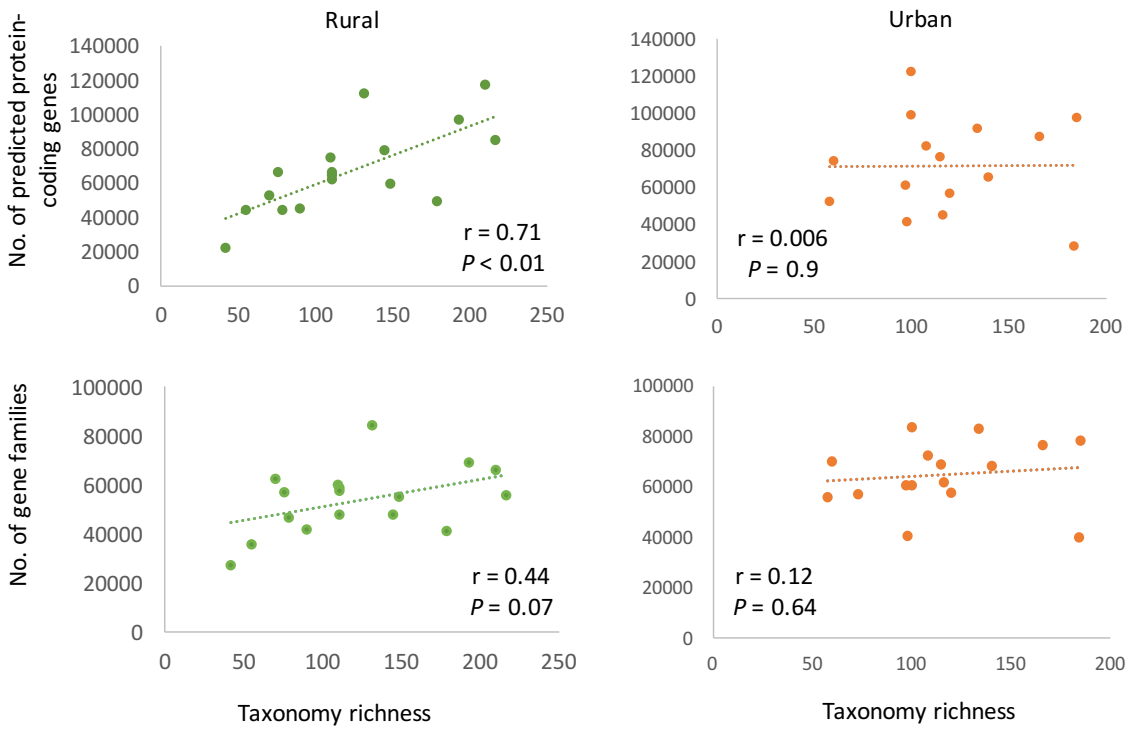


Figure B. 4. Comparison of the taxonomic and functional richness of metagenomes from rural and urban subjects.

The community richness was calculated as the number of observed OTUs based on 16S rRNA gene data and the functional richness corresponds to the total number of predicted proteins and KEGG Orthology gene families (KOs) in the metagenomes. Pearson's r values are indicated with their respective *P* value.

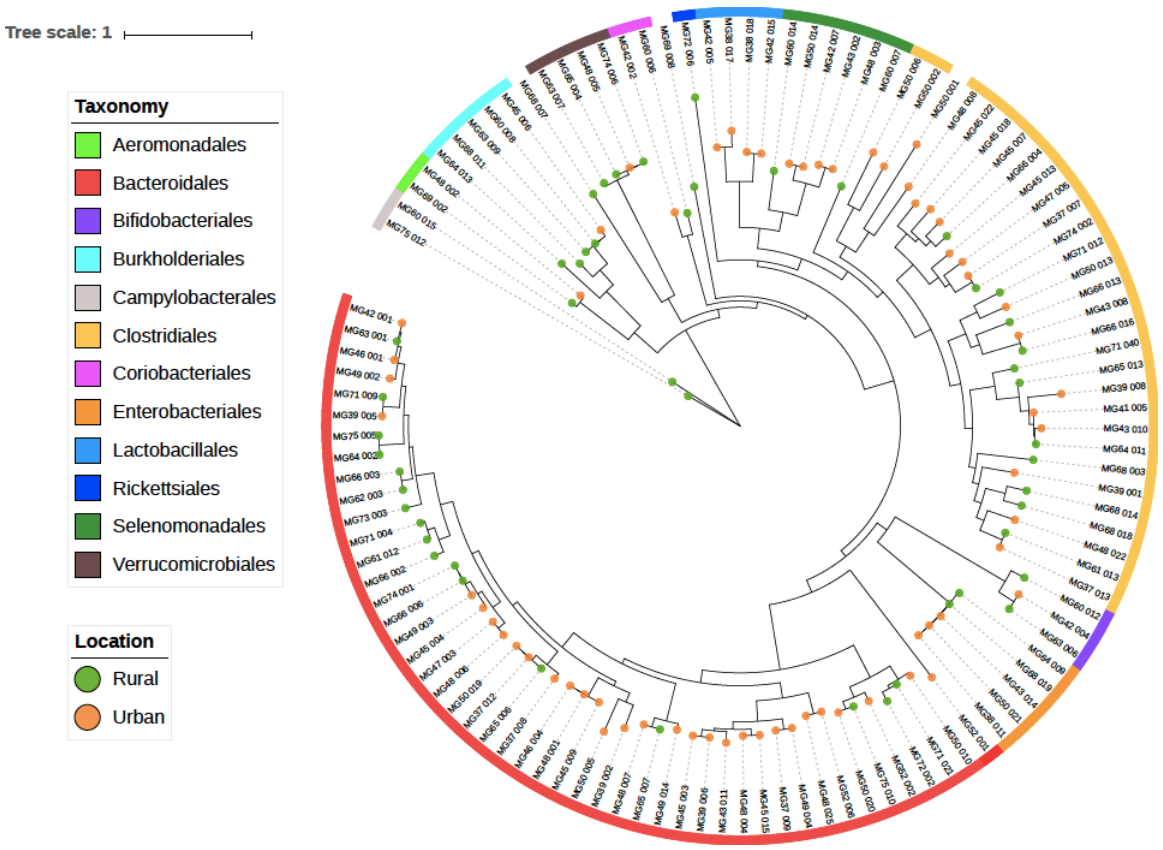


Figure B. 5. Overall bacterial diversity recovered across samples from rural and urban populations.

Phylogenetic reconstruction of MAGs based on the concatenated alignment of 8 universal single copy proteins. Maximum likelihood tree was built with RAxML v8.0.19 (Stamatakis, 2014). Colors indicate the order assigned to each MAG and its location (rural or urban).

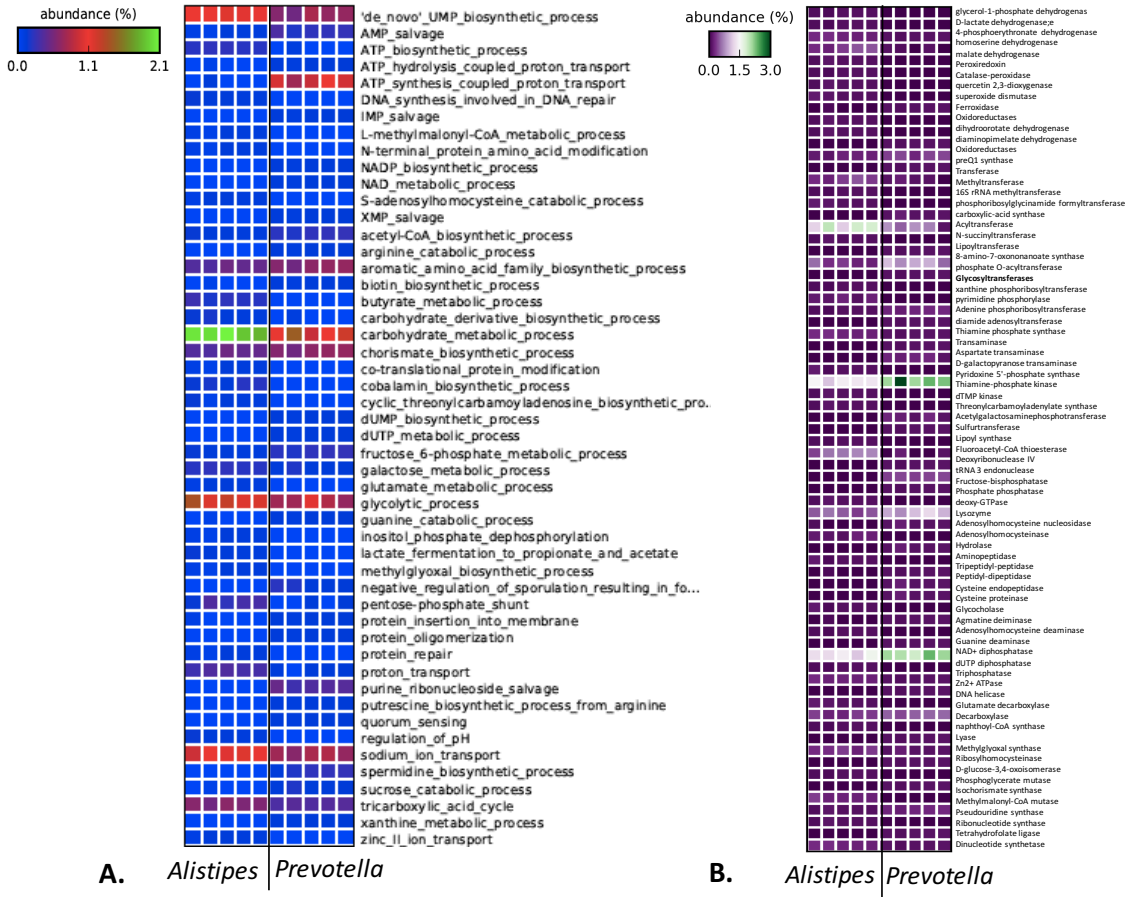


Figure B. 6. Differential functional profile of recovered *Prevotella* and *Alistipes* MAGs from healthy metagenomes.

Columns represent each MAG recovered that was classified either as *Prevotella* or *Alistipes* and rows represent the abundance of A. GO biological functions and B. Enzymes identified using the KEEG Enzyme Database that were significantly different between the two groups of MAGs (Welch's t-test, $P < 0.05$ with Benjamini-Hochberg FDR correction).

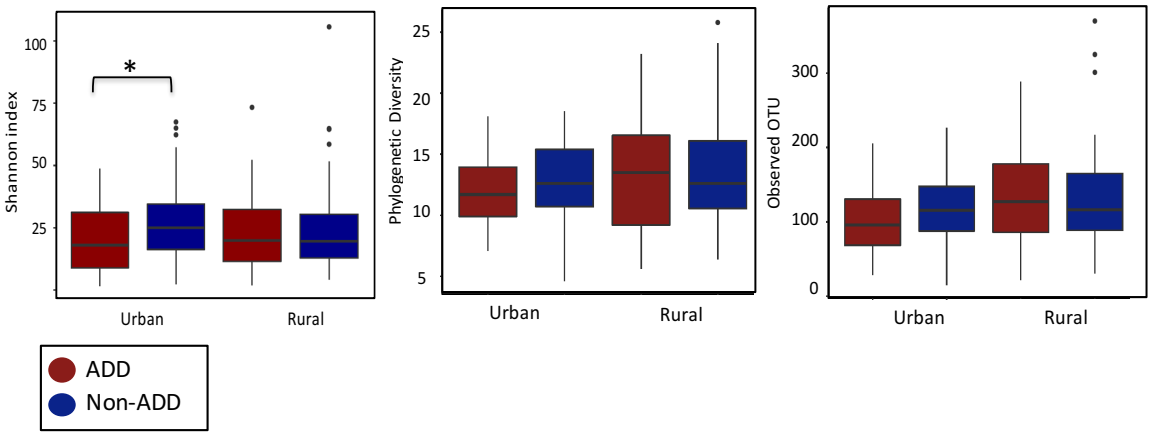


Figure B. 7. Changes in microbiota diversity during diarrheal episodes in rural and urban subjects.

16S rRNA gene-based OTU diversity boxplots (Shannon index, phylogenetic diversity, and Observed number of OTUs). *P < 0.05, **P < 0.01, ***P < 0.001.

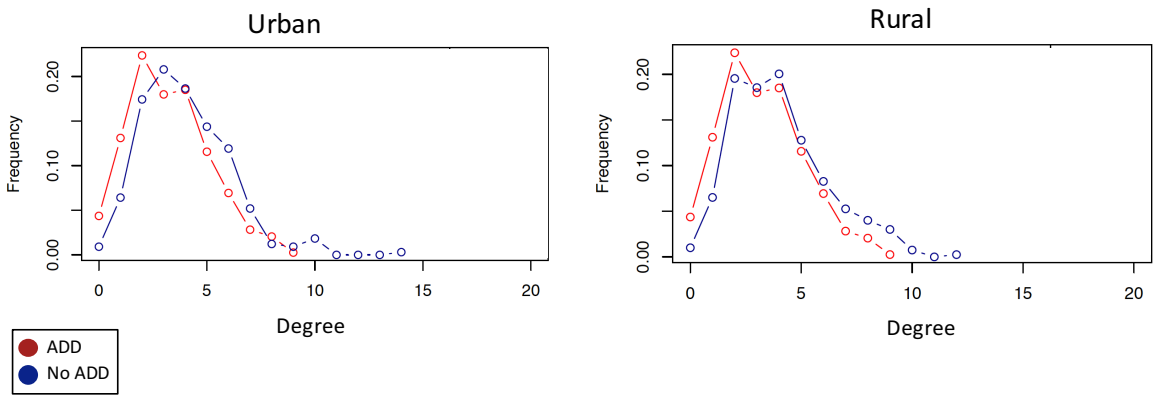


Figure B. 8. Degree distribution comparison between healthy and disease states from rural and urban OTU co-occurrence networks.

Degree distribution consists in the frequency of the number of connections of a node (degree) over the whole network. This parameter was calculated using igraph v1.2.1 (Csardi and Nepusz, 2006) package available in R v3.3.1.

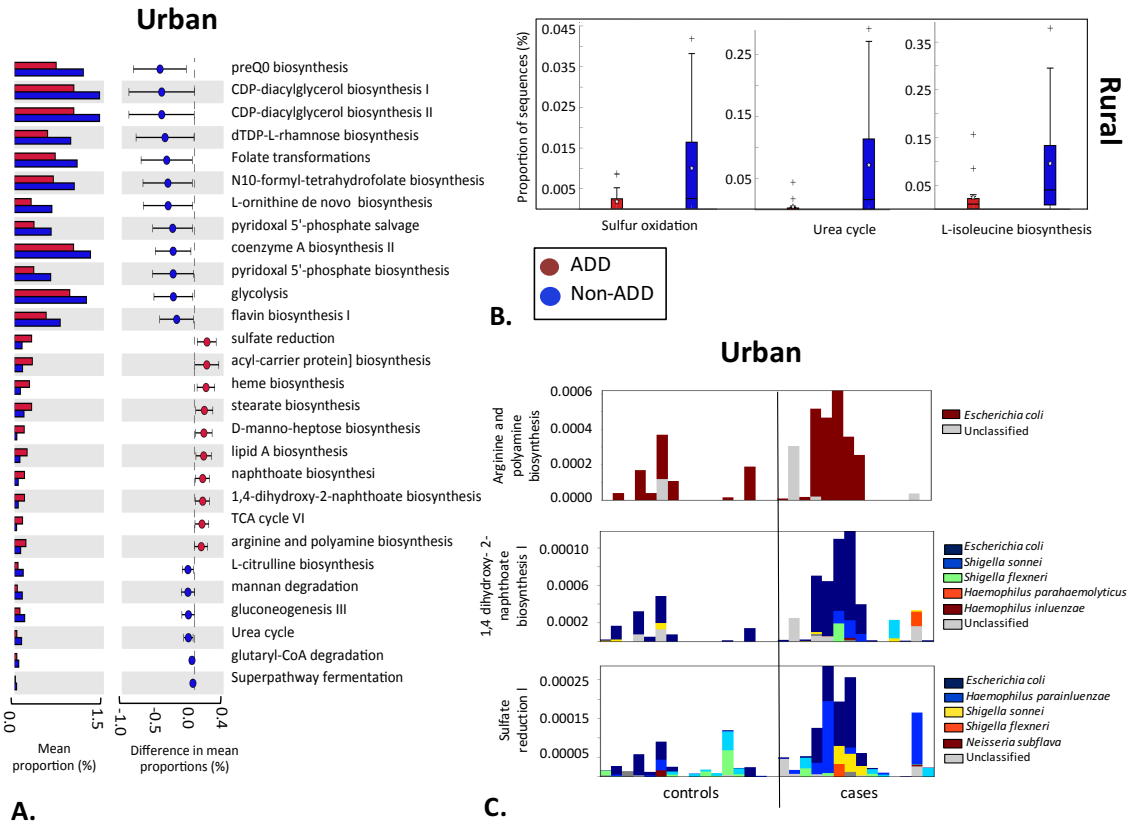


Figure B. 9. KEEG pathways differentially abundant during diarrheal episodes in metagenomes from rural and urban subjects.

Relative abundance of genes or pathways in ADD vs. non-ADD samples in urban metagenomes (panel A) and rural metagenomes (B panel). C. Taxonomic annotation of KEEG pathways that showed an increased abundance during ADD in urban samples.

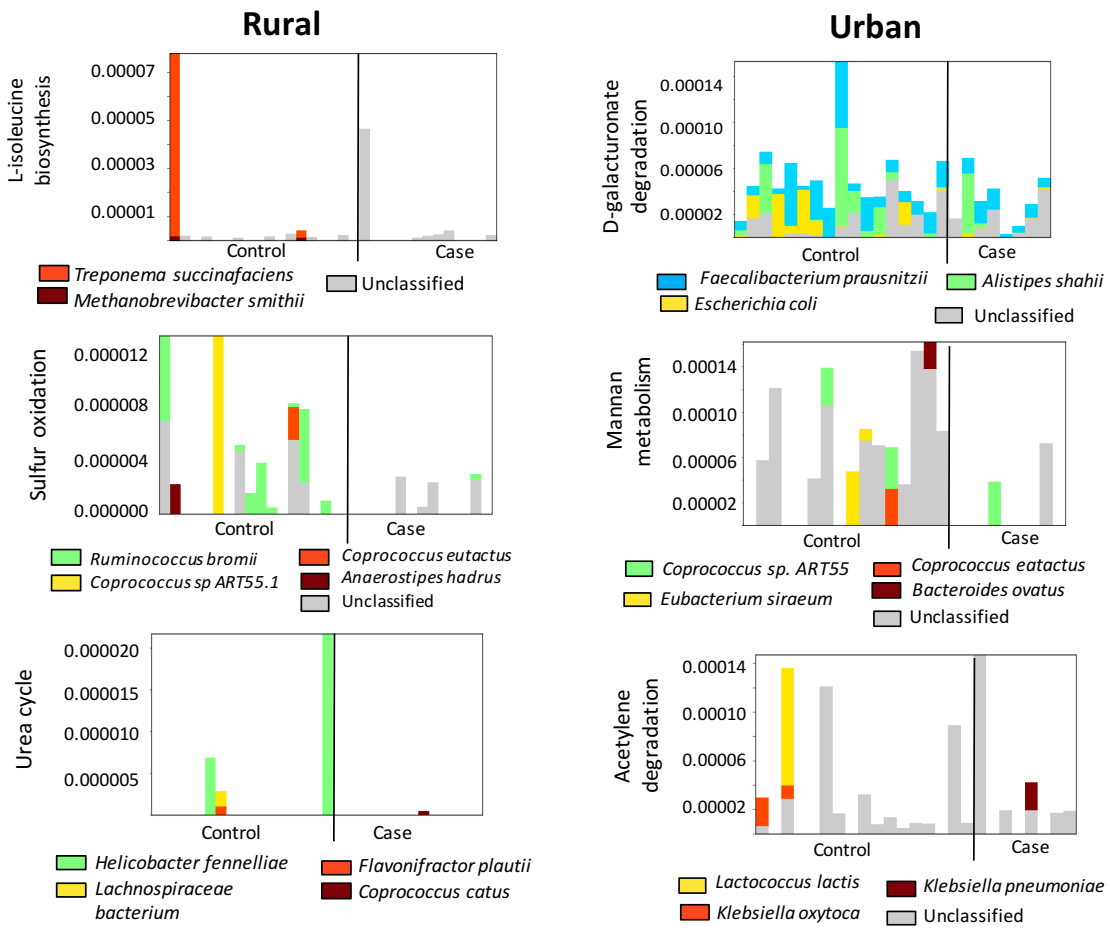


Figure B. 10. Taxonomic profile of KEEG pathways with significant shifts in abundance during ADD in selected metagenomes.

Each KEEG pathway includes the taxonomic profile of the microbial members that contributed to the relative abundance of the pathway in the community during ADD. Taxonomic annotations of KEEG pathways were extracted from the HUMAnN2 output file based on pre-defined clade-specific marker genes identified using MetaPhlAn2 (5).

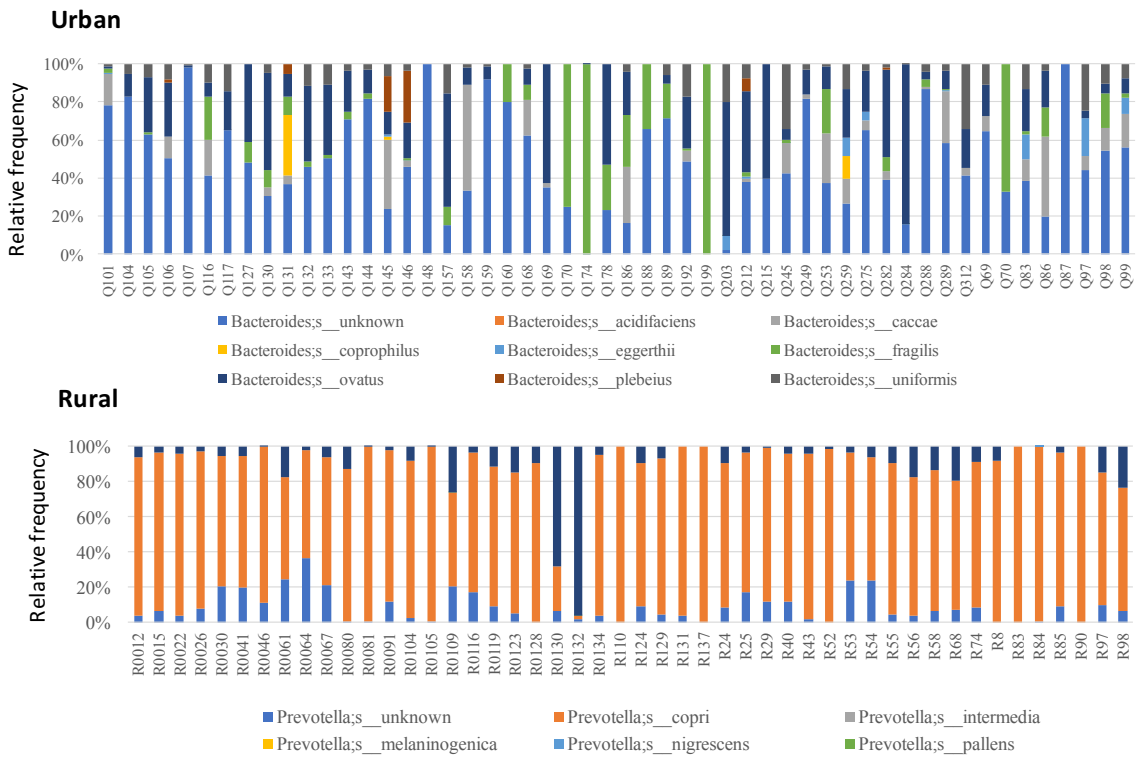


Figure B. 11. Distribution of 16S rRNA gene-based OTUs classified as *Bacteroides* and *Prevotella* among healthy samples from Quito and villages indicating that known and unknown species were identified in these two groups.

B.2. REFERENCES

1. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60.

2. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17:132.

3. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

4. Csardi C, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems:1695.

5. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12:902–903.

APPENDIX C. SUPPLEMENTARY MATERIAL CHAPTER 4

SUPPLEMENTARY TABLES AND FIGURES

Table C. 1. Akaike information criterion (AIC) values and number of parameters (K) for each equation model used to assess the relationship between genetic relatedness (*D*) and expected number of 100% identical gene (F_{100}). AIC was used to select the equation with the best fit of the data (lower AIC value).

Model	AIC	K
$1+\delta$	-6713.32	1
$1+\delta^2$	-6165.55	1
$1+\delta^3$	-5034.52	1
$1+\delta^4$	-3936.30	1
$1+\delta^5$	-3076.59	1
$1+\delta+\delta^2$	-6714.14	2
$1+\delta+\delta^3$	-6730.72	2
$1+\delta+\delta^4$	-6753.95	2
$1+\delta+\delta^5$	-6780.05	2
$1+\delta^2+\delta^3$	-6947.56	2
$1+\delta^2+\delta^4$	-7047.91	2
$1+\delta^2+\delta^5$	-7127.45	2
$1+\delta^3+\delta^4$	-7263.05	2
$1+\delta^3+\delta^5$	-7336.26	2
$1+\delta^4+\delta^5$	-7372.99	2
$1+\delta+\delta^2+\delta^3$	-7684.98	3
$1+\delta+\delta^2+\delta^4$	-7654.65	3
$1+\delta+\delta^2+\delta^5$	-7626.23	3
$1+\delta+\delta^3+\delta^4$	-7641.05	3
$1+\delta+\delta^3+\delta^5$	-7614.55	3
$1+\delta^2+\delta^3+\delta^5$	-7606.95	3
$1+\delta^2+\delta^3+\delta^4$	-7368.40	3
$1+\delta^2+\delta^3+\delta^5$	-7403.64	3
$1+\delta^3+\delta^4+\delta^5$	-7449.35	3
$1+\delta+\delta^2+\delta^3+\delta^4$	-7365.34	4
$1+\delta+\delta^2+\delta^3+\delta^5$	-7423.30	4
$1+\delta+\delta^2+\delta^4+\delta^5$	-7626.23	4
$1+\delta+\delta^3+\delta^4+\delta^5$	-7475.26	4
$1+\delta^2+\delta^3+\delta^4+\delta^5$	-7520.40	4
$1+\delta+\delta^2+\delta^3+\delta^4+\delta^5$	-7556.78	5

Table C. 2. List of bacterial taxa from the IMG database and previous studies used to estimate the degree of recent gene exchange.

Taxon	No. of genomes	Bacterial lifestyle
IMG database		
Fewer recent exchanges than the average		
Buchnera aphidicola	18	obligate intracellular bacterium
Salmonella_enterica	578	facultative anaerobe
Rickettsia rickettsii	8	obligate intracellular bacterium
Rickettsia prowazekii	11	obligate intracellular bacterium
Yersinia pestis	101	obligate pathogen
Mycobacterium tuberculosis	844	facultative intracellular pathogen
Lactobacillus casei	28	facultative anaerobe
Higher recent exchanges than the average		
Campylobacter jejuni	106	facultative opportunitic pathogen
Campylobacter coli	52	facultative opportunitic pathogen
Neisseria meningitidis	189	facultative opportunitic pathogen
Helicobacter pylori	324	facultative host-associated
Vibrio cholerae	180	free-living, opportunistic pathogen
Alteromonas macleodii	13	free-living, marine
Synechococcus sp	46	free-living, marine
Enterococcus faecalis	366	facultative opportunitic pathogen
Enterococcus faecium	263	facultative opportunitic pathogen

Table C.2. Continued

Previous studies			**Reference**
Buchnera aphidicola	4	obligate intracellular bacterium	(13)
Salmonella_enterica	8	facultative anaerobe	(14)
Klebsiella_pneumoniae	11	free-living, opportunistic pathogen	(15)
Shewanella baltica	3	free-living, marine	(16)
Staphylococcus aureus	7	free-living, opportunistic pathogen	(17)
Neisseria meningitidis	13	facultative opportunitic pathogen	(18)
Helicobacter pylori	62	facultative opportunitic pathogen	(19)
Campylobacter jejuni	65	facultative opportunitic pathogen	(6)
Escherichia coli	25	facultative opportunitic pathogen	(20)

Table C. 3. Results of genome comparisons of *H. pylori* strains isolated from the same and different individuals.

The table includes genome pairs with and without signatures of recent exchange as well as the estimated number of exchanged genes for those genome pairs with positive values of sigma. Genomes were reported in (19).

Genome pairs under recent gene exchange							
Same family	**ANI**	**Sigma**	**No. of exchanged genes**	**Within-host**	**ANI**	**Sigma**	**No. of exchanged genes**
SA161A-SA216A	98.92	6.77	492	SA163C-SA163A	99.83	0.72	77
SA227C-SA301C	98.72	5.49	397	SA210C-SA210A	99.87	0.31	36
SA162C-SA227C	98.46	4.47	311	SA161A-SA161C	99.83	2.84	316
SA162A-SA301A	98.31	4.37	303	SA162A-SA162C	99.66	5.46	506
SA210C-SA300C	99.88	1.75	201	SA146A-SA146C	99.82	2.36	258
SA210A-SA163A	99.74	2.27	238	SA227A-SA227C	99.80	1.86	198
SA210A-SA163C	99.83	0.94	101	SA45A-SA45C	99.88	1.05	115
SA158A-SA210C	99.87	1.32	150	SA144A-SA144C	99.85	0.58	65
Genome pairs without recent gene exchange							
Different family	**ANI**	**Sigma**		**Within-host**	**ANI**	**Sigma**	
SA220A-SA301A	96.22	-0.34		SA160A-SA160C	99.96	-0.53	
SA220A-SA30C	96.01	-0.57		SA300A-SA300C	99.96	-1.87	
SA233A-SA251C	97.35	-0.91		SA156A-SA156C	99.94	-0.01	
SA233A-SA47A	97.32	-0.85					
SA40A-SA47C	97.3	-0.15					
SA35C-SA45A	97.52	-0.10					

Table C. 4. P-values of multiple non-parametric goodness-of-fit tests used to test the spatial randomness in the genome of recent exchanges across the genome. KS refers to Kolmogorov-Smirnov.

Bacterial Sp	Moran's test	Cramér-von Mises test	Watson test	KS
C. jejuni				
110-117	0	0.06	3.79E-04	0.003
63-117	0	0.025	0	0.27
C. coli				
2516143061-2563366571	0.004	0.23	0.07	0.27
2516143087-2516143095	1.75E-09	0.115	0.002	0.032
N. meningitidis				
2537562117	0.039	0.18	0.071	0.15
2531839670	0.46	0.37	0.18	0.43
2547132294	0.005	0.49	0.24	0.32

Table C. 5. Running time of the Bayesian model and scripts to calculate the fraction of recently exchanged genes in bacterial genome pairs.

The running time was measured in datasets of different bacterial species with variable number of genomes in one CPU. OGs: Orthologous Groups of genes.

Dataset	No. of genomes	Estimation of the fraction under recent exchange		Estimation of OGs
		ANI (one pair)	Math model (group of genomes)	
C. jejuni	64	0m25.74s	0.026s	129.80m
C. coli	50	0m20.33s	0.048s	81.93m
N. meningitidis	189	1m8.0s	1.544s	1743.4m
B. aphidicola	18	0m9.24s	0.014s	4.6m
H. pylori	325	0m32.34s	20.74s	3858.3m

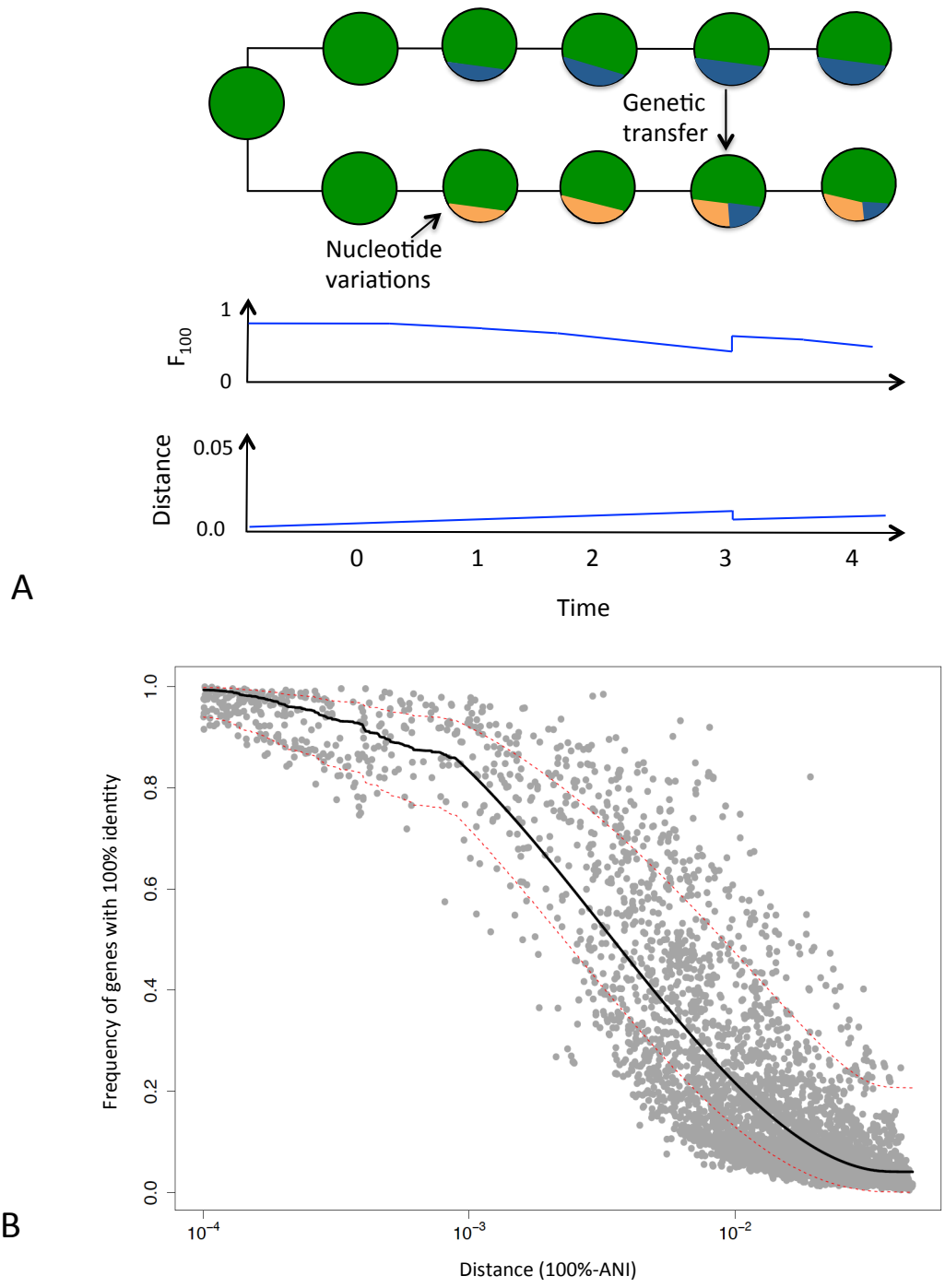


Figure C. 1. Schematic representation of the relation between genomic distance *D* and F_{100}.

A. Hypothetical scenario of genomic divergence associated with gene transfer over evolutionary time. Genomic distance is measured as ANI and observed F_{100}. Each step corresponds to a different period of time where a bacterial population (green circle) accumulates genetic nucleotide variations exemplified as blue and orange fragments inside the circle. Time 0 represents a clonal ancestral population; during time 1 and 2 minor genetic variation events (point mutations) accumulate, linked to a slight increase of genetic distance and a concomitant decrease of F_{100}. In time 3, genetic transfer events in a genome pair are illustrated, where F_{100} increases sharply and the nucleotide difference is slightly reduced since exchanged genes will have 100% nucleotide identity. Lastly, over time (time 4) additional mutations accumulate, the nucleotide composition of transferred genes become more similar to the one of the recipient genome, and F_{100} is reduced. **B.** F_{100} as a function of the distance *D*. Points represent pair of genomes with different genetic relatedness and their corresponding F_{100}. Expected value of F_{100} is represented as a solid black line and 95% credible intervals as dashed red lines.

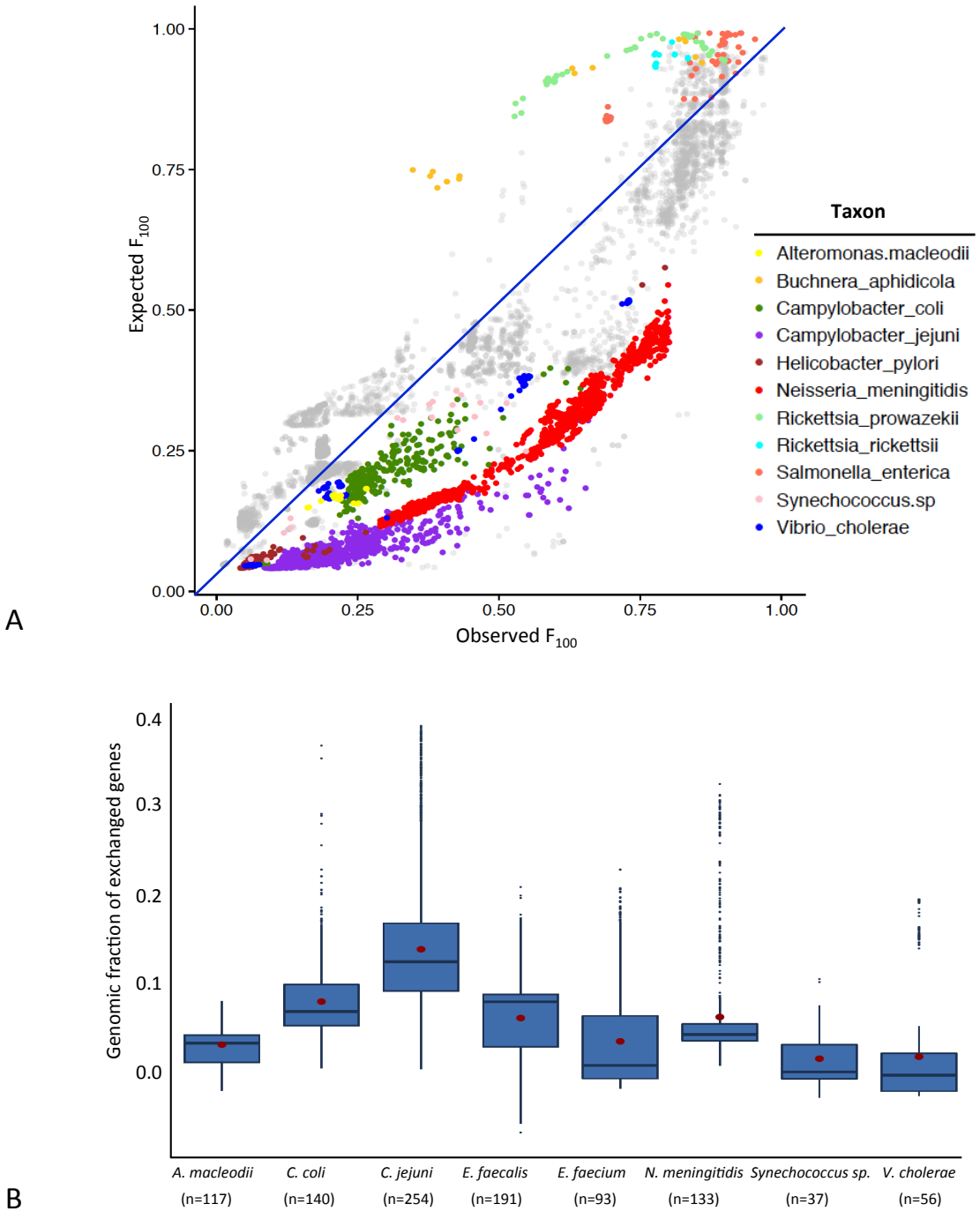


Figure C. 2. Signatures of recent exchange among genome pairs.

A. Distribution of observed (x-axis) vs. expected F_{100} (y-axis) of genome pairs from multiple bacterial species. Every point represents a genome pair that belongs to a bacterial species (color key). Grey points denote pairwise comparisons among remaining bacterial species that were not highlighted in color. The straight blue line indicates equal values for

the observed and expected F_{100}. B. Boxplots showing the distribution of the recently exchanged genes as a fraction of the total genomes (y axis) fractions for selected recombinogenic bacterial species. Red points indicate the average fraction for each species and the number in parentheses denotes the actual number of genes exchanged for each bacterial species.

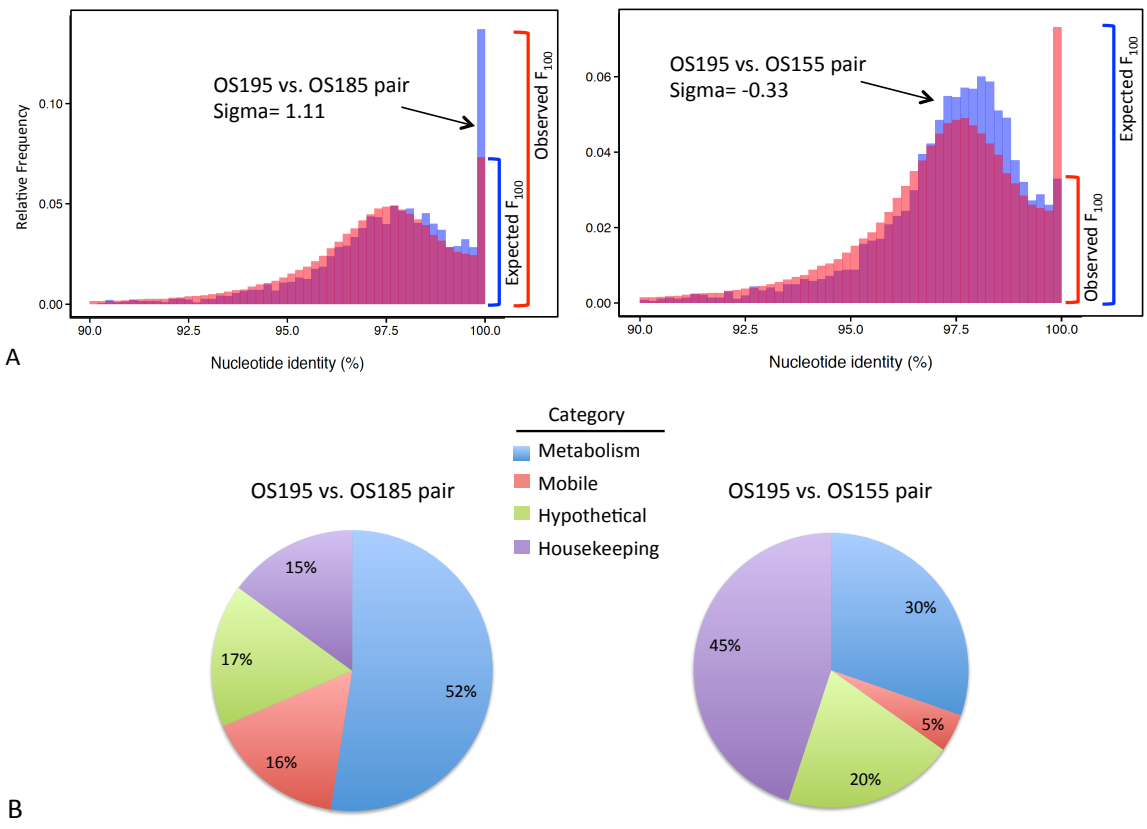


Figure C. 3. Comparison between *S. baltica* genome pairs with high (OS195 vs. OS185) and low (OS195 vs. OS155) frequency of recent genetic exchange. A. Each plot illustrates the relative frequency of Reciprocal Best Matches between the pair of genomes (RBMs) against their nucleotide identity for each genome pair (red) vs. the reference distribution from bacterial genomes pairs with a similar ANI value (~96.6%) (purple). Note the difference in scale on the y-axes. **B.** Functional annotation of genes with 100% nucleotide identity of each genome pair. Orthologous genes were annotated using UniProt database and grouped in four main categories: housekeeping, mobile, hypothetical and, metabolism. Pie charts represent the relative frequency of genes of each category.

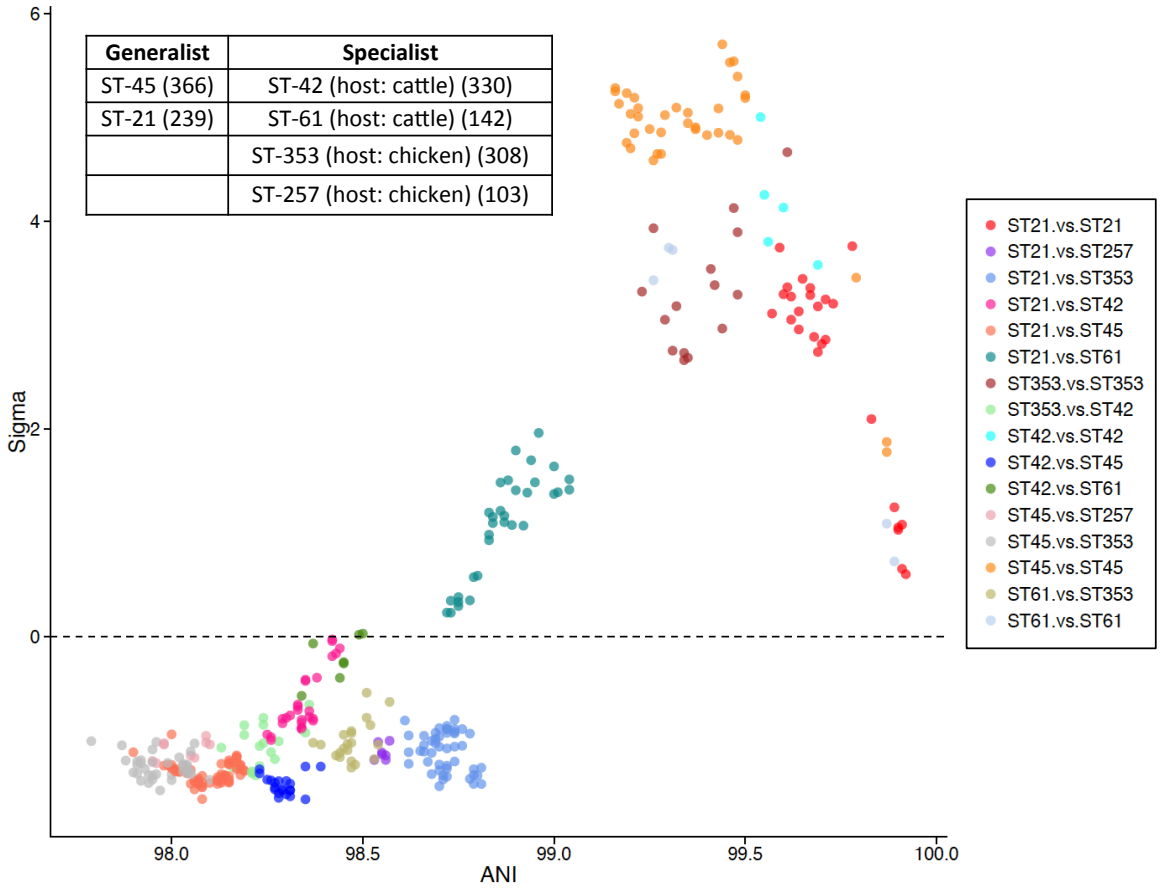


Figure C. 4. Comparison of recent exchanges between specialist and generalist *C. jejuni* strains from different host sources.

Every point represents a pair of genomes from the same or different clonal complex (see color chart). The table of the top left specifies whether the clonal complex is generalist (strains able to colonize multiple host species) or specialist (strains mostly restricted to a single host species and a specific source) according to (6, 7). The dotted line indicates equal values of the observed and expected F_{100}. Genome sequences of the *C. jejuni* strains were previously published in (6).

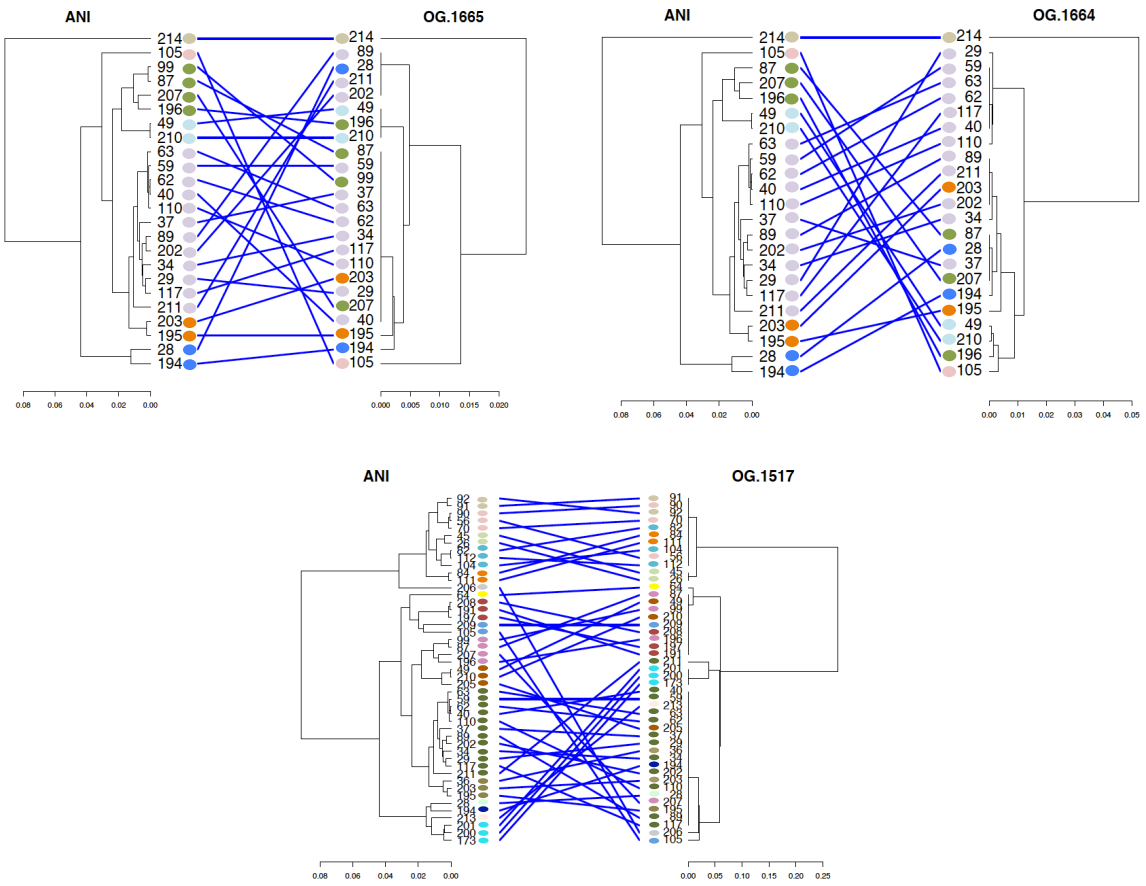


Figure C. 5. Examples of incongruence in tree topology in exchanged genes identified in the *C. jejuni* genome pair 63-117.

Tanglegrams comparing trees based on ANI distances (left) and maximum likelihood phylogenies (right) of three genes detected as recently exchanged from the variable genome. Crossing lines indicate recombination events. Small circles are colored by the cluster that each strain was previously assigned using PAM based on an ANI matrix (8).

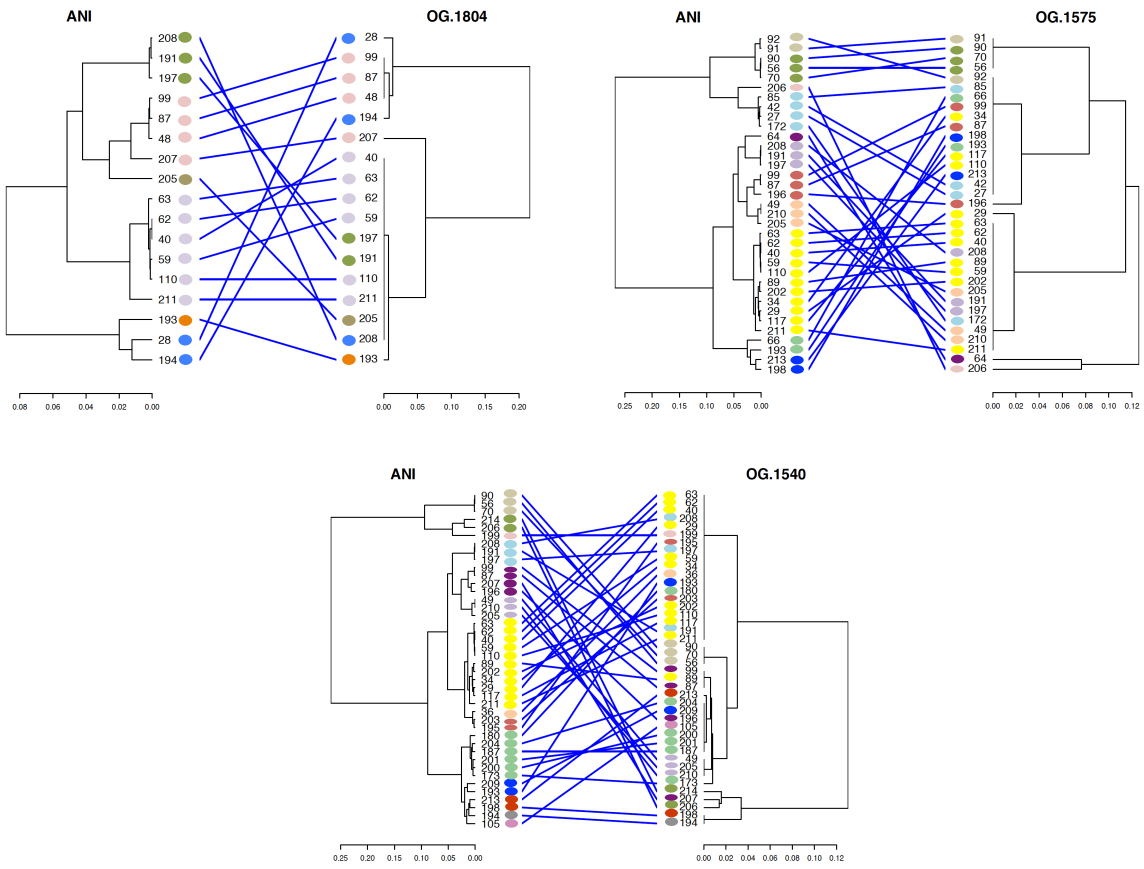


Figure C. 6. Examples of incongruence in tree topology in exchanged genes identified in the *C. jejuni* genome pair 62-191.

Tanglegrams comparing trees based on ANI distances (left) and maximum likelihood phylogenies (right) of three genes detected as recently exchanged from the variable genome. Crossing lines indicate recombination events. The corresponding cluster previously assigned to each strain, using PAM based on an ANI matrix, is indicated by the color of the small circle.

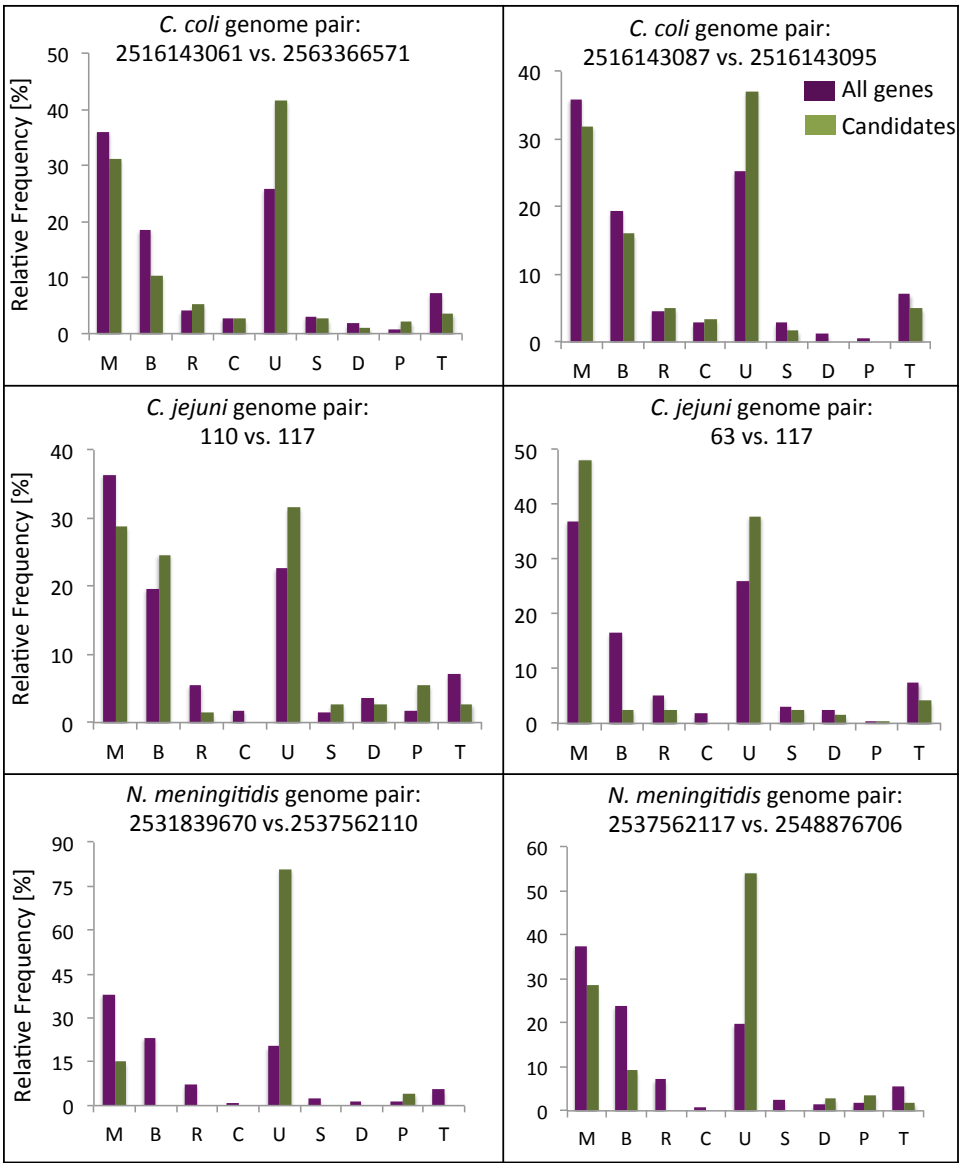
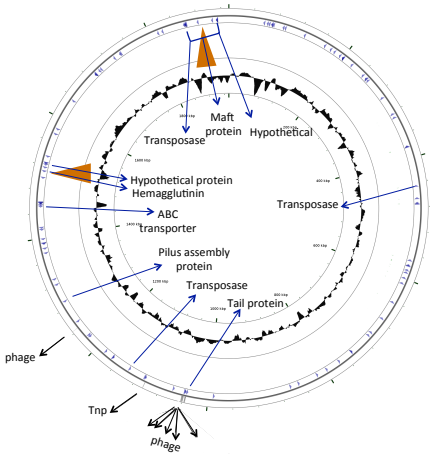


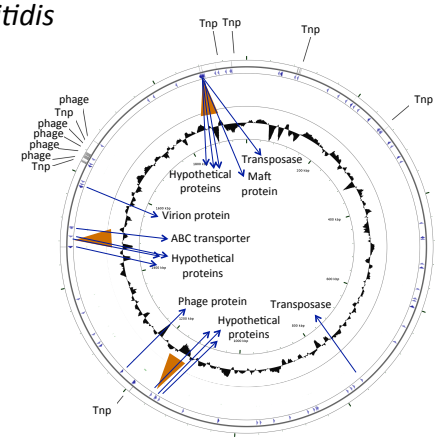
Figure C. 7. Functional comparison between all predicted genes in the genome (purple bars) and genes that have undergone recent exchange (green bars).

Each panel shows the percentage of the total genes in the genome (y-axis) assigned to each functional category (x-axis) for a pair of genomes pair (title), after excluding genes with high sequence conservation (100% nucleotide identity). Functional categories are based on EggNOG annotations (9) as follow: M (Metabolism), B (Cell cycle and Biogenesis), R (Recombination and Repair), C (Cell motility), U (Unknown function), S (Secretion and Transport), D (Defense mechanisms), P (Mobilome), and T (Signal transduction)

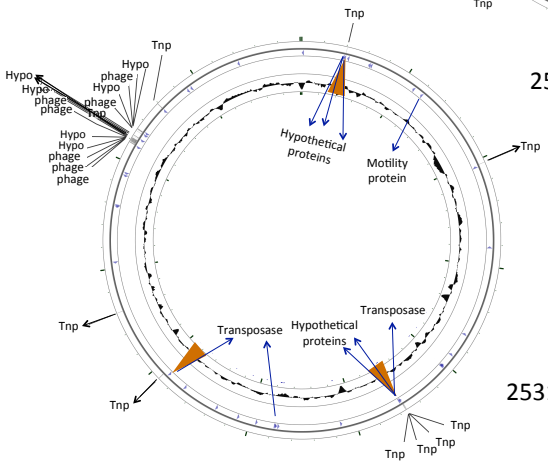
Species: *N. meningitidis*



Genome pair
2537562117-2548876706
(n=111)

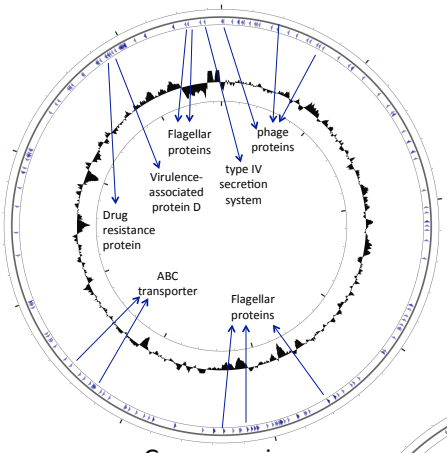


Genome pair
2547132294-646862337
(n=103)

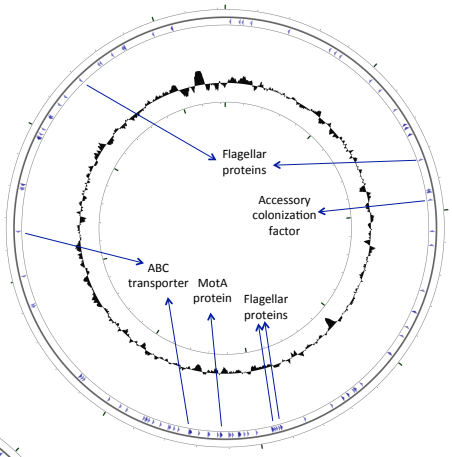


Genome pair
2531839670-2537562110
(n=47)

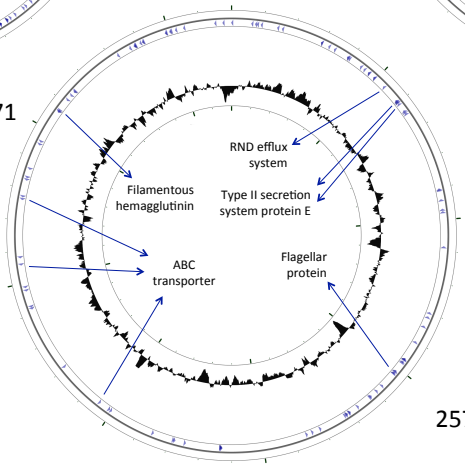
Species: *C. coli*



Genome pair
2516143061-2563366571
(n=193)



Genome pair
2516143087-2516143095
(n=119)



Genome pair
2576861668-638341044
(n=339)

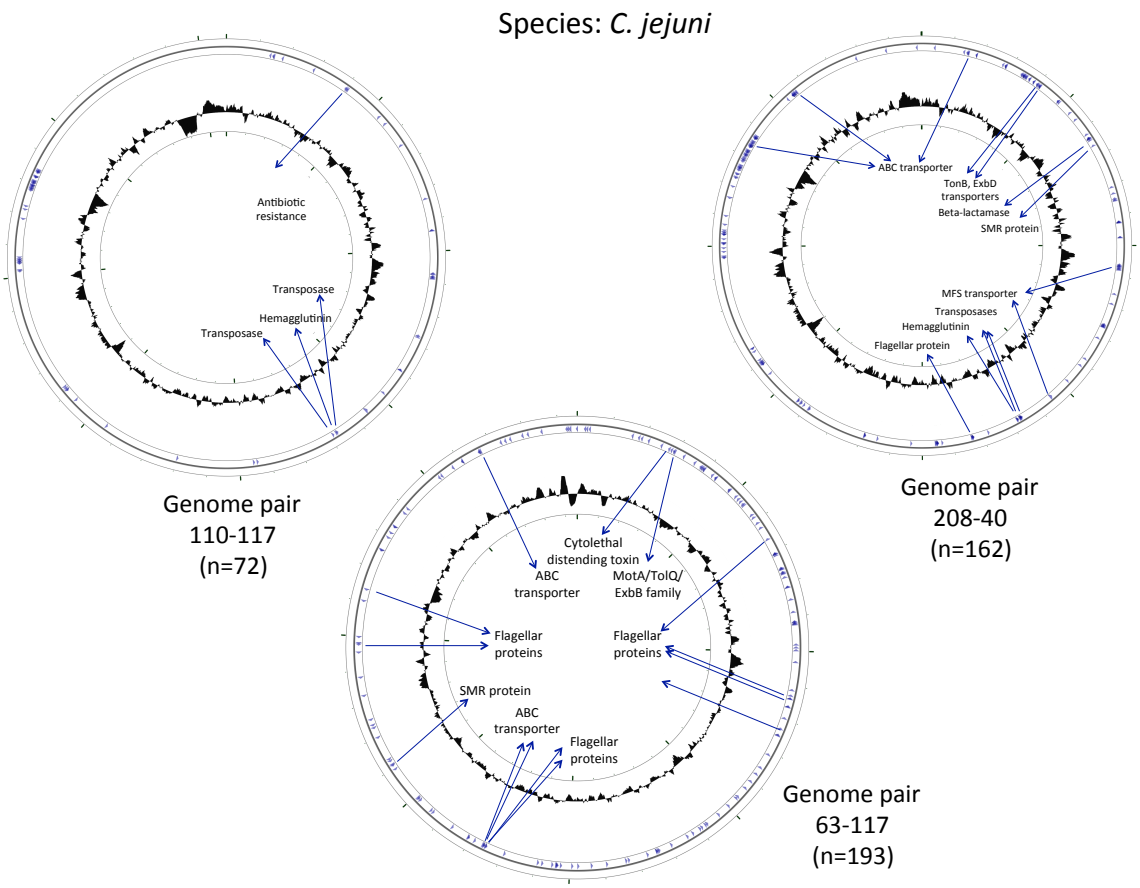


Figure C. 8. Circular plots indicating the position of the recently exchanged genes (blue arrowhead) of selected genome pairs.

Three examples per bacterial species are shown for the most recombinogenic species, i.e., *C. coli*, *C. jejuni*, and *N. meningitidis*. Conserved regions (genes with 100% nucleotide identity without signatures of gene exchange) were removed prior to drawing the plots. The number of recent exchanges for each genome pair is indicated in parentheses. The genes around exchanged genes in a window of ± 5 genes (upstream and downstream regions) are denoted with blue arrowheads within the outermost circle and those genes annotated as hypothetical proteins (Hypo), transposases (Tnp), and phages proteins (phage) are highlighted with black arrows. Pathogenicity islands are represented by brown triangles and were identified using PAIDB v2 (10) and IslandViewer (11). Plots were drawn using CGview (12).

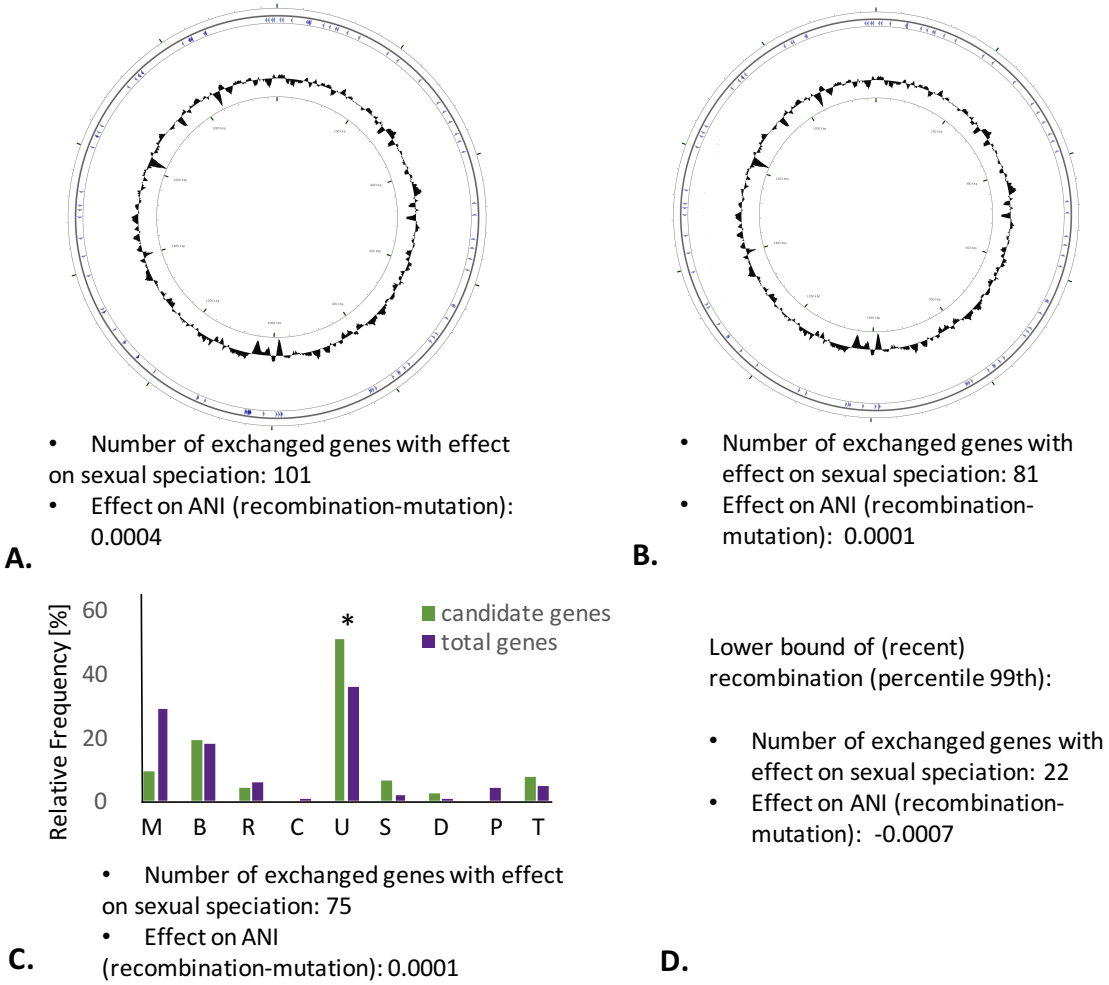


Figure C. 9. Effect of mutation and recombination on ANI in the genome pair *N. menigitidis* 2523533512-2534681689 at four different stringency scenarios. Four different scenarios for the effect of recombination and mutation on ANI were evaluated: All exchanged genes found to be recently exchanged in the genome (A); functional (B) and spatial (C) biases of the exchanged genes removed; and the distance between the most spatially distant exchanged genes in the genome (percentile 99th) was used to estimate the effect of recombination (D). Recently exchanged genes are denoted with blue arrowheads within the outermost circle in panels A and B. *Hypergeometric test, $P < 0.05$. See Materials and Methods for details.

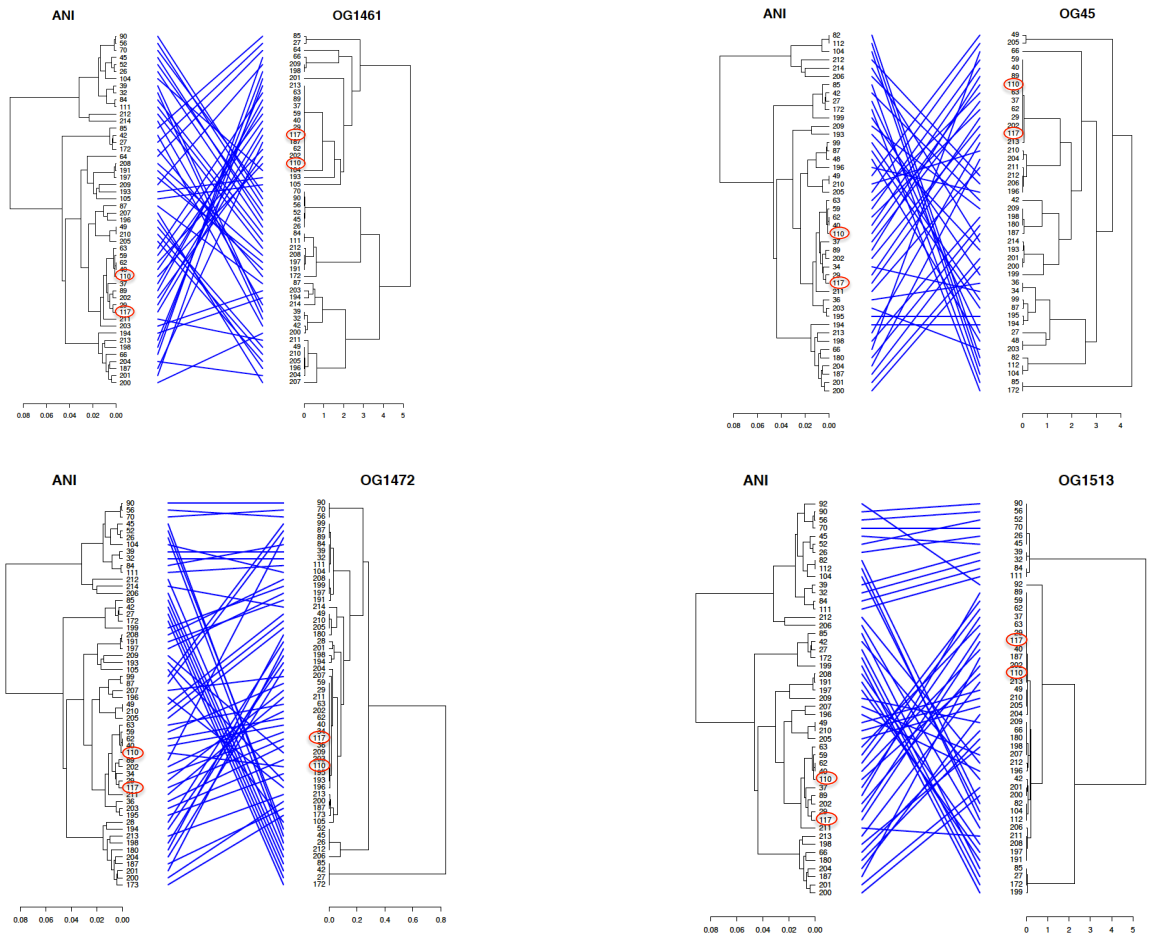


Figure C. 10. Examples of tree topology discrepancies in exchanged genes identified in the *C. jejuni* genome pair 110-117 by our method.

All recent exchanges in this genome pair were subjected to a manual inspection of their tree topology and four representative cases of detected recent gene exchanges are shown. Tanglegrams represent the genomic tree based on ANI distances (left) and the gene tree constructed with the maximum likelihood algorithm (right). Crossing lines indicate recombination events. Red small circles highlight the location of the query genomes.

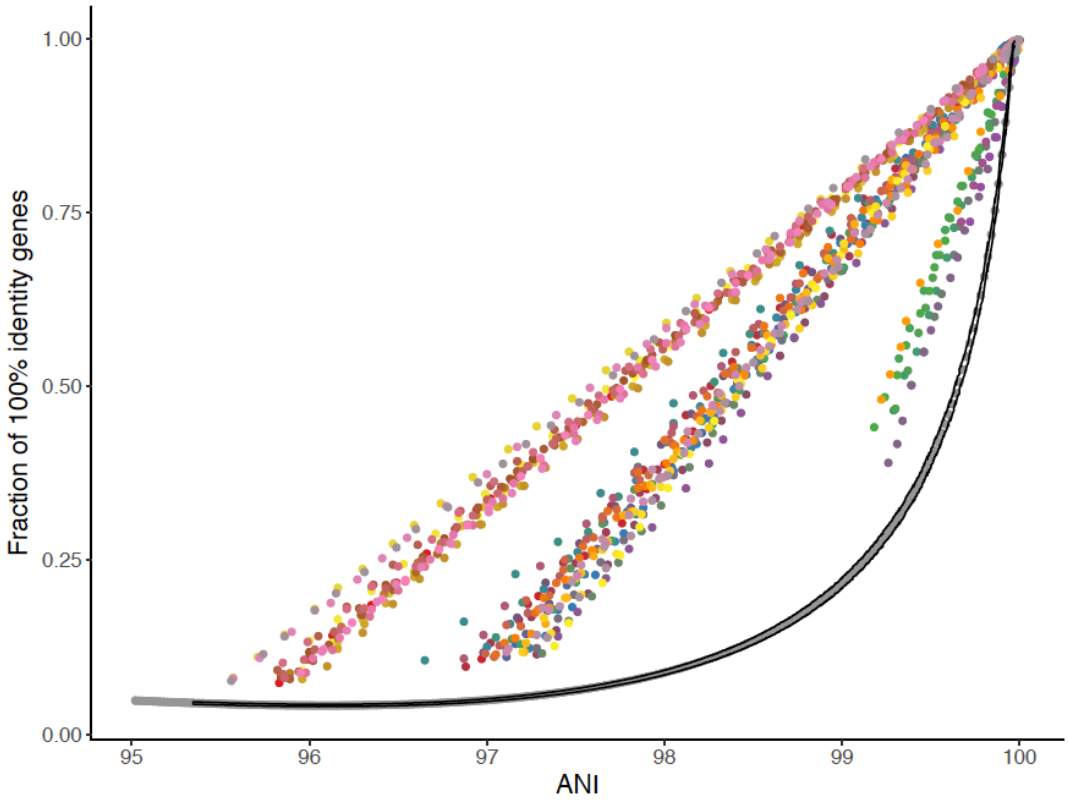


Figure C. 11 Assessing the ability to detect outliers as an effect of the frequency of 100% nucleotide sequence identity genes.

The fraction of 100% nucleotide sequence identity genes (F_{100}) of 64 genome pairs with ANI values around 96%, 97%, and 99% identified as outliers by our model was gradually increased by randomly introducing nucleotide changes in genes that were not originally 100% identical in order for these genes to become 100% identical (y-axis). Dots represent the observed F_{100} of the genome pair (y-axis) against the estimated (new) ANI value of the two genome (x-axis), and are colored by their genome pair. The average expected value of F_{100} is represented as grey dots and 95% credible intervals as solid black lines. Note that the genome pairs identified originally as outliers in the lower left part of the graph continue being detected as outliers in the right part of the graph, even when the fraction of 100% nucleotide identity genes increases up to about 99.8% ANI.

C.2. REFERENCES

1. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397.

2. Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91.

3. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122.

4. Rodriguez-R LM. MIGA: Microbial Genome Atlas. MIGA.

5. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. e1900v1. PeerJ Preprints.

6. Sheppard SK, Cheng L, Méric G, de Haan CPA, Llarena A-K, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, Strachan NJC, Forbes K, Colles FM, Jolley KA, Bentley SD, Maiden MCJ, Hänninen M-L, Parkhill J, Hanage WP, Corander J. 2014. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. Mol Ecol 23:2442–2451.

7. Gripp E, Hlahla D, Didelot X, Kops F, Maurischat S, Tedin K, Alter T, Ellerbroek L, Schreiber K, Schomburg D, Janssen T, Bartholomäus P, Hofreuter D, Woltemate S, Uhr M, Brenneke B, Grüning P, Gerlach G, Wieler L, Suerbaum S, Josenhans C. 2011. Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. BMC Genomics 12.

8. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. 2018. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res 46:W282–W288.

9. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293.

10. Yoon SH, Park Y-K, Kim JF. 2015. PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic Acids Res 43:D624–D630.

11. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, Brinkman FS. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res 45:W30–W35.

12. Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21:537–539.

13. Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol 2:RESEARCH0054.

14. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G. 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet 40:987–993.

15. Wyres KL, Gorrie C, Edwards DJ, Wertheim HFL, Hsu LY, Van Kinh N, Zadoks R, Baker S, Holt KE. 2015. Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the *Klebsiella pneumoniae* Clonal Group 258. Genome Biol Evol 7:1267–1279.

16. Caro-Quintero A, Auchtung J, Deng J, Brettar I, Hofle M, Tiedje JM, Konstantinidis KT. 2012. Genome Sequencing of Five *Shewanella baltica* Strains Recovered from the Oxic-Anoxic Interface of the Baltic Sea. J Bacteriol 194:1236–1236.

17. Driebe EM, Sahl JW, Roe C, Bowers JR, Schupp JM, Gillece JD, Kelley E, Price LB, Pearson TR, Hepp CM, Brzoska PM, Cummings CA, Furtado MR, Andersen PS, Stegger M, Engelthaler DM, Keim PS. 2015. Using Whole Genome Analysis to Examine Recombination across Diverse Sequence Types of *Staphylococcus aureus*. PLOS ONE 10:e0130955.

18. Yu D, Jin Y, Yin Z, Ren H, Zhou W, Liang L, Yue J. 2014. A Genome-Wide Identification of Genes Undergoing Recombination and Positive Selection in *Neisseria*. BioMed Res Int 2014:1–9.

19. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. 2013. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. Proc Natl Acad Sci 110:13880–13885.

20. Didelot X, Méric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256.