HABITAT ADAPTATION AND GENOME EVOLUTION IN THE GUT MICROBIOME

by

JESSE ROBERT REBOA ZANEVELD

B.S., University of Oregon, 2005

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Molecular, Cellular,

and Developmental Biology

2011

This thesis entitled:

HABITAT ADAPTATION AND GENOME EVOLUTION IN THE GUT MICROBIOME

written by

Jesse Robert Reboa Zaneveld

has been approved for the
Department of Molecular, Cellular, and Developmental Biology

_____
Ken Krauter

_____
Robin Dowell

Date_____

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.

Zaneveld, Jesse Robert Reboa (Ph.D., MCDB)
Habitat Adaptation and Genome Evolution in the Gut Microbiome

Thesis directed by Professor Rob Knight

ABSTRACT

We live in a world suffused with microbial life. Universal trees of life show that microbial bacteria, archaea, and eukaryotes constitute the vast majority of life's diversity. These diverse organisms perform many important ecological functions across a wide range of natural and man-made environments: photosynthesis in the world's oceans; nitrogen fixation and provision of carbohydrates in association with plant roots; and modification of the chemistry of the upper atmosphere by microbial communities in droplets of cloud-water. The bodies of animals are also colonized internally and externally by microorganisms, which play crucial roles in the development, homeostasis, and even behavior of their hosts. How have microbial bacteria, archaea, and eukaryotes adapted to survive and thrive across such a range of lifestyles and habitats?

I addressed one aspect of this question by using the bacteria inhabiting the

mammalian gut as a model for exploring how habitat adaptation impacts the evolution of microbial genomes. I characterized the relationship between 16S rRNA gene sequence similarity and overall levels of gene conservation in the genomes of four groups of species: gut specialists and cosmopolitans, each of which can be divided into pathogens and non-pathogens. At short phylogenetic distances, specialist or cosmopolitan bacteria found in the gut share fewer genes than is typical for genomes that come from non-gut environments, but at longer phylogenetic distances gut bacteria are more similar to each other than are genomes at equivalent evolutionary distances from non-gut environments, suggesting a pattern of short-term specialization but long-term convergence. Moreover, this pattern is observed in both pathogens and non-pathogens, and can even be seen in the plasmids carried by gut bacteria. This observation is consistent with the finding that, despite considerable interpersonal variation in species content, there is surprising functional convergence in the microbiome of different humans. Finally, I observed that even within bacterial species or genera 16S rRNA divergence provides useful information about average conservation of gene content . The results described here should be useful for guiding strain selection to maximize novel gene discovery in large-scale genome sequencing projects, while the approach could be applied in studies seeking to understand the effects of habitat adaptation on genome evolution across other body habitats or environment types.

DEDICATION

To my family, Zanevelds and Reboas, for your love and support.

ACKNOWLEDGEMENTS

I would like to thank my advisor Rob Knight for his advice and encouragement over the years; my collaborators and lab mates for their hard work, bright insights, and valuable suggestions; my classmates in MCDB for their companionship and commiseration; my thesis committee for their help guiding the direction of the research presented here; and my wife Micaela for her patience, love, and wisdom. In addition, I would like to make the following specific acknowledgements for the chapters presented in this thesis:

Chapter I

Is adapted from a multi-author review article, on which I was first author published in *Current Opinion in Chemical Biology*[1]. I would like to thank my co-authors: Pete Turnbaugh, Cathy Lozupone, Ruth Ley, Micah Hamady, Jeff Gordon and Rob Knight. This work was supported by NIH grant P01DK078669, the NIH/CU Molecular Biophysics Training Program T32GM065103, and the NIH/CU Signaling and Cell Regulation Training Program T32 GM08759.

Chapter II

Is adapted from a first-authored research paper that I wrote, published in *Nucleic Acids Research* [2] . I would like to thank my co-authors Cathy Lozupone, Jeff Gordon, and Rob Knight for their valuable assistance. I would like to thank Justin Kuczynski, Elizabeth Costello, Tony Walters, Daniel McDonald and Sara Nakielny for helpful comments on the manuscript. My classmates in "Genome Databases: Mining and Management", MCDB 5621, where this analysis was initiated as a class project, also provided valuable insight and support. This work was supported by National Institutes of Health pre-doctoral training (grant T32 GM08759 to J.Z.); National Institutes of Health (grant numbers P01DK078669, R01HG004872); the Crohn's and Colitis Foundation of America and Howard Hughes Medical Institute (HHMI).

Chapter III

Is adapted from an analysis of horizontal gene transfer that I performed in support of a broader effort to characterize genetic variation (the 'pangenome') in *Methanobrevibacter smithii*, an important gut archaon. The full analysis, on which Liz Hansen was first author, has been published in *PNAS* [3]. I would like

to thank Daniel McDonald and Julia Goodrich (also authors on the *PNAS* paper) for valuable assistance in running the analysis. This work was supported by NIH grant P01DK078669, the NIH/CU Molecular Biophysics Training Program T32GM065103, and the NIH/CU Signaling and Cell Regulation Training Program T32 GM08759. Portions of the introductory material on horizontal gene transfer are also drawn from a first-authored review that has been published in *Microbiology* [4]. I would like to thank my co-authors Diana Nemergut and Rob Knight for their help in seeing that project through to publication.

Chapter IV

Is adapted from a manuscript that is in revision at *Current Opinion in Microbiology* following favorable initial reviews. This was a multi-author manuscript on which I am first author. I would like to thank my co-authors: Laura Wegener Parfrey, Will Van Treuren, Catherine Lozupone, Jose C. Clemente, Dan Knights, Jesse Stombaugh, Justin Kuczynski, and Rob Knight. I would like to thank Mike Robeson for useful comments on the draft. The work from our laboratory described in this review was supported in part by the National Institutes of Health, the Crohns and Colitis Foundation of America, the Bill and Melinda Gates Foundation, and the Howard Hughes Medical Institute.

CONTENTS

CHAPTER

I. INTRODUCTION

The concepts of coevolution and codifferentiation....4

Identifying genes critical for symbiosis....7

Horizontal gene transfer and coevolution....9

The promise of metagenomic approaches....11

Combined phylogenetic and genomic approaches....14

Overview of the Thesis....15

II. EFFECTS OF ENVIRONMENTAL ADAPTATION ON GENOME CONTENT IN GUT-ADAPTED BACTERIA AND THEIR NON-GUT RELATIVES....18

Introduction....18

Methods....22

Selection and classification of genomes....22

Gene conservation....25

BLAST analysis....26

Tree construction....27

Results....28

A scale relates gene content to 16S rRNA distance....28

Habitat adaptation and genome size alter aggregate gene conservation 31

16S rRNA distance predicts genome diversity within bacterial species 41

Habitat adaptation in bacterial plasmids.....45

The effects of habitat adaptation on gene conservation occur in both pathogens and non-pathogens.....49

Discussion.....50

III. HORIZONTAL GENE TRANSFER AND GENOME EVOLUTION IN THE *METHANOBREVIBACTER SMITHII* PAN-GENOME 57

Background 57

M. smithii's role in the gut 58

Genes involved in *M. smithii* habitat adaptation.....59

Methods for HGT detection 60

Methods 63

Compositional analysis of HGT.....63

Phylogenetic analysis of HGT 64

Results and Discussion.....65

HGT has contributed to both the core and variable components of the *M. smithii* pangenome 69

Functional contribution of horizontally transferred genes to the *M. smithii* pangenome....69

Evidence for large-scale horizontal gene transfer of adhesin-like proteins (ALPs)....75

Conclusion....76

IV. HIGH-THROUGHPUT STUDIES OF MICROBIAL HABITAT ADAPTATION: PRINCIPLES AND PROGRESS....78

Setting the stage for high-throughput studies of habitat adaptation....78

High-throughput studies of habitat adaptation....83

Genome Reduction....83

Challenges in defining environment....85

Metadata annotation....88

Ordination methods....90

Application of machine-learning techniques....92

Phylogenetic comparative methods....93

Ancestral state reconstruction....97

Relating co-occurrence patterns to bacterial genomes....98

Horizontal gene transfer....100

Source/sink dynamics....102

Conclusion....103

V. CONCLUSION 105

REFERENCES.....110

TABLES

TABLE

Table 1. Compositional evidence of Horizontal Gene Transfer in the M. smithii pangenome.....68

Table 2. Distribution of HGT genes in the M. smithii core, variable, and pangenome by detection method.....70

Table 3. KEGG functional categories of genes with compositional evidence of horizontal gene transfer 74

Table 4. Compositional evidence for horizontal gene transfer of *M. smithii* ALP genes.....75

Table 5. Links to software and resources discussed in the text 81

FIGURES

FIGURE

Figure 1. Undestanding the microbial part of ourselves 3

Figure 2. Processes affecting host-microbe coevolution. 5

Figure 3. A comprehensive understanding of our microbial ecology requires integration of many data sources.....13

Figure 4. Classification of species by habitat and pathogenicity 25

Figure 5. Gene conservation by evolutionary distance 30

Figure 6. Gene conservation in gut-adapted bacteria 36

Figure 7. 16S rRNA percent identity vs. phylogenetic distance (Clearcut) 39

Figure 8. Regression of phylogenetic distance on 16S rRNA percent identity at short distances 40

Figure 9. Greater 16S rRNA divergence implies greater divergence in gene content within bacterial species. 49

Figure 10. Gene conservation in plasmids borne by gut-adapted bacteria 49

Figure 11. Gut pathogens, like gut commensals, exhibit different patterns of gene content from non-gut genomes 54

Figure 12. Compositional analysis of Horizontal Gene Transfer 67

Figure 13. Recurrent themes in the analysis of microbial habitat annotation 87

Figure 14. The importance of phylogenetic correction in comparing traits across habitats.....96

CHAPTER I

INTRODUCTION

The ~10^{14} microbes that live in and on each of our bodies belong to all three domains of life on earth — bacteria, archaea and eukarya. They outnumber our own cells by a factor of 10, and contribute many physiological capabilities, including metabolism of glycans and amino acids, synthesis of vitamins and isoprenoids, and biotransformation of xenobiotics [5]. A deeper understanding of our human biology thus requires understanding of our microbial communities and the genes that they harbor ('our' microbiome) (Figure 1**)**. The notion that we have a 'meta-genome' composed of microbial and human components, and a 'meta-metabolome' that reflects metabolic activities carried out by both our microbial and our *H. sapiens* cells, has implications for the definition of health, discernment of disease susceptibilities, and diagnosis of human pathologies. This view of ourselves also opens up another dimension to therapeutics, including treatment strategies that accommodate microbial metabolism of drugs that target our human cells, and a new generation of therapeutics that affect the structure and function of our indigenous microbial communities. The vast majority of our microbes live in the gut. Thus, the current challenge is to understand the extent to which each individual's gut microbiota affects the bioavailability and host/microbial responses to orally or

parenterally administered drugs, and the impact of interventions that alter our microbial ecology.

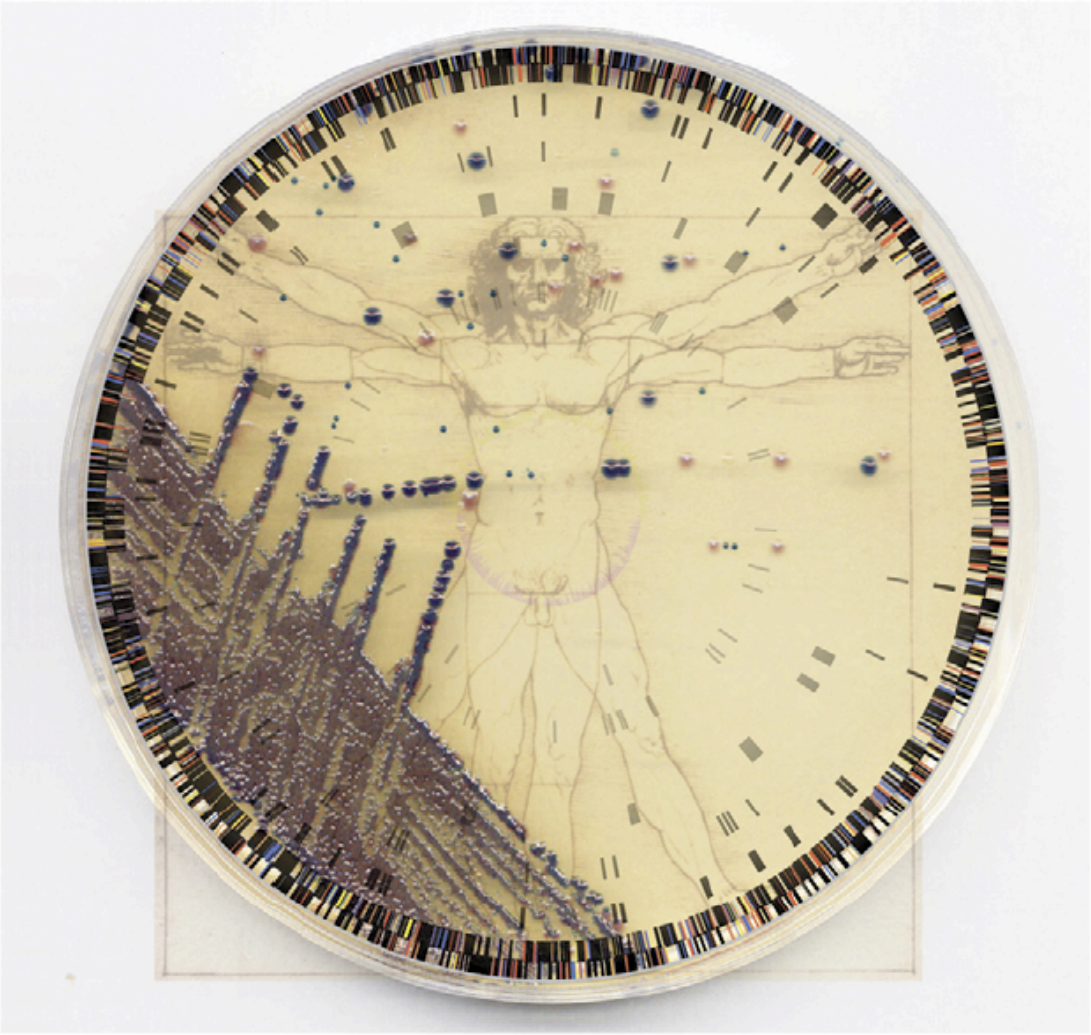


Figure 1. Understanding the microbial part of ourselves. A key part of understanding human metabolic capabilities is to understand our microbial symbionts, and the genomes of those symbionts.

The concepts of coevolution and co-differentiation

One key step in understanding our microbiota is to identify lineages that have coevolved with humans (or with mammals in general), and to identify the genomic consequences of this coevolution. Coevolution between a host and a beneficial symbiont, or a pathogen, is defined as reciprocal adaptation of each lineage in response to the other[6]. For example, genetic changes that increase production of a metabolite by an intestinal bacterium may trigger selection of changes in the host genome that promote uptake or prevent synthesis of that metabolite. Coevolution can also result in co-differentiation. Co-differentiation is defined as the diversification of host and symbiont lineages in parallel through a history of constant association; however, coevolution and co-differentiation can occur independently of one another [6]. Co-differentiation can be detected by showing that host and symbiont phylogenies match (see [7, 8] for detailed reviews of methods). Some methods for detecting co-differentiation can also generate hypotheses about processes causing specific differences between host and symbiont phylogenies. These differences (Figure 2) include (i) the absence of a symbiont in a host lineage, due to extinction of the microbial species in the microbiota that occupies a given body habitat in the host, or due to under-sampling; (ii) speciation of a symbiont within the same host, so that the host contains two closely related species of the symbiont; (iii) failure of the symbiont to speciate when the host speciates, so that two closely related species of the host contain the same symbiont; and (iv) host-switching, i.e. transfer of a symbiont to a different host[7]. For example, in the case of (iv), unrelated

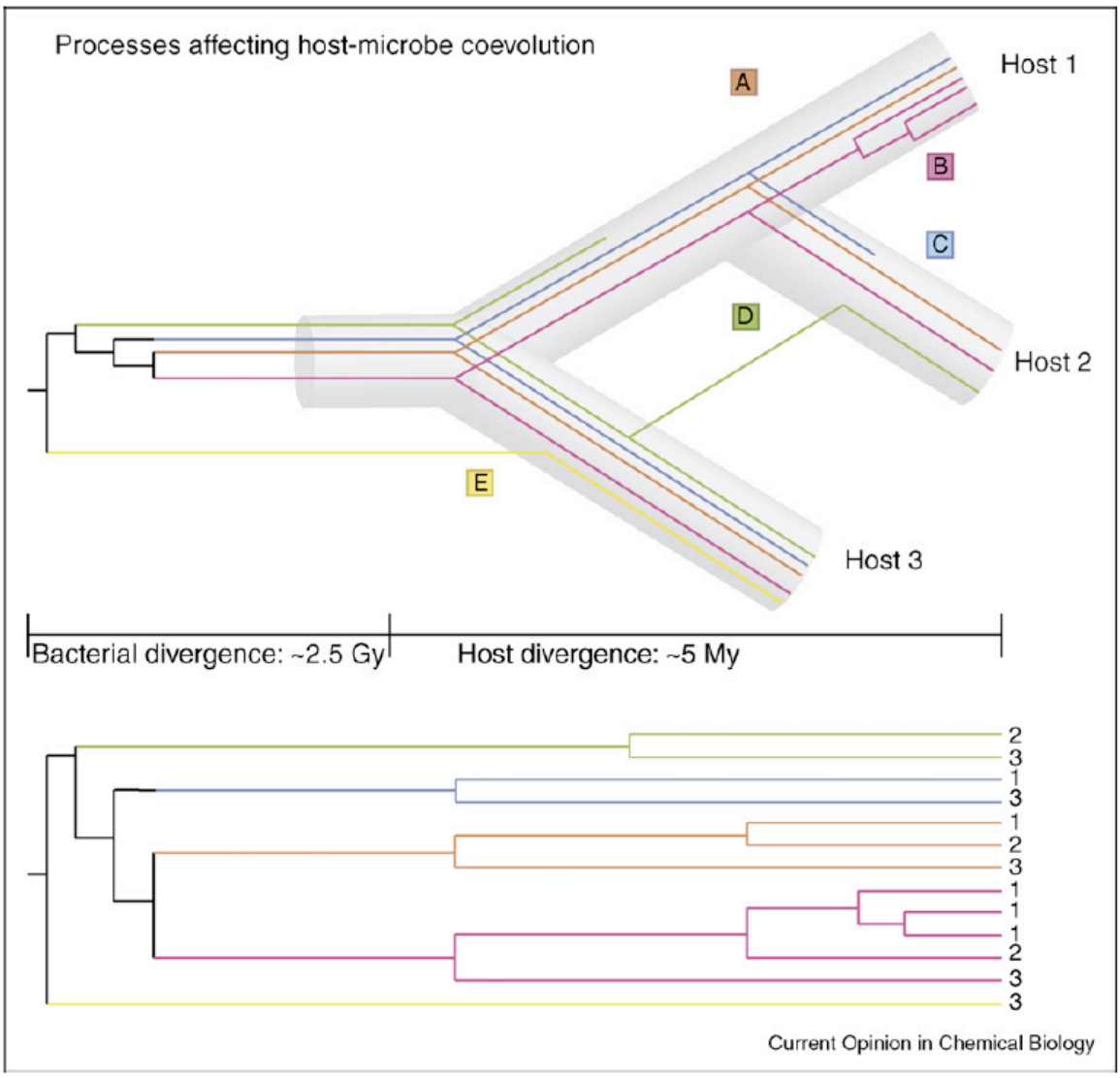


Figure 2. **Processes affecting host-microbe coevolution.** Large gray tubes indicate relationships between host lineages, while thin colored lines indicate relationships between microbial lineages (top tree). Importantly, divergence between bacterial divisions (black lines) occurred over a much longer time-scale than divergence between metazoan host lineages such as, for example, humans, chimpanzees and bonobos (see scale). Processes depicted include co-divergence of host and microbial lineages (A), diversification of a microbial lineage while associated with a host (B), extinction (C), host-swapping (D) and association of a free-living microbial lineage with a host (E). Co-divergence, adaptation to a novel host, host-swapping or diversification within a lineage all produce splits on the microbial phylogenetic tree (bottom tree). Although great caution and deep sampling are required, studies of the distribution of microbial lineages among hosts (numbers at tips of bottom tree), and comparison of the microbial phylogeny to the host phylogeny, may help to resolve which evolutionary processes are responsible for observed divergence between host-associated microbial lineages.

Mammal species that independently acquired similar diets may support the same symbiont species, which evolved in one and then found conditions in the other to be suitable.

An extension of the concept of co-differentiation between host and symbiont is co-differentiation of an entire microbial community with a host animal lineage. Here, an entire microbial community would be passed vertically from host to offspring. Over the course of speciation events in the host lineage, the microbial communities would differentiate in a way that would mirror host phylogeny. Such a scenario would be expected in host lineages where parents inoculate their offspring with a microbial consortium that is highly adapted to a specialized diet. An example of such hosts is the Koala bear: mothers inoculate their young with "pap", a specialized dropping that allows the young to make the transition from milk to a folivorous diet of *Eucalyptus* leaves and branches[9].

Coevolution has been invoked to describe the relationship between mammals and their gut microbial communities, because communities differ between species (mouse, cow, pig and human[10]). However, these differences could instead stem from selection of microbial lineages by a host's diet or immune system. Little is known about variation within each species, so differences between samples could also primarily reflect differences between individuals rather than between species. Similarly, gut microbial communities may be composed of environmental microbes pre-adapted to the chemical milieu of the gut, rather than microbes that have co-evolved with their hosts. Unambiguous demonstration of co-diversification of mammals and their gut microbes has not yet been achieved: for example, patterns of

community similarity obtained by comparing the gut microbiotas of a range of mammals (e.g. using distance measurement algorithms such as UniFrac[11]), might mirror the phylogeny of the mammals. Such a test requires a systematic survey of the microbial communities associated with animal hosts representing a range of taxonomic orders and diets.

Identifying genes critical for symbiosis

Functional and comparative genomic analyses of human gut symbionts are revealing genes critical for adaptation to the gut environment, and mechanisms for horizontal transfer of these genes. These studies, along with in-depth analyses of symbionts of invertebrate hosts, provide a necessary framework for designing and interpreting metagenomic studies of the human microbiome.

In contrast to the human gut, which houses a diverse microbial community, many invertebrates (e.g. aphids, sharpshooters, and stinkbugs) have simple communities that are either maternally transmitted directly to the offspring[12, 13] or are eaten as maternally-deposited capsules shortly after hatching[14]. These symbiont genomes have dramatically reduced gene content, but retain genes for key metabolic capabilities that complement host physiology, including vitamin and amino biosynthesis[12, 13]. However, some human gut symbionts, including members of a large division of bacteria known as the Bacteroidetes, have maintained a larger genome size[15], perhaps because they must survive outside the host to be transmitted.

Studies of mice raised to adulthood in sterile isolators without any exposure

to microbes ('germ-free' animals) are especially useful complements to genomic approaches for understanding the function of gut microbes. For example, germ-free mice colonized with a prominent human gut symbiont, *Bacteroides thetaiotaomicron,* demonstrate that this bacterium can selectively induce a set of its genes that degrade otherwise indigestible dietary polysaccharides[16]. Genomic analysis of *Methanobrevibacter smithii* showed that this species likely promotes energy harvest in hosts by consuming a range of fermentation products of other gut bacteria, and may be a good target for anti-obesity drugs[17]. Comparative genomic analyses of gut and non-gut Bacteroidetes[15] revealed that gut Bacteroidetes possess large arsenals of genes that sense the nutrient environment. These nutrient sensors are linked to gene clusters encoding proteins involved in acquiring specific classes of glycans, and degrading these glycans by glycoside hydrolases and polysaccharide lyases. The products of these polysaccharide utilization gene clusters are used by other members of the microbiota that are ill-equipped to degrade complex glycans, but are well-endowed with genes involved in importing monosaccharides and converting them to fermentation products that can be utilized by other components of the microbiota, and the host[18]. These types of studies can be expanded to model communities of sequenced gut symbionts that are introduced into normal or genetically engineered germ-free mice: the effects of diet and or drugs can be carefully monitored in these 'gnotobiotic' mouse models under conditions where potentially confounding variables, such as host genotype and diet, can be constrained. Gnotobiotic mouse studies will provide better understanding of the rules and forces that govern the assembly and operations of microbial communities,

proof-of-principle experiments that ascertain the contributions of specified groups of microbes to community and host operations, and proof-of-concept tests of the efficacy of new types of anti-microbial drugs that target horizontal gene transfer between members of a microbiota or the activities of virulence factors embedded in a microbiome (see below).

A preliminary study of microbial gene content in the fecal microbiota, which mirrors the microbiota of the distal gut, of several healthy humans showed that compared to our *H. sapiens* genome and the genomes of all sequenced microbes, there is an enrichment of the representation of genes involved in vitamin biosynthesis, degradation of diet- and host-derived polysaccharides as well as xenobiotic metabolism[5]. Further analysis revealed that the gut microbiome is also enriched in a family of conjugative transposons, consistent with a pronounced role of horizontal gene transfer (HGT) in shaping gut microbial genomes[19].

Horizontal gene transfer and coevolution

HGT is an important factor in the evolution of microbial communities that promotes adaptation to novel or changing environments, including mammalian host environments. HGT is of intense medical interest, not only because of its contribution to the spread of antibiotic resistance genes, but also because it can cause closely related strains to differ drastically in clinical parameters. For example, type III secreted effectors may contribute to differences in host specificity between strains of *Salmonella enterica*[20]. On a longer time scale, the acquisition of the type III secretory systems encoded by the SPI-1 and SPI-2 pathogenicity islands is a

defining feature of host adaptation for *S. enterica* as a whole[21].

Several novel strategies for drug development are being pursued in response to the challenge posed by the horizontal transfer of genes involved in both antibiotic resistance and virulence. These strategies include the development of compounds that directly inhibit gene transfer[22] or virulence[23].

Targeting virulence factors with small-molecule inhibitors directly presents several potential advantages. Such targeting may cause less collateral damage to the indigenous microbiome than traditional antibiotics, may exert less selective pressure for the evolution and transfer of resistance, and may be effective against divergent organisms that have acquired a particular virulence factor by HGT.

Genomic islands contain a rich source of genes of unknown function[24] that may yield novel virulence factors appropriate for small-molecule inhibitors. For example, a recent screen for *Salmonella enterica* serovar Typhimurium genes involved in survival and replication within macrophages (a key feature of persistent infection by *Salmonella*) found that such genes were dramatically overrepresented within putatively transferred regions, such as prophages and pathogenicity islands[25].

Gene transfer systems themselves are also being targeted with small-molecule inhibitors. Such inhibitors could be co-administered with antibiotics to prevent the *in-vivo* acquisition of resistance factors by susceptible pathogens during the course of antibiotic therapy. For example, the bisphosphonate compounds clodronate and etidronate inhibit the F plasmid TraI relaxase *in vitro* and conjugative transfer of F plasmid *in vivo*. These findings are particularly significant

because relaxases are essential components of conjugative transfer systems, and the F plasmid TraI relaxase is closely related (~99% sequence identity) to the relaxases of many plasmids known to transfer antibiotic resistance genes[22].

The promise of metagenomic approaches

The vast majority of phylogenetic diversity in microbial communities associated with the human body (and other ecosystems) is represented by organisms that are difficult or impossible to culture in the laboratory using currently available methods. Standard culture methods are especially problematic for understanding symbiosis, because microbes may express completely different sets of genes or may not grow at all outside the host. Metagenomics allows us to observe the genes contained in this vast uncultured majority through the isolation and sequencing of DNA directly from the community. Typically, 16S rRNA gene sequences are used as a phylogenetic marker to probe community structure and diversity ('who's there, and in what abundance?'). The rest of the genes in the microbiome are characterized through shotgun sequencing of whole microbial community DNA. Although the short 200-250 nucleotide reads currently obtained by the latest generation of massively parallel DNA sequencers (i.e. pyrosequencers) are sufficient for characterizing communities based on their 16S rRNA gene content[26], characterizing other genes typically requires a combination of short pyrosequencing reads, longer Sanger-sequencing reads, and, ideally, complete reference genomes (i.e. genomes of cultured representatives of major phylogenetic lineages present in the community) (Figure 3). In simple communities, such as those present in certain invertebrates or

in the environment, metagenomic data can be sufficient for assembling the genomes of their constituent microbes. Nonetheless, assembly of sequenced genomic DNA fragments from a microbiome into relatively large contiguous sequences of physically linked genes derived from a given organismal genome remains challenging, especially in complex communities such as the human gut microbiota [5, 19, 27]. Despite these challenges, metagenomic studies have revealed specific genes that are enriched in the gut microbiome of humans (see above), as well as microbial genes whose representation is enriched in mouse models of human diseases, including obesity[27].

Because assembly is difficult, metagenomic analysis of complex communities is currently 'gene-centric': DNA sequences are mapped to known genes and genomes to infer the relative abundance of different genes and metabolic pathways[27, 28]. One major challenge in metagenomics is to link these 'gene centric' functional predictions to the organism that contained each gene. This goal is complicated by a lack of complete reference genomes and complete/consensus reference taxonomies, and by the short DNA fragment generated by pyrosequencing. Functional assignments typically rely on homology searches. Taxonomic assignment is more complex, and usually involves aligning homologs, and building phylogenetic trees. Annotations can then be assigned based on the best matches or closest homologs in a reference set of genes. Thus, taxonomic assignment is challenging and computationally expensive even with full-length sequences, and results are strongly affected by the reference set.

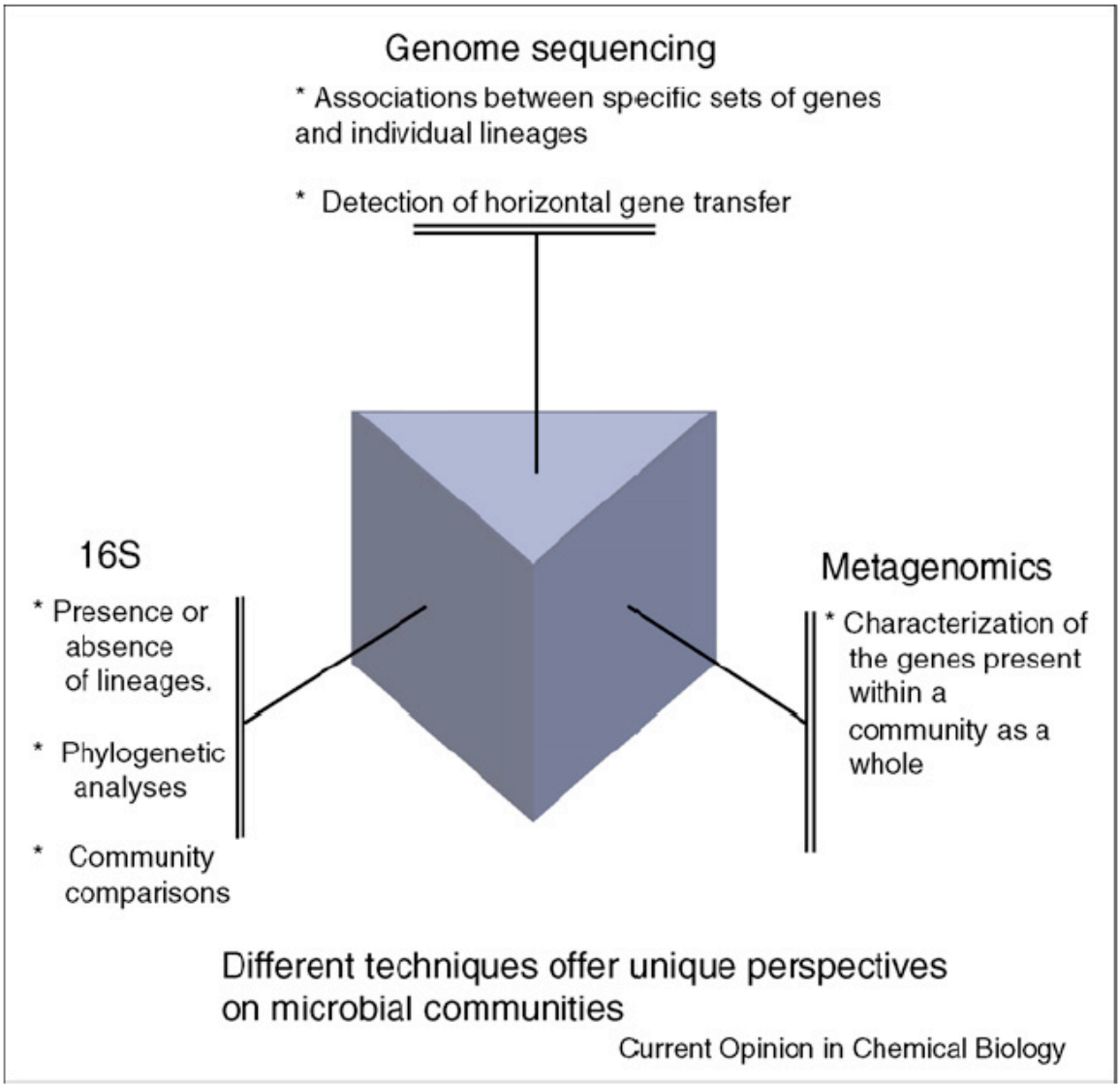


Figure 3. A comprehensive understanding of our microbial ecology requires integration of many data sources. Community profiles with 16S rRNA indicate which types of organisms are present, metagenomic profiling allows us to identify specific functional categories of genes that are critical for differences in function, and complete genomes act as scaffolds for understanding changes in gene content through loss, amplification, and HGT that allow microbes to adapt to functional roles in different environments.

Another major challenge is that individual labs usually do not have the capability to completely characterize a complex community through metagenomics: many key analysis tools do not scale to large datasets, and the cost of sequencing and the computational resources can be formidable. The most successful studies have come from collaborations spanning a range of disciplinary expertise, including physiology, microbiology, molecular biology, statistics, molecular evolution, ecology and high-performance computing. Collaboration and coordination at different project scales are clearly needed[18].

Combined phylogenetic and genomic approaches

Due to the limitations of 16S rRNA community surveys, genome sequencing, or metagenomic surveys individually **(**Figure 3**)**, there is a pressing need to develop methods that allow for the information generated by each of these techniques to be related. In particular, although 16S rRNA surveys can readily identify differences between microbial communities, it is difficult to reliably infer information about specific uncultured lineages identified in these surveys. In general, this problem is addressed by extrapolating infromation from the closest relative for which a complete genome sequence or cell culture is available. However, given the prevalence of horizontal gene transfer (see Chapter III) and other forms of microbial genome plasticity, there are many unanswered question

regarding the reliability of such inferences. For example, how closely related must an uncultivated bacterium be to a known pathogen before we can infer with 95% confidence that it is also a pathogen? Currently such problems are not easily addressed. Thus, one key focus of my graduate research was to describe the relationship between the genes shared between microbial genomes and the relatedness of those microorganisms in phylogenetic trees. By conducting descriptive studies of how habitat adaptation, changes in genome size, and horizontal gene transfer affect the relationship between phylogeny and genome contents, I hope I have helped to set the stage for predictive methods that will allow us to accurately assess how much we know (or don't know) about the vast numbers of uncultivated microorganisms based on their position on the tree of life. While addressing this methodological issue, these descriptive studies have also yielded new insights into the process by which microorganisms have evolved to live in association with their human hosts, either as commensals or pathogens.

Overview of the Thesis

This thesis is organized around studies of the process by which microbial organisms have evolved to inhabit the mammalian intestinal tract. The chapters present a subset of my published work as a graduate student [1-4, 29-32]. Chapter II addresses the question of how adaptation to life in the human intestinal tract changes the genomes of gut-adapted bacteria relative to their non-gut neighbors. Surprisingly, I found that adaptation to life in association with the gut produces a common pattern of genomic changes (in terms of the presence or absence of genes)

in both disease-causing and commensal bacteria. Adaptation to the gut environment appears to have opposite effects in closely-related and distantly related bacteria. In closely related bacteria, adaptation to the gut tends to promote divergence of gene content, which I hypothesize may be due to niche specialization. For more distantly related pairs of bacteria, however, adaptation to the gut instead tended to produce unexpected commonalities in gene content. I propose that these commonalities amongst distantly related bacteria may represent convergent evolution of gene content in response to the challenging gut environment. Based on investigations of the gene content of plasmids, I found that the genomic changes associated with adaptation to the gut are more intense in these mobile elements. This suggested that gene transfer may play an important role in producing the patterns of habitat adaptation to the gut that I observed. Chapter III follows up on the idea that gene transfer may play a critical role in microbial habitat adaptation to the gut. In it, I present a study of horizontal gene transfer (HGT) in 22 strains of *Methanobrevibacter smithii,* an abundant gut archaeon. I describe the application of multiple independent HGT detection algorithms to understand the role that gene transfer has played in *M. smithii*'s adaptation to the human gut, as well as the differences between strains. Consistent with the findings in bacteria from Chapter II, I found that gene transfer has played an important role in introducing gene functions important for *M. smithii* adaptation to the human gut. Gene transfer appears to have both introduced novel traits that are now common to all known *M. smithii* strains (i.e. part of the *M. smithii* core genome), as well as contributed to massive differences between strains in particular protein families

(specifically a class of adhesin-like proteins). Chapter IV synthesizes some of the lessons learned from the studies described in Chapters II and III, and proposes a generalized workflow for combining phylogenetics and comparative genomics to gain new insights into microbial habitat adaptation. Chapter IV also reviews recent large-scale studies of bacterial habitat adaptation, thus helping to situate the research in the thesis within the field as a whole. Finally, Chapter V contains conclusions drawn from the thesis as a whole, and suggests directions for future research.

CHAPTER II.

EFFECTS OF ENVIRONMENTAL ADAPTATION ON GENOME CONTENT IN GUT-ADAPTED BACTERIA AND THEIR NON-GUT RELATIVES

Introduction

As discussed in the introduction, our microbiota (including bacteria, archaea, and eukaryotes) plays many important roles in human health. Thus, it is important to understand how both commensal and pathogenic members of our microbiota have evolved to live on and in our bodies. In this chapter I describe my investigations of genome evolution in bacteria that have evolved to live in the human intestinal tract, in contrast to their non-gut relatives. The results presented in this chapter are derived from a first-authored paper that I published in *Nucleic Acids Research* in collaboration with Catherine Lozupone, Jeff Gordon and Rob Knight. In the chapter, I investigate several aspects of the relationship between phylogeny and genome content. This relationship is interesting because most microbial organisms are uncultured, their physiological properties only known by their position in the tree of life and the samples in which they were discovered [33]. Thus, determining the extent to which phylogenetic relatedness (to a well-studied organism) predicts genome content is important for the interpretation of many

microbial ecology studies that rely on 16S rRNA gene surveys. I address four inter-related questions: (i) How well does phylogeny predict gene content in gut-adapted bacteria? (ii) Does phylogeny yield useful information about gene content within (as well as between) bacterial species? (iii) How is the relationship between phylogeny and gene content affected by habitat adaptation to the gut? (iv) How do these relationships change when we examine plasmids, which are subject to frequent horizontal gene transfer? The answers to these questions shed light on the process by which bacteria adapt to life in the gut, and also provide useful lessons for the interpretation of 16S rRNA gene surveys.

The human gut harbors the largest collection of microbes in any of our body habitats; its microbiome is of great interest because the microbiota appears to have pervasive effects on health and disease, including the development of a functional immune system, vitamin synthesis, and nutrient processing [18]. Culture-independent methods for the discovery of novel microbial lineages using 16S rRNA gene sequencing have revolutionized our understanding of microbial diversity [33-35]. The 16S rRNA gene is an excellent marker of average genomic evolution because it is a core gene that seldom undergoes horizontal gene transfer and has a phylogeny that matches other core genes, because it appears to evolve largely independently of ecological diversification, and because it contains both fast- and slow-evolving regions and can thus be used to resolve relationships among taxa at different phylogenetic depths (see [33, 36, 37] and [38] for reviews on the topic). 16S rRNA based-surveys indicate that bacterial communities of the mammalian gut differ more from

non-gut communities than even the most extreme free-living communities differ from one another [39]. This observation suggests that life in the intestinal environment may have demanding and distinctive functional requirements. Understanding whether 16S rRNA surveys that reveal which species (or higher taxa) are present relate directly to diversity in functional gene repertoires is critical for Human Microbiome Projects [18]: these projects generally seek to relate variation in the phylogenetic composition of the microbiome, as profiled by 16S rRNA surveys, to health and disease [40-44]. To begin addressing this question, I ask whether gut-dwelling species have converged on more closely related gene repertoires than expected from their phylogenetic relationship. In particular, is the degree of overlap in the gene repertoire of gut dwellers greater than that for non-gut dwellers after a given amount of evolutionary time?

Differences in 16S rRNA gene sequences between genomes are related to overall levels of gene conservation between those genomes and to the average nucleotide identity (ANI) of genes conserved between them [45], although whether the same trends hold true for very closely related genomes (e.g. those within the same bacterial species) is unknown. Several mechanisms alter genome content, including genome reduction, gene duplications, and horizontal gene transfer. These have been extensively studied. However, the effect of differences in habitat on the rate of evolution of gene content has only been systematically studied using a small number of species, primarily from non-host-associated habitats [46]. Substantial variation in

gene content has been observed within individual bacterial species, whether isolated from many environments (such as *Escherichia coli*) [47] or highly habitat-restricted (such as *Helicobacter pylori* [48]).

These observations that bacterial species vary in their degree of gene conservation [46, 49, 50], raise the question of whether the differences are due to differences in population structure [48], diversity within and/or between habitats, or ecological interactions with other organism [47]. For example, the rate at which gene content varies with phylogenetic distance [46] might be due to any of the mechanisms outlined above. Two well-characterized examples of associations between specific environments and mechanisms of genomic change are the extreme genome reduction observed in obligate intracellular symbionts and intracellular pathogens [51-53] as well as microbial adaptation to hypersaline environments through enrichment of proteins throughout the proteome with the acidic amino acids aspartate and glutamate [54, 55]. However, signatures of adaptation to specific environments have generally been difficult to obtain.

The mammalian gut provides an attractive model to explore these issues, because it harbors an especially restricted group of lineages [39]. If this restriction results from a highly selective environment, we might expect that different species adapt to the gut by convergent evolution in gene content. More generally, there are several reasons why bacteria sharing a habitat may share more or fewer genes than phylogenetic distance alone would predict [46]. For example, adaptation to a shared

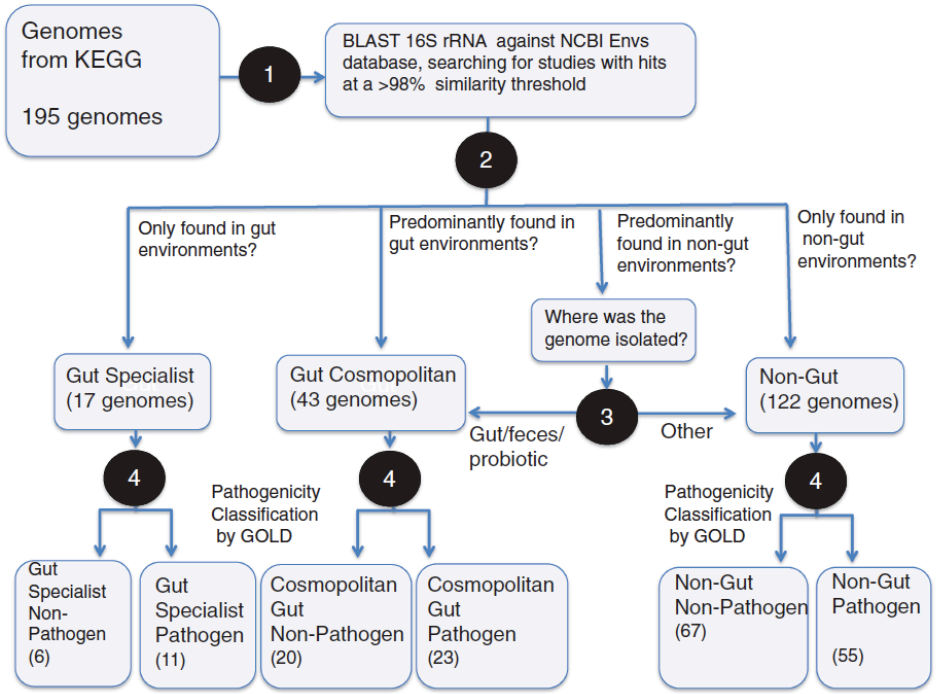
environment might enrich for the same genes necessary for growth and survival in that environment, and horizontal gene transfer may increase in densely packed communities, leading to more shared genes (for example, the distal mammalian gut can contain up to 10^{12} cells/ mL lumenal contents). Alternatively, competition within a shared environment could produce niche specialization [56-58] as strains diversify their gene content and exploit underutilized resources. Thus, Ireason that inferring the relationship between evolutionary distance, as measured by 16S rRNA sequence divergence, and functional relatedness, at the level of overlap in gene repertoires, could assist in discriminating among these various mechanisms.

Methods

Selection and classification of genomes. I sought to identify genomes representing abundant gut lineages that were specialist or cosmopolitan, and nonpathogenic or pathogenic. To do so, I downloaded 195 genomes from the KEGG database that were members of the Actinobacteria, Bacteroidetes, Firmicutes (separating the Clostridiales and the Lactobacillales), δ-Proteobacteria, ε-Proteobacteria, and the γ-Proteobacteria (Enterobacteria). The bacteria from which these genomes were sequenced were then characterized according to their habitat and pathogenicity status (Figure **1**) according to the following workflow: **(1)** To obtain information on the lifestyle of the isolates from which genome sequences were obtained, Catherine Lozupone (my co-author in this project) determined which 16S rRNA-based environmental surveys of microbial assemblages had deposited

sequences in GenBank that were nearly identical to the 16S rRNA sequence in the corresponding complete genome. She first downloaded the gbenv files from the NCBI ftp site on 12/31/07 and used them to create a BLAST database. These files contain GenBank records for the ENV database, a component of the non-redundant nucleotide database (nt) where 16S rRNA environmental survey data are deposited. GenBank records for hits with >98% sequence identity over 400 bp to the 16S rRNA sequence of each genome were parsed to obtain a list of study titles associated with the hits. **(2)** These study titles were used to determine whether close relatives of each of the isolates had been found only in the gut (gut specialist), never in the gut (non-gut), or in the gut as well as a diversity of free-living communities (gut cosmopolitan). **(3)** In ambiguous cases, where close relatives of the isolate were found in many environmental samples and only rarely in gut samples, I used isolation information from the Genomes Online Database (GOLD) to determine whether a genome represented a cosmopolitan member of the gut or a non-gut organism.

(a)



(b)

Taxon	Lifestyle	Pathogenicity	n	Example Organism
Actinobacteria (all)	G	N	1	*Bifidobacterium adolescentis* ATCC 15703
	GC	N	1	*Bifidobacterium longum* NCC2705
	N	N	20	*Corynebacterium efficiens* YS-314T
	N	P	19	*Corynebacterium diphtheriae* gravis NCTC 13129
Bacteroidetes (all)	G	N	3	*Bacteroidetes thetaiotaomicron* VPI-5482
	N	N	4	*Salinibacter ruber* M31
	N	P	1	*Flavobacterium psychrophilum* JIP02/86
Firmicutes – Bacilli (Lactobacillales)	G	N	1	*Lactobacillus reuteri* F275, JCM 1112
	GC	N	16	*Lactobacillus sakei sakei* 23K
	GC	P	2	*Streptococcus sanguinis* SK36
	N	N	2	*Oenococcus oeni* PSU-1
	N	P	22	*Streptococcus pneumoniae* D39
Firmicutes - Clostridia	G	P	4	*Clostridium perfringens* ATCC 13124
	N	N	14	*Desulfitobacterium hafniense* Y51
	N	P	5	*Clostridium tetani* Massachusetts E88
Proteobacteria - Delta	G	P	1	*Lawsonia intracellularis* PHE/MN1-00
	N	N	15	*Desulfovibrio vulgaris* vulgaris DP4
Proteobacteria - Epsilon	G	N	1	*Campylobacter hominis* ATCC BAA-381
	G	P	5	*Campylobacter jejuni* jejuni NCTC 11168
	GC	P	2	*Helicobacter hepaticus* ATCC 51449 (= 3B1)
	N	N	3	*Nitratiruptor* sp SB155-2
Proteobacteria - Gamma (Enterobacteria)	GC	N	3	*Escherichia coli* K12- MG1655
	GC	P	19	*Salmonella enterica* sv Paratyphi A SARB42
	N	N	8	*Buchnera aphidicola* Cc
	N	P	8	*Yersinia pestis* Nepal516

Figure 4. Classification of species by habitat and pathogenicity. (A) All genomes for the Actinobacteria, Bacteroidetes, Firmicutes (separating the Clostridiales and the Lactobacillales), δ-Proteobacteria, ε-Proteobacteria, and the γ-Proteobacteria (Enterobacteria) present in the KEGG database were downloaded (195 genomes total). The genomes were classified as follows (see Methods for detailed description): 1) BLAST was used to compare 16S rRNA sequences for each genome against the NCBI Envs database to determine the environmental distribution of the species. 2) Genomes were characterized by examination of the study titles of hits: genomes found exclusively in gut or fecal samples were labeled 'gut specialist', those found in several studies of the gut, but also in other environments were categorized as 'gut cosmopolitan', while those never found in the gut were labeled 'non-gut'. 3) In borderline cases where genomes were found in several environmental samples and only a small number of gut samples, isolation information from the GOLD database was used to determine whether the genome should be categorized as 'gut cosmopolitan' or 'non-gut'. Probiotic bacteria, or those isolated from the gastrointestinal tract or feces in this abundance class were taken to be 'gut cosmopolitan'. 4) Finally, genomes in each category were categorized by pathogenicity using the GOLD (26) annotations for 'phenotype' and 'disease'. Commensal microbes capable of only opportunistic infection were treated as non-pathogens in this analysis. Additionally, 13 genomes where annotation information was ambiguous or conflicted with observations from 16S rRNA observations were removed from the analysis. (B) Example output of this annotation process, and numbers of genomes in each subcategory. Abbreviations are as follows: 'G', gut specialist, 'GC' cosmopolitan resident of the gut, 'N' non-gut. Pathogens are denoted 'P' and non-pathogens 'N'. from the GOLD database was used to decide how a genome should be categorized. In these ambiguous cases, strains annotated as probiotic, or strains isolated from the distal gut or feces, were categorized as 'gut cosmopolitan' whereas others were categorized as non-gut.

I removed thirteen genomes from subsequent analysis because their isolation and phenotypic annotations from GOLD were ambiguous or conflicted. This classification process yielded 17 gut specialists, 43 gut cosmopolitan and 122 non-gut bacteria. **(4)** Within each of these four categories, I identified pathogens using GOLD annotations downloaded 10/08/2009 [59].

Gene Conservation. Gene conservation was measured as the proportion of genes in the query genome with at least one homolog conserved in the subject genome (see BLAST analysis, below). This measure is asymmetric because the query and subject genome can be of different sizes (for example, if genome A contains 500 genes,

genome B contains 5000 genes, and they share 250 genes, B contains 50% of the genes in A, but A contains only 5% of the genes in B). Because comparisons between genomes with large size differences was found to produce aberrant clusters of high or low gene conservation (see Results), I placed genomes into three size categories +/- one standard deviation from the mean genome size: these categories were small (< 1783 genes), medium (1783– 4964 genes), and large (> 4964 genes). Comparisons between genomes in different size categories were then excluded from the analyses where noted below (see figure legends). Because plasmids are subject to frequent horizontal gene transfer, and because the absence of plasmids in the strain chosen for genome sequencing does not indicate their absence in the corresponding natural populations, queries from plasmids were excluded from the analysis for comparisons of gene content to evolutionary distance. To assess the significance of correlations between evolutionary distance and gene content conservation, Mantel tests with 10,000 permutations were run on either the full matrix of comparisons for each taxon analyzed, as well as subsets of those matrices subdivided by environment, pathogenicity or chromosome type (chromosome or plasmid). Tests were performed using the Mantel test implementation in the PyCogent toolkit [60].

BLAST analysis. BLASTp analyses were conducted using a custom python script based on PyCogent [60] to run NCBI BLAST [61]. Analyses were run using the BLOSUM62 matrix (-M BLOSUM62) with maximum hits was set to 1 (-m 1). Hits were then filtered to an e-value threshold of 10^{-10} (analyses using alternative e-value

thresholds altered the slope of results but not the qualitative outcome, data not shown), and hits with alignable regions shorter than 75% of the length of both query and subject were rejected.

Tree Construction. 16S rRNA sequences for each of the genomes under study were identified by BLASTing the E.coli rrsG gene against the nucleotide (.nuc) file from KEGG, (http://www.genome.ad.jp/), for each genome with an e-value threshold of 1e-20 and word length of 11. Some genomes contain multiple 16S rRNA sequences. Iverified manually that the BLAST settings used identified all 16S rRNA sequences from several such genomes (and no others) that had been identified in a previous study [62]. 16S rRNA sequences identified in this manner were then aligned using NAST [63].

In cases where multiple 16S rRNA sequences in a single genome passed the NAST screen, sequences were selected randomly. The Lane mask [64] from GreenGenes [65] was applied to the selected NAST-aligned sequences. Phylogenetic trees were constructed in ClearCut [66] using traditional neighbor-joining and the Kimura two-parameter distance correction. In order to determine whether short reads such as those generated by pyrosequencing would suffice for analyses of gene content and evolutionary distance, trees were also constructed using simulated pyrosequencing reads. In this case, trees were also constructed by the same procedure, but instead using only the regions of the 16S rRNA corresponding to 250 bases of the regions

amplified by V2, V4, and V6 primers [67]. These were generated by taking only the corresponding regions from the full-length 16S rRNA sequences. The gaps were then removed, and the sequences realigned. The coordinates in the GreenGenes 7682 bp format for these regions were: V2, 1869 to 2353; V4, 2310 to 4100; and V6, 4625 to 5877.

Results

A scale relates gene content to 16S rRNA evolutionary distance. I calculated gene conservation for all pairs of bacterial genomes in the KEGG database from within the Actinobacteria, Bacteroidetes, Firmicutes (separating the Clostridiales and the Lactobacillales), δ-Proteobacteria, ε-Proteobacteria, and the γ-Proteobacteria (Enterobacteria). These taxa were selected because they contain prominent members of the mammalian gut microbiota [68]. Plotting proportions of shared genes against tip-to-tip distances on a 16S rRNA neighbor-joining tree for the resulting 5737 intra-taxon genome-to-genome comparisons allowed us to infer a model for the relationship between 16S rRNA distances and protein conservation. The proportion of shared genes was determined by performing protein BLAST queries for each gene in that genome against a database composed of all genes in each other genome within the taxon at an e-value threshold of 10^{-10}. The proportions of genes with homologs below the e-value threshold were then plotted against the tip-to-tip distance between the two genomes on a neighbor-joining tree. Initial studies indicated that the BLAST stringency varied only the steepness of the slope but not

the overall patterns; therefore only data for the 10^{-10} threshold is shown although 10^{-4} and 10^{-7} were also used. Gene conservation as measured by protein BLAST was found to decrease exponentially with 16S rRNA distance, in agreement with previous observations [45, 69]. Exponential regression of 16S rRNA distance alone explained only 29% of the overall variance in gene conservation levels. This regression also suggested that gene conservation falls at a rate of $0.62\ e^{-4.326\,\mathbf{d}}$ where **d** is the corrected tip-to-tip distance on a 16S rRNA neighbor-joining phylogeny.

To test whether patterns of gene conservation over evolutionary distance were universal or varied by bacterial taxon, the results were broken down by taxonomy (Figure 5). For all taxa in the analysis, the negative correlation between evolutionary distance and gene content conservation was statistically significant by Mantel Test ($p < 0.05$). However, the

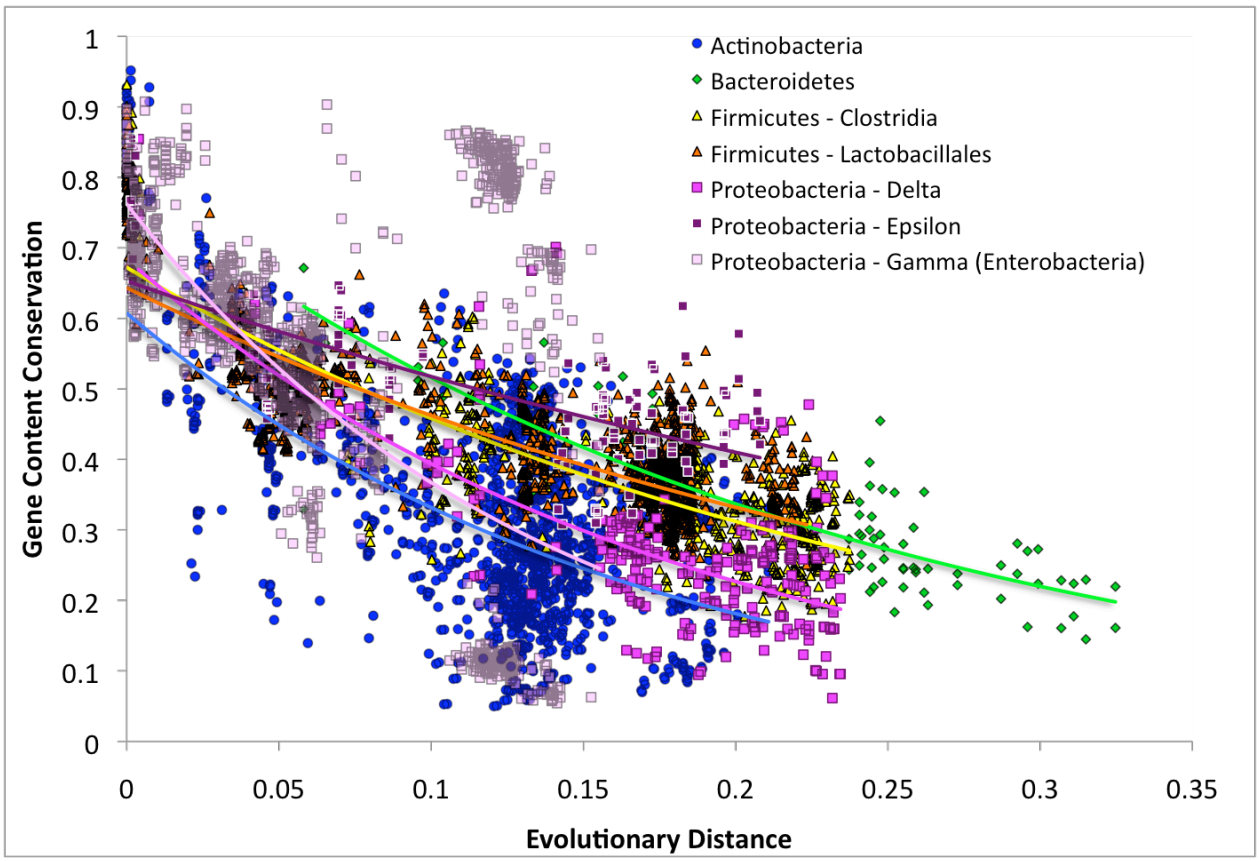


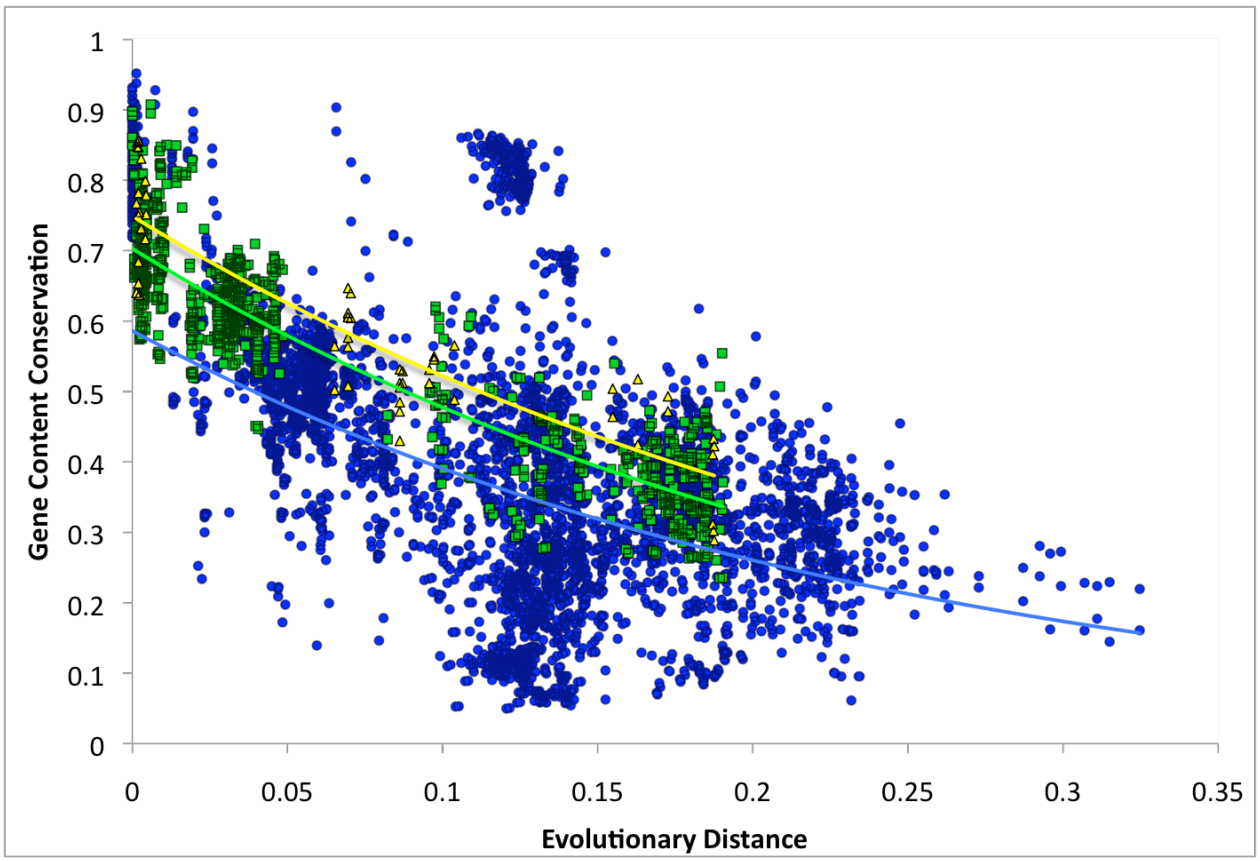
Figure 5. Gene conservation by evolutionary distance. Gene content conservation at the protein level. Each point represents a BLAST comparison between two genomes at an E-value threshold cutoff of 10^{-10}. The x-axis represents the 16S distance between the two genomes, while the y-axis represents the proportion of proteins from the query genome that match proteins from the subject genome. Genome – genome comparisons are subdivided by taxonomic group. Comparisons between members of the same taxonomic group are represented by the same shape and similar colors. Each colored line represents the exponential regression of the points within a single taxon. R^2 values for exponential regression of each taxon were: Actinobacteria, $r^2 = 0.28$; Bacteroidetes, $r^2 = 0.70$; Clostridia, $r^2 = 0.57$; Lactobacillales, $r^2 = 0.70$; δ-Proteobacteria, $r^2 = 0.38$; ε-Proteobacteria $r^2 = 0.48$; γ-Proteobacteria $r^2 = 0.24$.

explanatory power of 16S rRNA gene distance varied greatly between the taxa studied, explaining as little as 28% (Enterobacteria) to as much as 70% (Bacteroidetes) of the variance in gene conservation levels (Figure 5). This heterogeneity could arise from several mechanisms, including different rates of horizontal gene transfer, genome reduction, or habitat specialization in different taxa, which I investigate below.

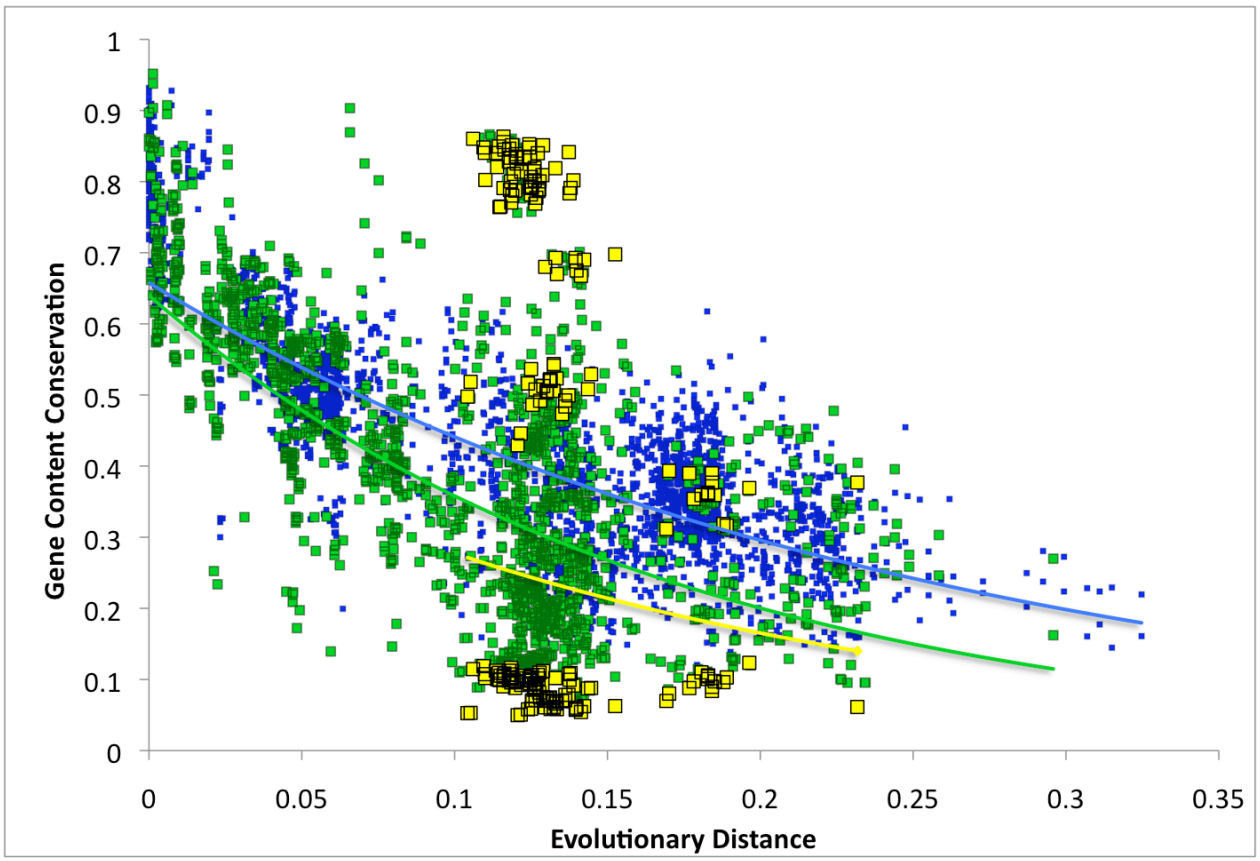
Habitat adaptation and genome size alter aggregate gene conservation. In order to test whether the shared lifestyle of gut-adapted bacteria altered the relationship between gene conservation and evolutionary distance, the genomes in this analysis were categorized based on how often they have been observed in the gut relative to other environments in 16S rRNA studies, combined with information about isolation sources and pathogenicity status derived from the GOLD database [59] (See **Methods** and Figure 4**)**. Species found exclusively in the gut were labeled 'gut specialist', while those frequently found in both the gut and other environments were labeled 'gut cosmopolitan', and those rarely or never observed in the gut but plentiful in other environments were labeled 'non-gut', with isolation information being used to decide borderline cases [59].

Gene content fell exponentially with increasing evolutionary distance for both specialist, cosmopolitan and non-gut species (Figure 6a). In each taxon, and each habitat category, the correlation between gene content conservation and

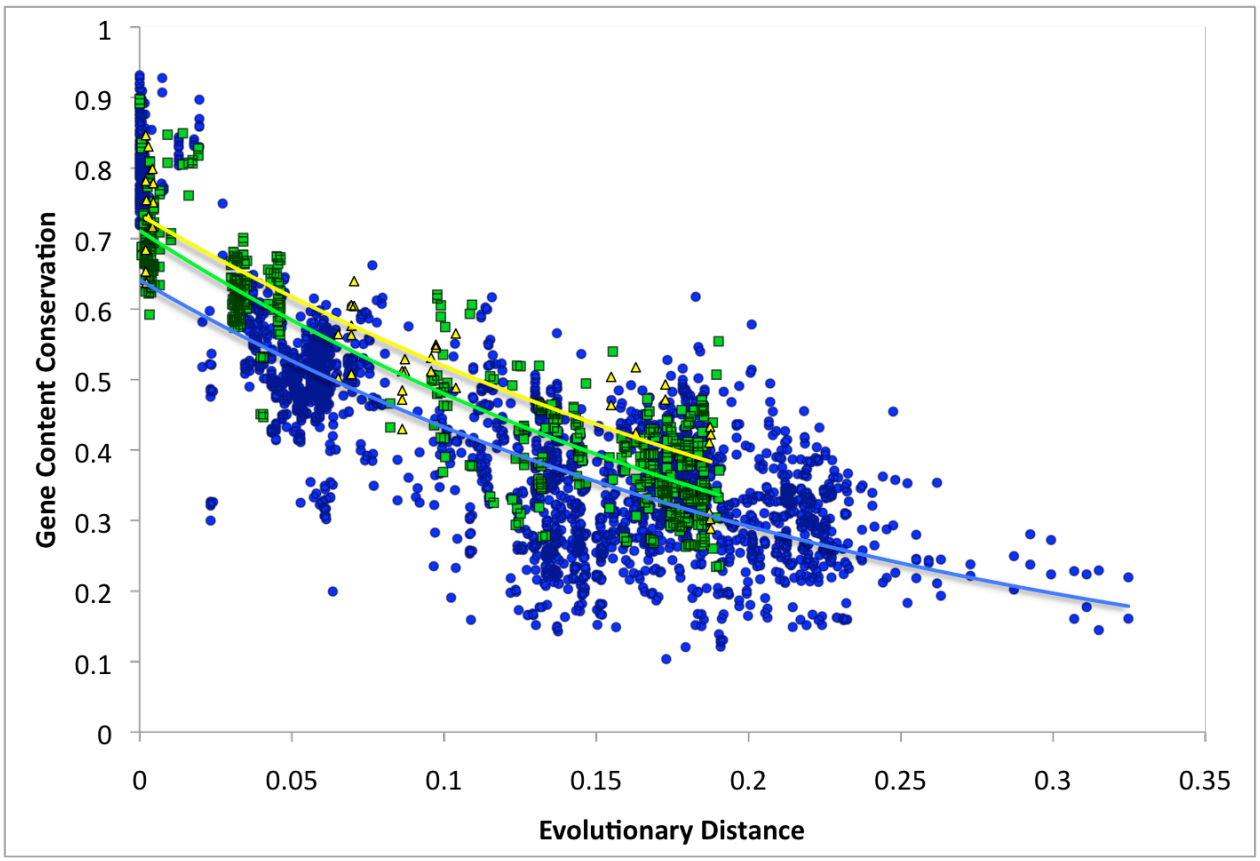
evolutionary distance was statistically significant ($p < 0.05$, Mantel test), except in subcategories for which very few ($n < 5$) genomes were available. Differences in gene content were well explained by evolutionary distance for gut-adapted bacteria (specialists: $r^2 = 0.82$; cosmopolitan: $r^2 = 0.80$), but


Gene Content Conservation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Evolutionary Distance

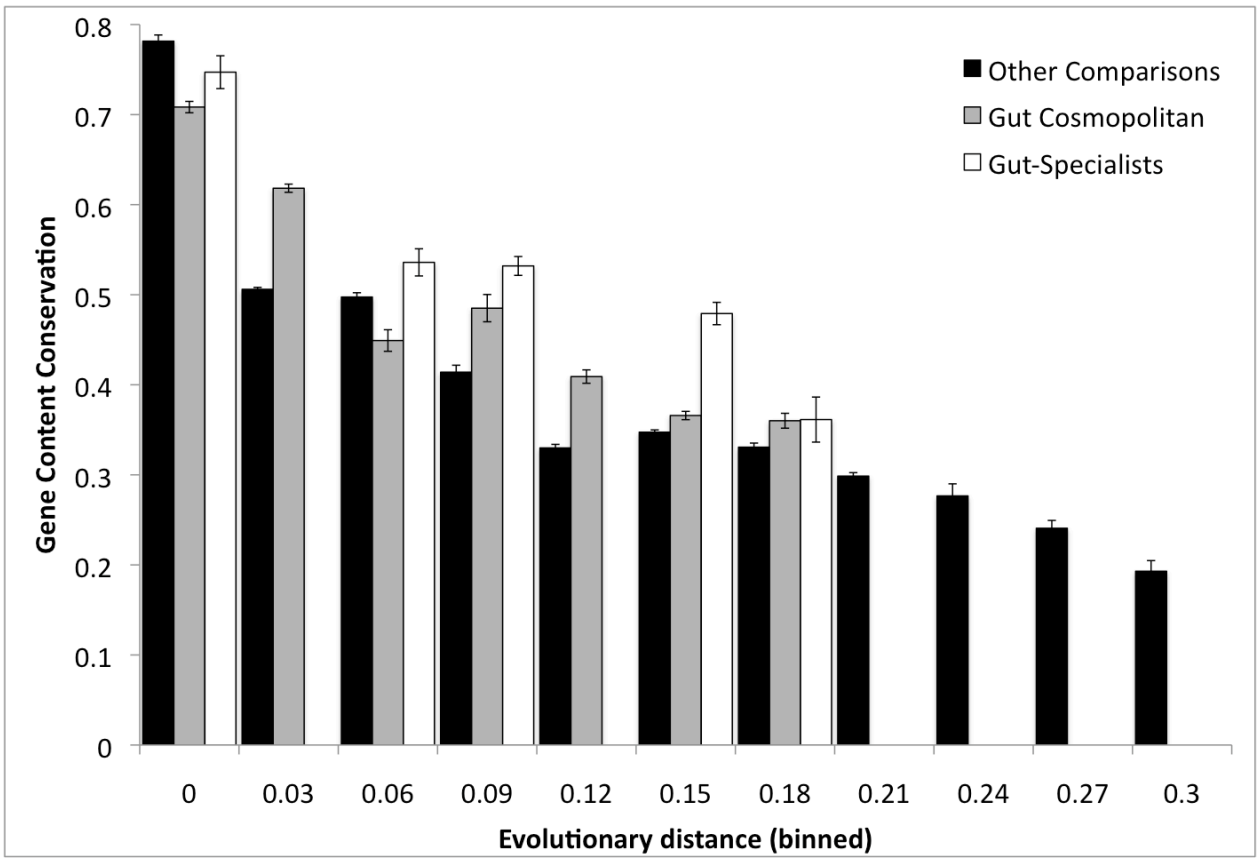
(A)


Gene Content Conservation
Evolutionary Distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35

(B)


Gene Content Conservation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Evolutionary Distance

I



(D)

Figure 6. Gene conservation in gut-adapted bacteria. Relationship between evolutionary distance in terms of 16S rRNA divergence and gene content conservation. For these graphs, the x-axis shows evolutionary divergences in terms of nucleotide substitutions per site in the 16S rRNA gene, and the y-axis shows the fraction of genes in the first species that are found in the second species using BLASTP on the translated sequences. (A) Each point represents a comparison between two genomes. Yellow points are comparisons between two genomes that are both gut specialists, green points are comparisons between two genomes that are both cosmopolitan members of the gut microbiota, while all other comparisons are considered together and colored in blue. Although much variation in gene conservation is explained by phylogenetic distance, examples of genomes that vary little or greatly in gene conservation can be found at any given distance. R^2 = 0.82 for gut specialists; 0.80 for gut cosmopolitans; 0.22 for other comparisons. (B) Effects of relative genome size on conservation of gene content (size categories are defined in Methods above). Genome-genome comparisons were plotted separately for pairs of genomes where both are in the same size category (blue squares), where one genome is medium and the other is either large or small (green squares), or where one genome is large and the other is small (yellow squares). (C) Gene content conservation in pairs of gut-adapted bacteria with similar genome sizes. When only gut specialist or gut cosmpolitan genomes are considered, and when both genomes in each pair are similarly sized, phylogenetic distance is predictive of gene content conservation: r2 = 0.81 gut specialists; 0.78 gut cosmopolitan; 0.57 for other comparisons. (D) Depicts the same data as in panel C, but binned into increments of 0.03 corrected substitutions per site in the 16S rRNA, to clarify trends in conservation. Specialist (white bars) and cosmopolitan (gray bars) bacteria inhabiting the gut have somewhat lower levels of gene conservation at evolutionary distances below 0.03 substitutions per site than non-gut bacteria (black bars), but elevated levels between approximately 0.06 to 0.18 substitutions per site. Error bars depict standard error. Average numbers of genome pairs per bin were: gut specialists, 9.8; gut cosmopolitan 180; non-gut, 567.7.

poorly explained for other comparisons ($r^2 = 0.22$). Importantly, regression analysis indicated that, for a broad range of phylogenetic distances, gut-adapted bacteria possess higher levels of gene conservation than their non-gut relatives, with cosmopolitan members of the gut community being intermediate between gut-specialists and other species.

Because the measure of similarity in gene content (i.e. conservation) used was asymmetric (see Methods), averages of pairwise comparisons among genomes of different sizes can be misleading. Differences in gene conservation attributable to genome reduction are captured in Figure 5 and Figure 6a. Clusters of very high gene conservation were found when comparing reduced genomes to large genomes, and conversely clusters of very low levels of gene conservation were found when comparing large genomes to their reduced relatives.

To investigate the effect of relative genome size on the relationship between evolutionary distance and gene content, the genome – genome comparisons in Figure 6a were re-plotted according to relative genome size (Figure 6c). Each genome was categorized as small, medium, or large according to the criteria defined in Methods. The results from Figure 6a were then re-plotted according to whether the genomes being compared belonged to the same size category (Figure 6b).

Comparisons between genomes with very unequal sizes explain many of the outliers

from the overall trend in gene conservation over phylogenetic distance reported in the analyses above. While phylogenetic distance explained ~ 60% of the variance in gene conservation between genome pairs within the same size category, it explained only 27% of the variance between genome pairs that differed by one size category, and only 1% of the variance in genome pairs that differed by two size categories. This result suggests that controlling for genome size is critical for prediction of gene conservation from phylogenetic distance. Moreover, this is a difference that would be missed if gene conservation were calculated symmetrically. Recalculating the results from Figure 5 to include only genome-genome comparisons (Figure 7) within the same size category yields an r^2 of 0.60 , a ~2 fold improvement in the degree to which variance in gene content can be explained by phylogenetic distance. This improvement applies only to lineages where variation in genome size is substantial. For example, the enterobacteria, rather than appearing as an outlier to the overall trend appear entirely typical once differences in genome size are corrected for (γ-Proteobacteria $r^2 = 0.60$; see Figure 7)

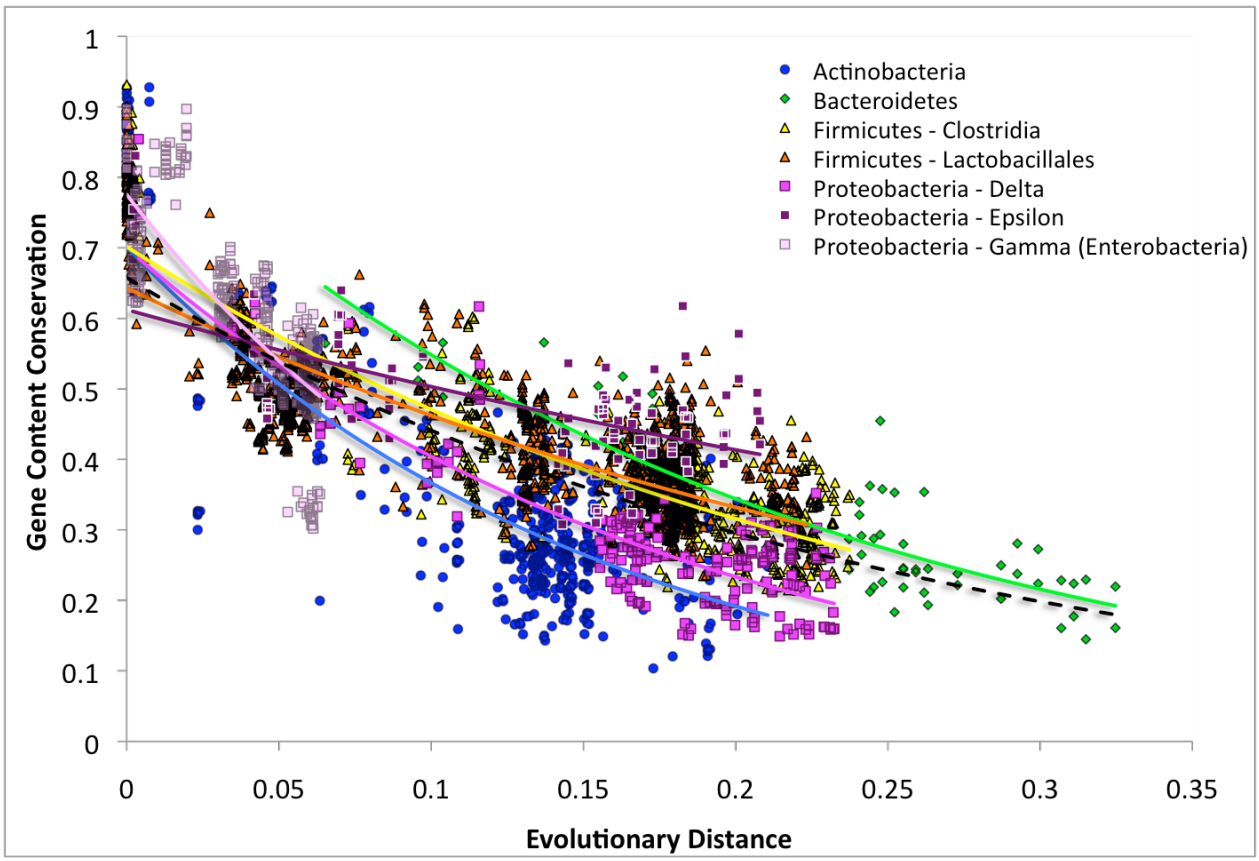


Figure 7. 16S rRNA distance predicts genomic diversity within bacterial taxa in the study, when corrected for genome size.

To test whether the elevated gene conservation in gut-adapted genomes seen in Figure 6a is an artifact caused by wide variation in genome sizes amongst non-gut genomes, I repeated the analysis in Figure 6a excluding genome-genome comparisons from different size categories. Similar patterns emerged to those observed in the full dataset (Figure 6c), indicating that differences in the evolution of gene content between gut and non-gut genomes were not simply attributable to trends in genome size. In order to quantify the effects of adaptation to the gut habitat on gene conservation at various phylogenetic distances, and to test whether this difference was significant, genome-genome comparisons were binned into

increments of 0.03 corrected substitutions/site in the 16S rRNA (Figure 6d). This analysis revealed that gut-specialist and gut-cosmopolitan lineages have greater gene conservation for evolutionary distances between 0.06 and 0.18 substitutions/site. However, at distances of < 0.03 16S rRNA substitutions per site (roughly corresponding to the traditional bacterial species boundary, see Figure 8), gut genomes tended to have much lower gene conservation than is present at greater distances. This could reflect increased niche specialization in very closely related gut genomes, or increased convergence in other environments.

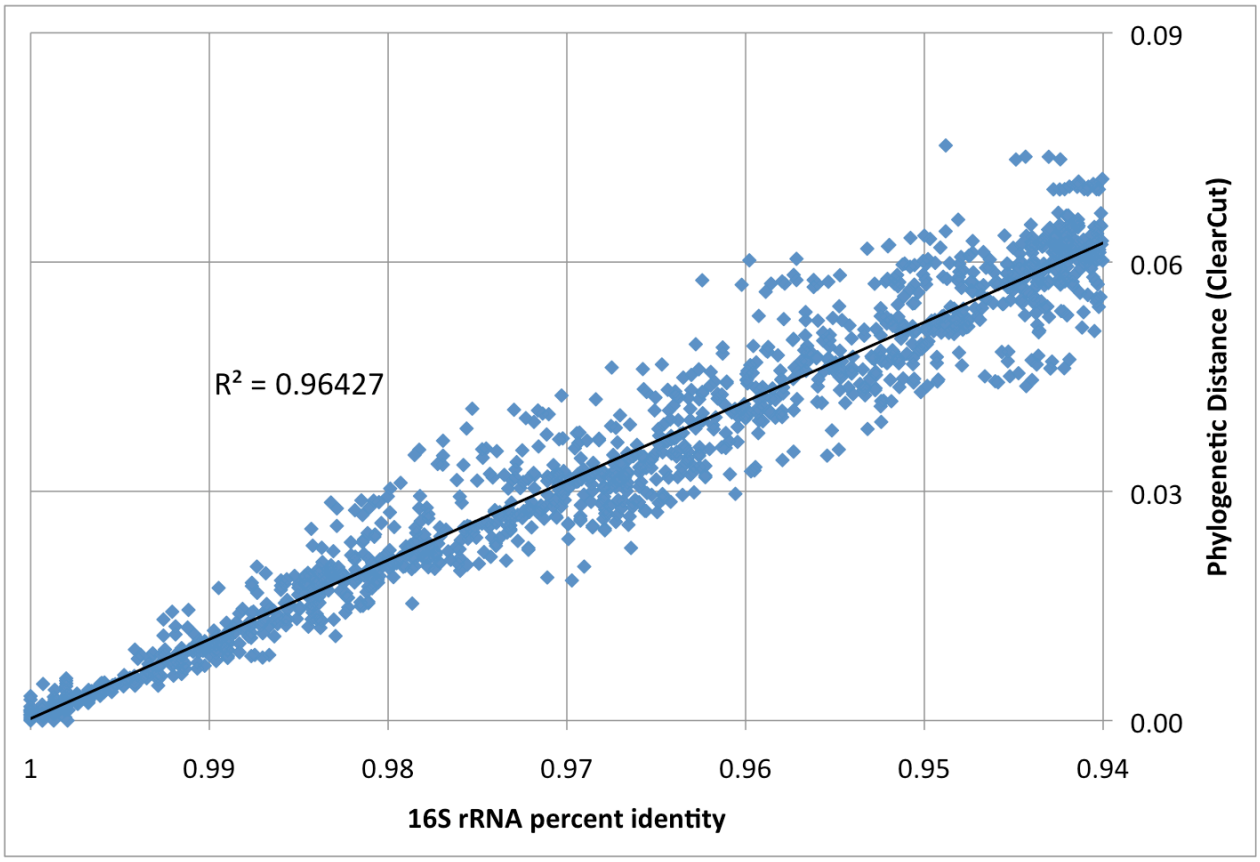
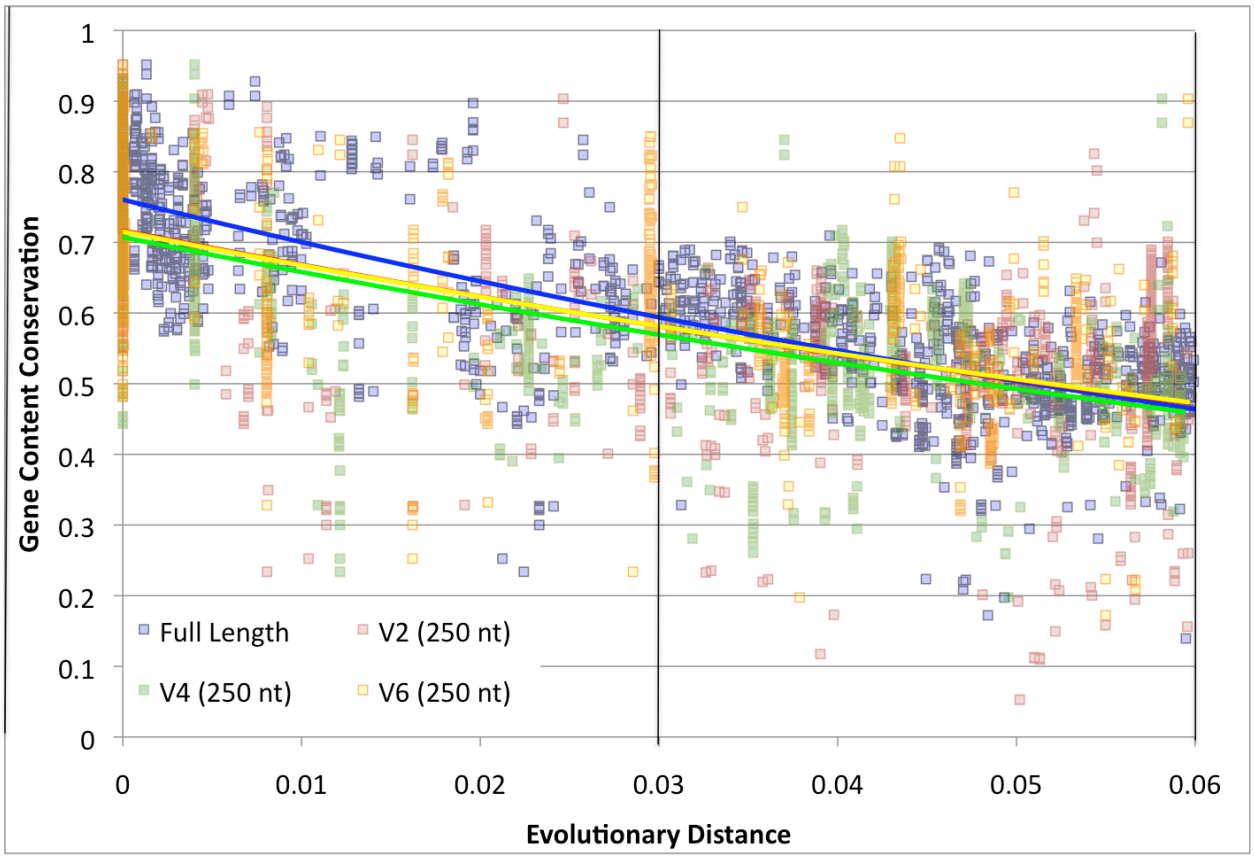


Figure 8. Regression of phylogenetic distance on 16S rRNA distance at short distances.

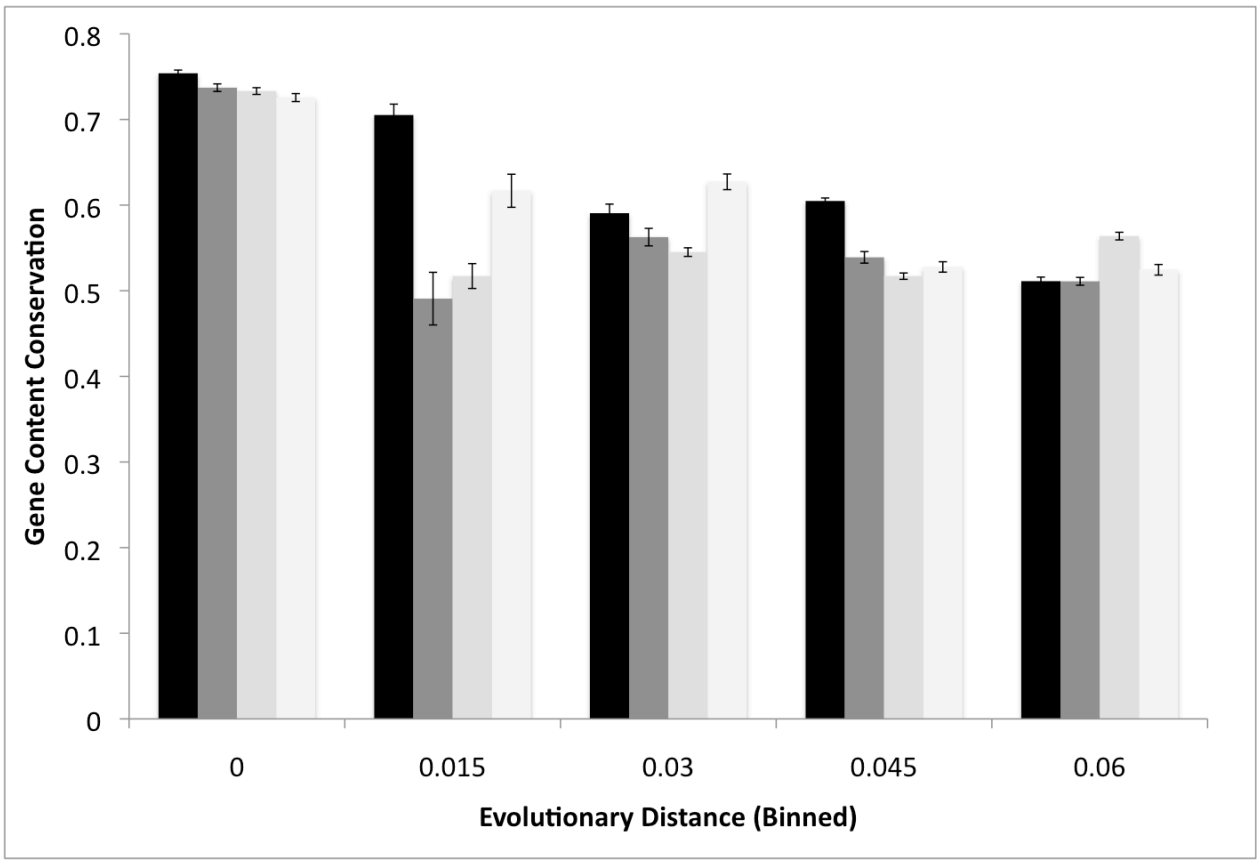
16S rRNA distance predicts genome diversity within bacterial species. Patterns of niche specialization within and between bacterial species may operate according to different principles, which could provide insight into the ecological mechanisms that underlie them within a given habitat. To follow up on this question of niche specialization, I next examined the ability of 16S rRNA distances to predict gene content within bacterial species. This analysis is interesting for two reasons. *First,* because barriers to horizontal gene transfer are believed to be lower between closely related genomes [70], it might be expected that the phylogenetic signal would have little effect on gene content within bacterial species. *Second*, although genome sequencing is increasingly affordable, criteria for choosing strains that maximize divergence in genome content so as to maximize the discovery of new components of the pan-genome are essential. If 16S rRNA distance had little effect on gene conservation within bacterial species, then it would be preferable to select strains based on other criteria, or at random to maximize statistical power.

Even when examining gene conservation at scales that correspond to the most commonly used cutoff for bacterial species (16S rRNA distances below 3% divergence), I found that 16S rRNA gene distance is an important predictor of gene conservation. Gene conservation between strains of the same species fell as evolutionary distances approached 0.03 nucleotide substitutions per site (Figure 9a and b). These results are consistent with those of Konstantinidis and Tiedje, who found a relationship between 16S rRNA divergence, overall gene content, average

nucleotide identity in orthologous genes, and DNA rehybridization kinetics [46].


Gene Content Conservation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.01
0.02
0.03
0.04
0.05
0.06
Evolutionary Distance
Full Length
V2 (250 nt)
V4 (250 nt)
V6 (250 nt)

(A)



(B)

Figure 9. Greater 16S rRNA divergence implies greater divergence in gene content within bacterial species. A) Trees constructed from either the full length 16S rRNA or 250 nucleotide stretches of its V2, V4 or V6 regions . The vertical bar corresponds to the species boundary, using the traditional bacterial species definition of > 97% 16S rRNA identity. (This boundary was determined by regressing the corrected 16S rRNA distances displayed here against 16S rRNA percent identity. See **Figure 8.**) The results demonstrate that even within the same bacterial species, the average gene conservation of a genome pair falls as phylogenetic distance increases. B) Binning the results from Panel A to bins of 0.015 16S rRNA substitutions per site allows quantification of the effects of phylogenetic distance on gene conservation. Black bars represent average gene conservation at a given distance when distances are calculated using the full-length 16S rRNA gene sequence, while progressively lighter gray bars represent gene conservation when calculating distance with fragments of the V2, V4 or V6 regions, respectively.

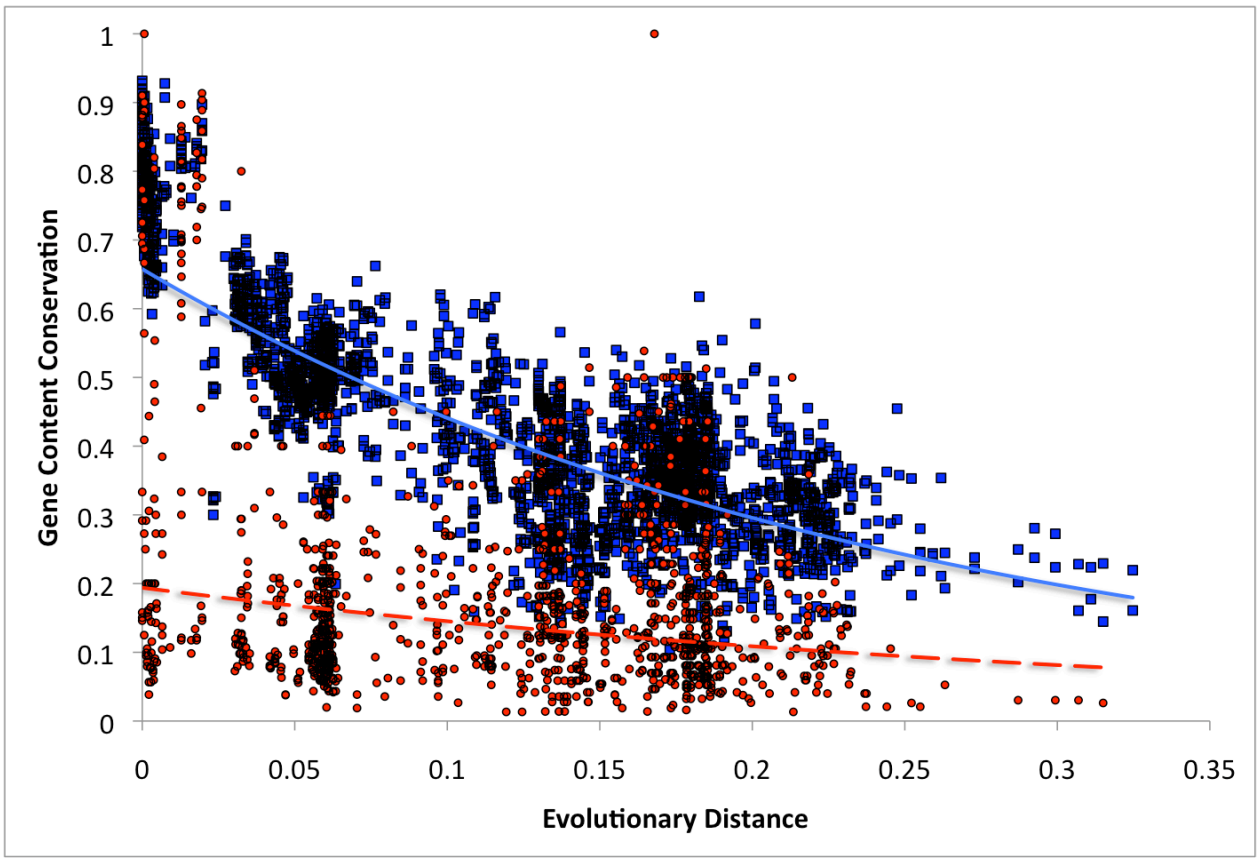
In addition, these trends can be recovered using not just full-length 16S rRNA, but also using 250 nucleotide reads from the V2, V4, or V6 regions of this gene. This result reveals that even short 16S rRNA gene reads, such as those produced with pyrosequencing, are associated with genomic differences (Figure 9). On average, selecting a strain with 16S rRNA distance between 0.015 to 0.03 from the nearest known strain will produce ~ 9% fewer conserved genes (and, conversely, greater gene novelty) than selecting a random genome within the species, while a similar criterion applied to phylogenies constructed from 250 nucleotide reads from V2, V4, or V6 primers will yield an average 17%, 16% or 4% reduction in conserved genes, respectively (Figure 9b). A similar concept applies when selecting species within the same genus (using the >94% rRNA percent identity threshold). Selecting the most divergent strains within a genus (i.e. those with 94-95% percent identity in the 16S rRNA) provides an average 8-12% reduction in gene conservation relative to randomly chosen species belonging to the same genus, depending on the primers used. It should be noted, however, that variation is sufficiently high in either case that this technique is most useful when sequencing a large number of genomes; although choosing divergent lineages at the genus or species level provides access to a pool of strains or species with reduced gene conservation, it is not the case that gene conservation for every genome pair will be reduced.

Habitat adaptation in bacterial plasmids. Bacterial plasmids are frequently subject to horizontal transfer. Because plasmids supplement an existing bacterial

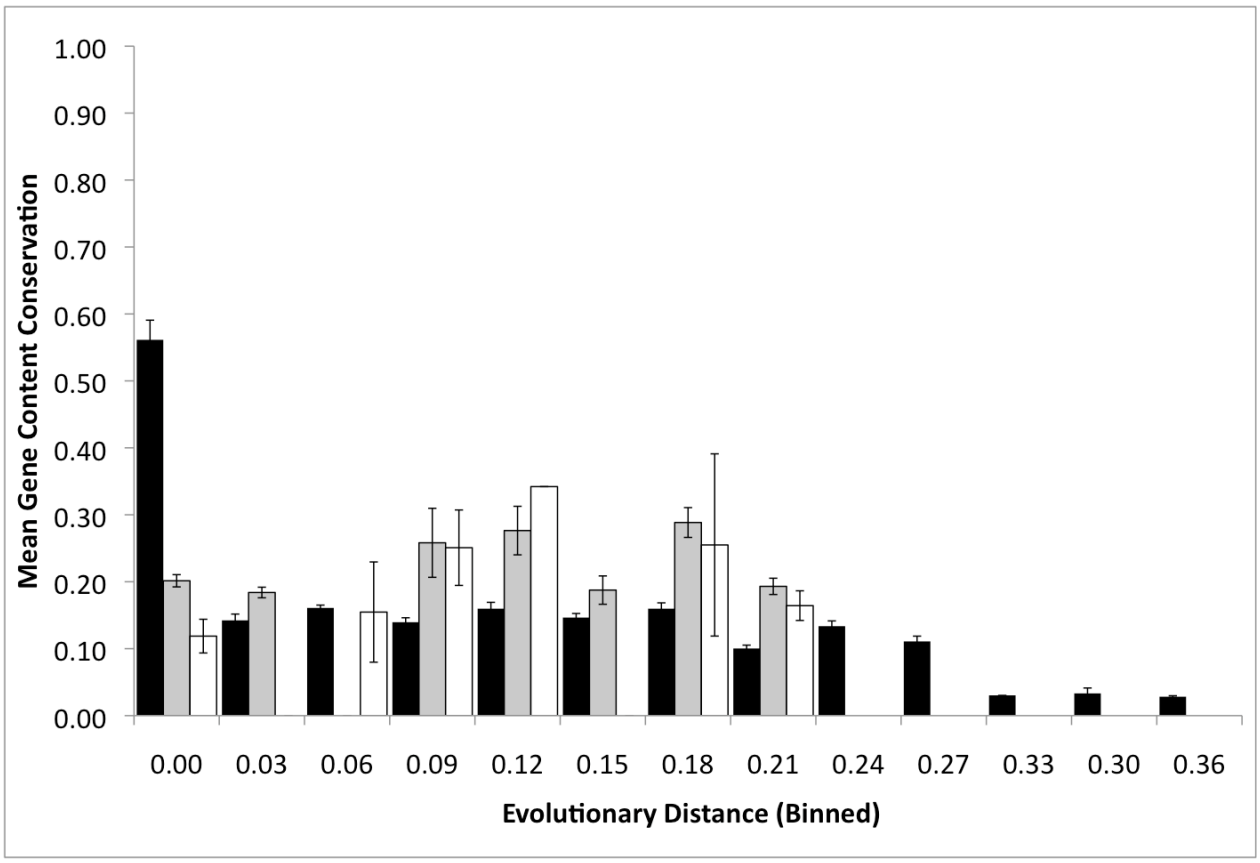
genome, they are not constrained to contain genes essential for cellular life. The 132 plasmids sequenced with the genomes included in this analysis thus provide a window into gene conservation amongst frequently transferred genes. Icompared the genes carried on each plasmid to the combined pool of genes carried on the chromosomes and plasmids of each other isolate in the analysis (Figure 10a). Both overall gene conservation and the ability to predict gene conservation from phylogenetic distance were dramatically reduced in plasmids. This contrast between conservation of plasmid-borne genes and those located on bacterial chromosomes suggests that horizontal gene transfer in genomes is not so frequent that phylogeny and gene conservation are uncoupled (in which case the ability of phylogenetic distance to predict gene conservation would be similar for both plasmids and chromosomes). Instead, once Iaccount for differences in overall genome size, the gene content of chromosomes is substantially more predictable than that of plasmids ($r^2 = 0.60$ chromosomes; $r^2 = 0.06$ plasmids). Surprisingly, despite explaining little of the variation in gene content conservation, the correlation between evolutionary distance and gene content conservation in plasmids is still statistically significant for the taxa in the analysis ($p < 0.05$, Mantel test), except in cases where the number of plasmids is very small ($n < 5$).

Given the observation that the dense bacterial community of the mammalian gut presents ample opportunities for horizontal gene transfer, and horizontal gene transfer is thought to be a process promoting habitat adaptation, Itested whether the effect of environmental adaptation on gene conservation observed in bacterial

chromosomes also occurs on plasmids. The plasmids of gut-cosmopolitan genomes clearly show a similar effect of habitat on gene content to that observed in bacterial chromosomes (Figure 10b). That is, at short phylogenetic distances gene content conservation is reduced for comparisons within the same environment, while at longer phylogenetic distances gene conservation is enriched, suggesting that the same pattern of short range specialization and long range convergence observed for bacterial chromosomes may be acting on plasmids. For gut – specialist plasmids the dataset is limited to a small number of examples, but overall the results appear consistent with the patterns observed for the full chromosomes. Indeed, the effect of habitat on gene content conservation over short phylogenetic distances appears to be even more dramatic in plasmids than in bacterial chromosomes (Figure 10b).


Gene Content Conservation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Evolutionary Distance

(A)



(B)

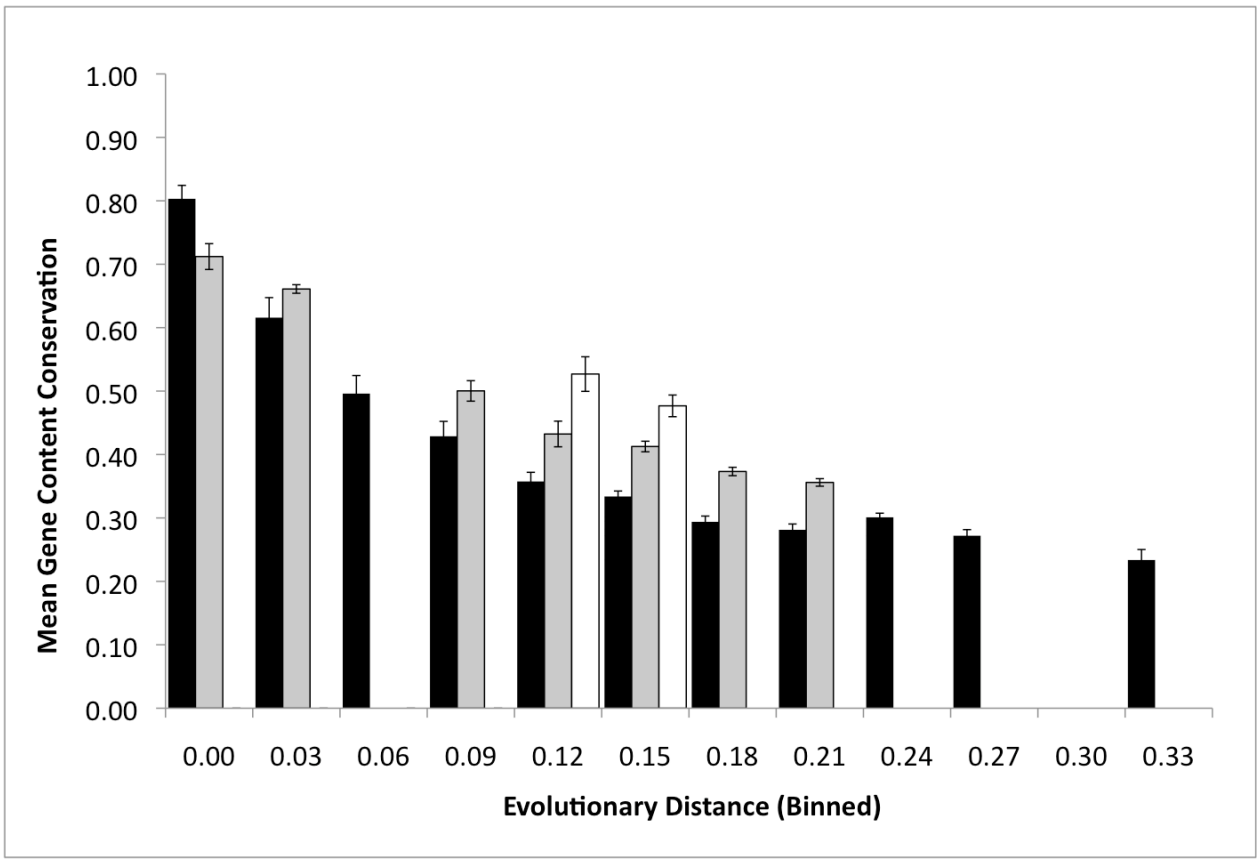
Figure 10. Gene conservation in plasmids borne by gut-adapted bacteria. A) Gene conservation in bacterial chromosomes (blue squares) or plasmids (red squares). Plasmids show both lower average gene conservation than bacterial chromosomes, and, as would be expected given frequent conjugative exchange, a weaker relationship between evolutionary distance and gene conservation (r^2 = 0.60 genomes; r^2 = 0.06 plasmids). B) Plasmids borne by specialist (white bars) or cosmopolitan (gray bars) bacteria tend to have higher gene conservation at evolutionary distances between 0.09 and 0.21 16S rRNA substitutions per site than those borne by non-gut bacteria (black bars). These plasmids also exhibit markedly reduced gene conservation at distances under 0.03 substitutions per site.

The effects of habitat adaptation on gene conservation occur in both pathogens and non-pathogens. Finally, I tested whether the effects of shared habitat, phylogenetic distance and genome content were common across commensal and pathogenic genomes. When I divide the genomes into more categories, the

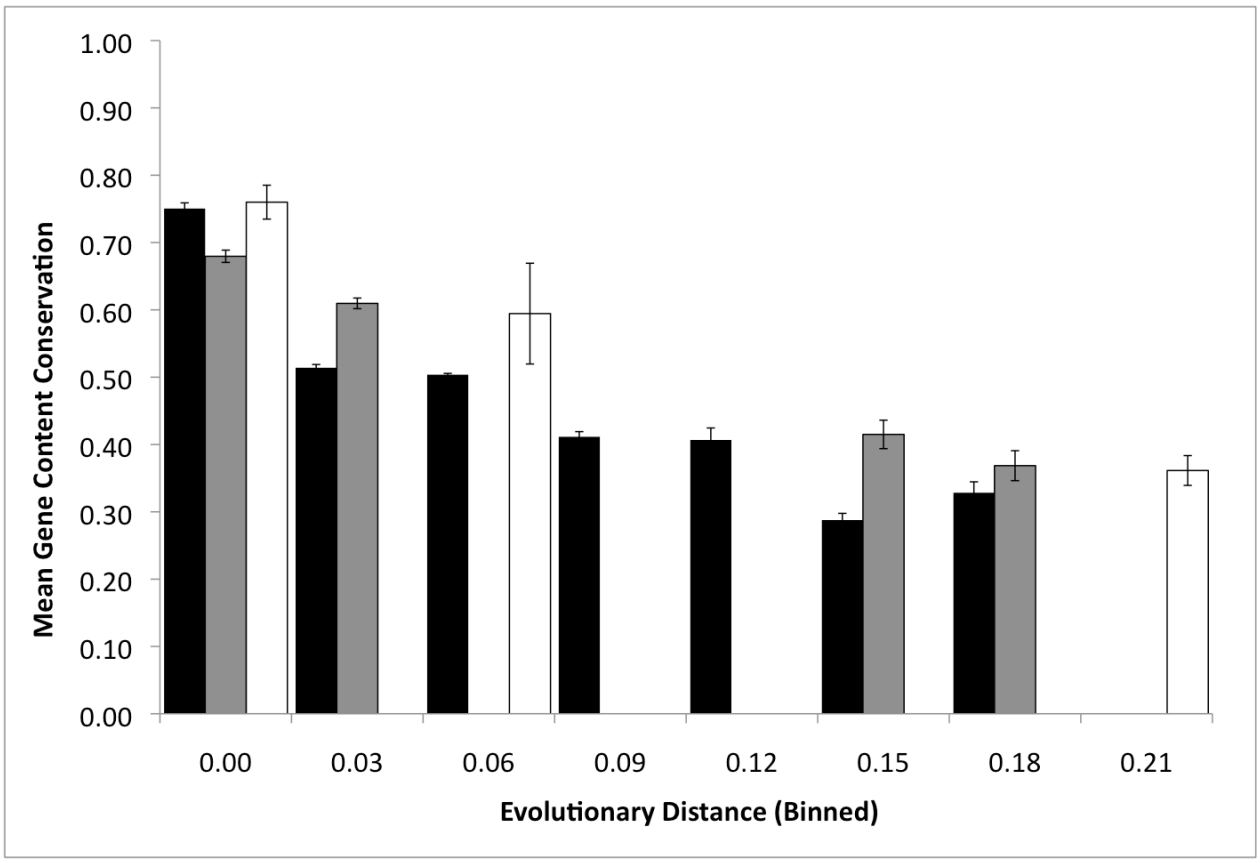
statistical power is reduced, but in cases where data are available gut-adapted commensal (Figure 11**a**) and pathogenic (Figure 11b) genomes generally display the same elevated levels of gene conservation at intermediate phylogenetic distances relative to non-gut genomes. This effect persists when also limiting the data to comparisons between genomes of similar size (Figure 11c and d).

DISCUSSION

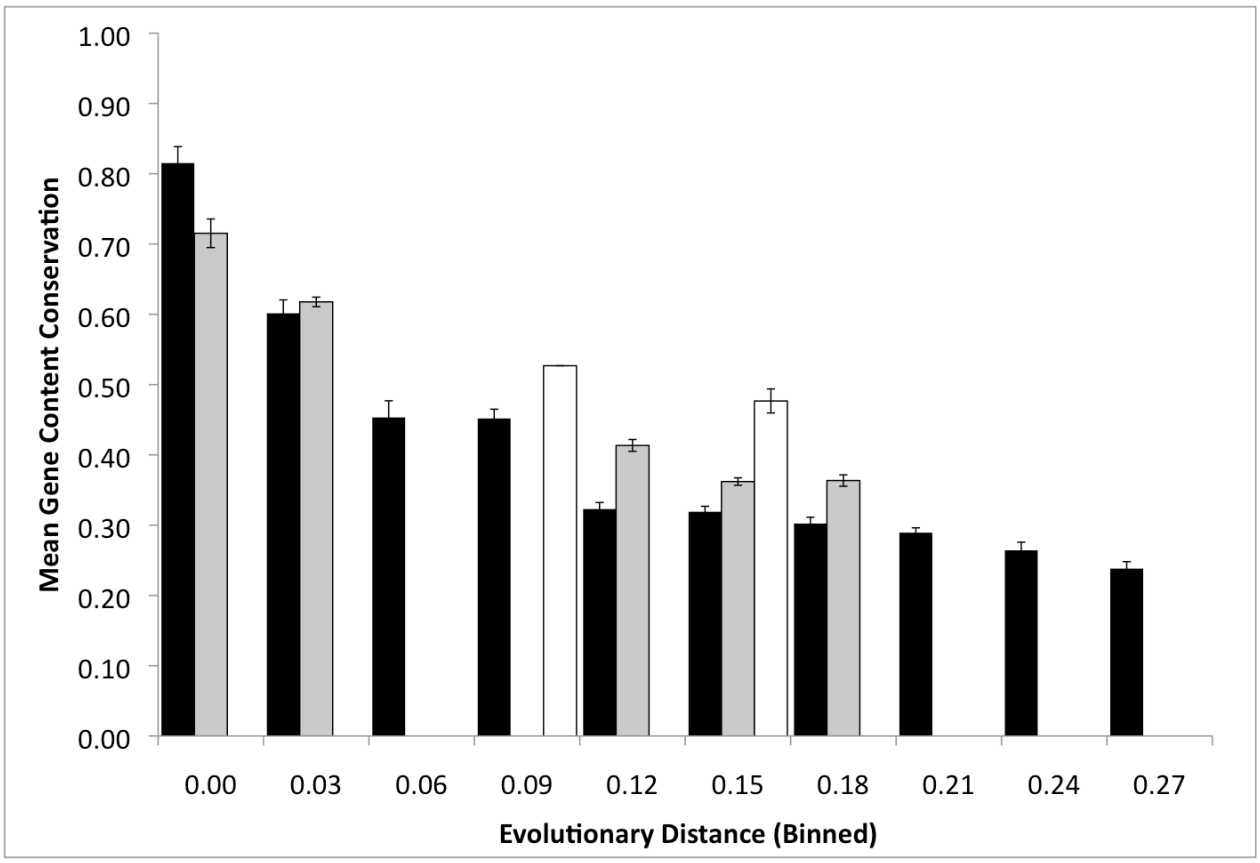
This study reveals that gut-adapted genomes are more similar in gene content at a given evolutionary distance than non-gut genomes. Thus, common functional requirements or increased horizontal gene transfer cause similarities in gene content within the gut habitat. This trend holds over a broad range of phylogenetic distances. However, niche specialization at short phylogenetic distances (e.g. of strains within the same bacterial species) is also important in the mammalian gut. The well-known result that genome content can vary radically for genomes with identical 16S rRNA sequences [45, 71], and studies that report high levels of horizontal gene transfer [72, 73] have raised doubts about our ability to understand genome and community functions based on phylogeny.


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Mean Gene Content Conservation
0.00
0.03
0.06
0.09
0.12
0.15
0.18
0.21
0.24
0.27
0.30
0.33
Evolutionary Distance (Binned)

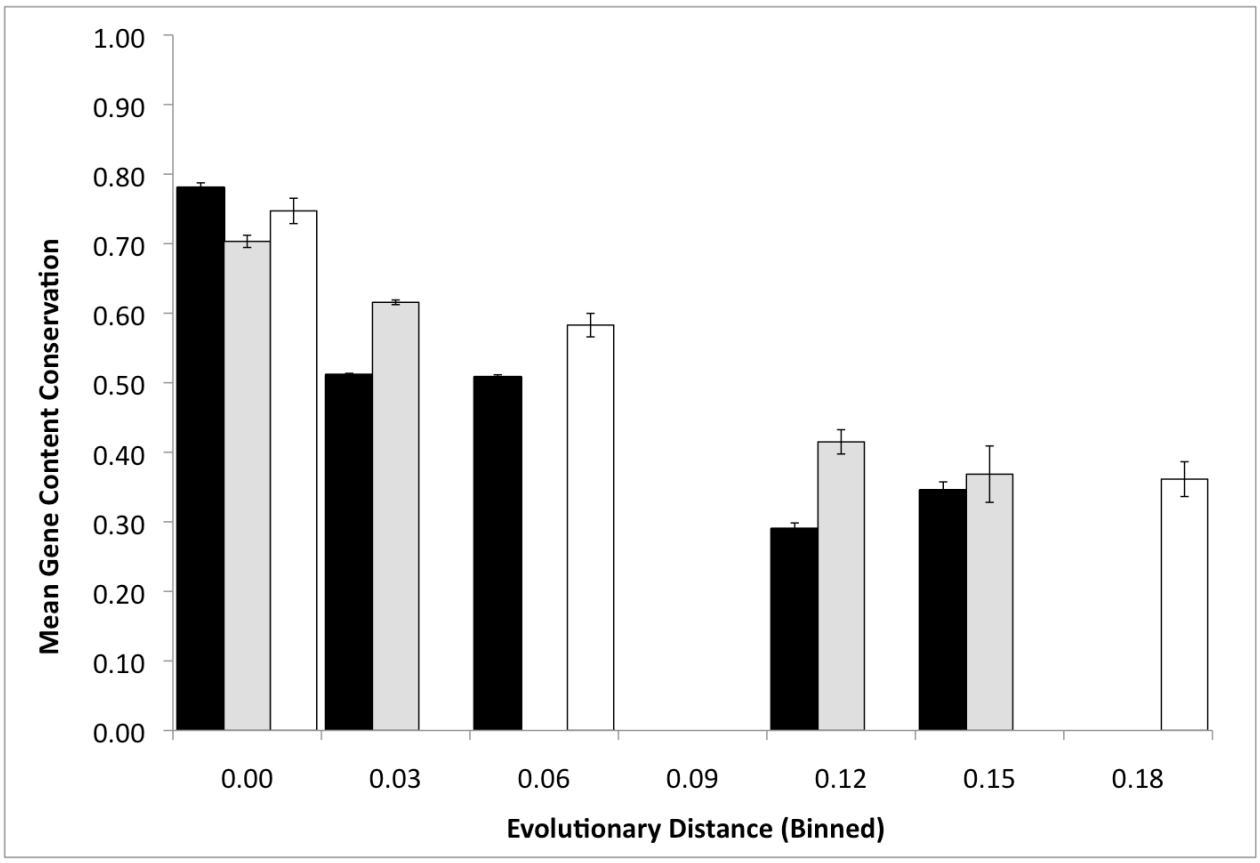
(A)


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Mean Gene Content Conservation
0.00
0.03
0.06
0.09
0.12
0.15
0.18
0.21
Evolutionary Distance (Binned)

(B)


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Mean Gene Content Conservation
0.00
0.03
0.06
0.09
0.12
0.15
0.18
0.21
0.24
0.27
Evolutionary Distance (Binned)

(C)



(D)

Figure 11. Gut pathogens, like gut commensals, exhibit different patterns of gene content conservation from non-gut genomes. Each panel depicts average levels of gene content conservation, binned in ranges of 0.03 16S rRNA substitutions per site. Values for comparisons between pairs of non-gut bacteria are shown in black, pairs of gut cosmopolitan bacteria in gray, and pairs of gut specialists in white. A) Gene conservation in non-pathogens, including comparison between pairs in all size categories. B) As in Panel A, but showing only comparisons between pairs of genomes in the same size category. C) As in A, but for pathogenic bacteria. D) As in Panel B, but for pathogens. Error bars depict the standard error of the mean.

The results presented here, together with the demonstration from GEBA (http://www.jgi.doe.gov/programs/GEBA/) that phylogenetically chosen genomes maximize novel gene lineage discovery, suggest that these effects, while important, do not obscure the overall trend that evolutionarily related organisms tend to share genomic features and, presumably, ecological niches.

The finding that gene conservation between gut-adapted bacteria is reduced over very short phylogenetic distances but elevated at greater distances suggests that gene content filters the persistent lineages of microbes in the gut [74]. The reduced gene conservation at short phylogentic distances might thus indicate that competitive exclusion amongst bacteria with very similar functional profiles dominates amongst closely related bacteria, while the gene content of more divergently related gut bacteria is more strongly influenced by the shared selective pressures imposed by life in the gut. This interpretation is further supported by the convergence of very different species assemblages on similar functional repertoires in the human gut, as revealed by metagenomic studies [75].

A survey of microbial communities across 27 body habitats in healthy individuals has emphasized the importance of body habitat in determining community composition relative to interpersonal or temporal variation [76]. If there is more convergence in function in the gut due to extreme selective pressure and/or

horizontal gene transfer, would this be mirrored by more consistent metagenomic profiles and/or more divergence at fine phylogenetic scales in the gut than in other body habitats? Although difficulties with low sample biomass currently preclude metagenomic studies of these other body habitats, large-scale sequencing of strains associated with other body habitats could address these important questions by allowing the application of the techniques introduced here.

A key and pressing challenge is to understand how, if the gut is such a selective environment, some species are able to establish and maintain a broadly cosmopolitan lifestyle. To that end, it would be profitable to deliberately choose closely related gut and non-gut strains both for sequencing and for careful experiments to test survival across a broad set of conditions and environments where common metabolic themes such as fermentation may be represented. Ideally these would be newly isolated from well-characterized environments, sidestepping the issue of dubious provenance of many existing strains. As these species are being sequenced, our ability to gain insight will improve as annotations converge on improved standards such as Minimal Information about a Genome Sequence (MIGS; [77] and Minimal Information about an ENvironmental Sequence (MIENS; http://darwin.nerc-oxford.ac.uk/gc_wiki/index.php/MIENS). This combination of data and metadata will enable more general tests of the effects of environmental adaptation on genome composition and evolution.

CHAPTER III.

HORIZONTAL GENE TRANSFER AND GENOME EVOLUTION IN THE *METHANOBREVIBACTER SMITHII* PAN-GENOME

Background

In the previous chapter, we saw that habitat adaptation to the gut had important consequences for the evolution of bacterial genomes. These consequences varied by phylogenetic distance: closely related gut bacteria sharing fewer genes as expected (which I hypothesized might be due to niche specialization), and more distantly related gut bacteria sharing a greater proportion of genome contents (which I interpreted as due to convergence in gene content due to the common challenges faced by gut-adapted bacteria). Interestingly, these effects of habitat adaptation on gene presence or absence in bacterial genomes appeared to an even greater extent in bacterial plasmids. Because plasmid-encoded genes are frequently transferred, this suggested that horizontal gene transfer might be an important process by which bacteria adapt to life in the gut. Although the analysis that I presented in the previous chapter dealt exclusively with bacteria, one might also hypothesize that similar processes of habitat adaptation are at work in gut-adapted archaea.

In this chapter, I present work which assesses the extent and functional consequences of horizontal gene transfer on the genome of *Methanobrevibacter*

smithii, an important gut archaeon. The full paper, on which Liz Hansen was lead author, was recently published in *PNAS* as part of a larger analysis of *Methanobrevibacter smithii* distribution, ecology, and genomics. In this subanalysis, I address the role of horizontal gene transfer in shaping the *M. smithii* pangenome, and the extent to which gene transfer explained the differences between strains isolated from different families or individuals. The portion that I discuss here is adapted from supplementary results for the *PNAS* paper, in which I report an analysis of gene transfer conducted by myself, Daniel McDonald, and Julia Goodrich (see [3] for the full analysis). I have also included additional explanatory material drawn from a review of horizontal gene transfer that I wrote for *Microbiology* with Diana Nemergut and Rob Knight as coauthors [4], and a book chapter on detecting horizontal gene transfer using the CodonExplorer web server that I co- first authored with Micah Hamady for *Methods in Molecular Biology* [29].

***M. smithii's* role in the gut.** The hydrogen concentration in the human gut is believed to be an important factor governing the metabolic efficiency of the gut microbiota. Fermentation of dietary polysachharides by the gut microbiota produces short-chain fatty acids (SCFAs), CO_2, alcohols, formate, and H_2. Methanogens are one of three microbial groups that can remove H_2 from the human gut. Other groups include the sulfate reducing bacteria (SRB) and phylogenetically diverse bacterial acetogens. Because high H_2 concentrations can inhibit fermentation of dietary polysaccharides ([3], [78] and reference contained therein), it has been hypothesized that removal of H_2 may play an important role in

promoting the overall efficiency of metabolism by the gut microbiota. Thus, theraputic manipulation of H_2 – removing lineages could potentially be useful in future efforts to develop drugs to regulate the overall metabolic function of the gut community during obesity or malnutrition. Such future efforts, however, require a much greater fundamental understanding of the physiology, ecology, and genetics of H_2 – consuming bacteria. *M. smithii* is a particularly interesting case: both because it is the most abundant archaon in the human gut, and because its threshold for H_2 utilization is lower than that of bacterial acetogens, it may be an efficient H_2-consumer.

Genes involved in *M. smithii* habitat adaptation. A previous comparison [17] of gene functions enriched in the gut-adapted *M. smithii* relative to environmental methanogens and all other archaea identified gene categories that are likely to play a role in *M. smithii* adaptation to the gut. These included the KEGG functional categories for cofactor/vitamin synthesis and central metabolism (both categories include many methanogenesis genes), as well as that for surface variation.

Given the importance of *M. smithii* in modulating the gut ecosystem, and the suggestion from the previous chapter's results that gene transfer may play an important role in microbial adaptation to the gut, I sought to test whether genes involved in *M. smithii* metabolic pathways previously hypthesized to be important for life in the gut had been affected by horizontal gene transfer. Genes involved in surface variation had previously been indentified as enriched in *M. smithii* relative to its non-gut relatives. These surface-variation genes include a novel class of

proteins called adhesin-like proteins (ALPs) [17]. These genes were identified by sequence similarity to bacterial adhesins, and have been hypothesized to play a role in associating *M. smithii* either with the host intestinal wall, or with syntrophic bacterial partners. Because previous steps of the *PNAS* analysis identified these adhesin-like proteins (ALPs) as highly variable in distribution and expression (assayed by RNA-seq) across sequenced *M. smithii* isolates, I also hypothesized that horizontal gene transfer might play a role in shaping the distribution of these proteins across strains. Frequent gene transfer of adhesin-like proteins amongst *M. smithii* strains would also be consistent with the pattern of niche specialization amongst closely related gut specialists inferred from the results presented in the previous chapter.

Methods for HGT detection. In order to test these ideas, I felt that it was important to conduct an analysis of horizontal gene transfer using the most reliable methods available. However, all existing methods have certain drawbacks. The detection of HGT in genomic data is currently accomplished using two main approaches: phylogenetic and compositional. Phylogenetic methods attempt to examine the distribution of, or relationships between, genes in multiple taxa. For example, discordance between the phylogeny of a gene and the species phylogeny may indicate horizontal transfer. Such results should be treated with caution, however, since gene duplication and differential loss or selection can generate similar patterns [79, 80]. Additionally, even in a genome in which all genes share the same phylogeny, the sequence of some genes may not support the true phylogeny.

This effect is especially important where the phylogenetic signal is weak, and thus susceptible to being overwhelmed by random noise.

Techniques related to these phylogenetic methods include distributions of BLAST hits (e.g. [81, 82]), inference of gene presence/absence on species phylogenies (e.g. [83, 84]), and ratios of evolutionary distances (e.g. [85, 86]). These techniques, however, have been shown to detect only a subset of horizontally transferred genes, and therefore should be treated as approximations to, not replacements for, phylogenetic methods [87 88 89]. Compositional methods, in contrast, examine sequence characteristics that vary in different taxa but are relatively conserved within a taxon. If a gene or genomic fragment possesses unusual sequence characteristics, it may have been transferred from a taxon in which those characteristics are typical. Unlike phylogenetic methods, compositional methods do not require alignment of homologous sequences, and are therefore better suited to examine poorly conserved or very rapidly evolving genes.

Sequence characteristics that have been used to study HGT include GC content (in the whole gene or at the third position in each codon), the codon adaptation index (CAI), amino acid usage, relative synonymous codon usage (RCSU), and dinucleotide usage [90-94]. Markov models, including frame-dependent Markov models [72] and variable order Markov models [95] have also been used. Codon usage and other compositional information can be used to detect horizontal gene transfer because different organisms differ in the composition of their genes. Thus, horizontal gene transfer can be detected, if sufficiently recent [91], by looking for genes of unusual composition [72, 96]. However, some caution must be exercised in this

approach since highly-expressed genes can also show codon bias, and gradients in composition can appear along a genome due to replication-coupled biases [97, 98]. One disadvantage of compositional methods is that transfer between related or unrelated strains with similar compositional characteristics will not be detected. Conversely, if a sequence characteristic varies within a genome, then classes of genes, such as ribosomal proteins, that have unusual compositions may appear to have been horizontally transferred. Even when the composition of a transferred gene is initially distinct from the composition of the recipient genome, the composition of the gene will drift toward that of the recipient genome over time. Thus, the traces of ancient transfers may be obliterated in sequences that have had time to equilibrate fully [91].

One of the major remaining challenges in the area of HGT research is to achieve accurate estimates of the overall rate of HGT. Although there are many well-established individual cases of HGT, and general agreement that particular classes of genes that are more or less frequently transferred, estimation of the overall frequency of HGT on the tree of life is a difficult problem and an active area of research. Several recent attempts to estimate the global extent of HGT produced strikingly different results, probably as a result of the different methodologies used [72, 99, 100].

In order to avoid the limitations of any individual horizontal gene transfer detection algorithm applied in isolation (I address this issue at length in [4]), I applied a combination of phylogenetic, compositional, and mobile-element based HGT detection methods to provide a more comprehensive analysis of gene transfer

in each of the 22 *M. smithii* strains sequenced for this analysis.

Methods

Compositional analysis of horizontal gene transfer. For each gene call, compositional statistics were calculated by using the PyCogent code base [60]. The statistics included the GC content at each position, three versions of the dinucleotide use (overlapping, nonoverlapping, or "3-1"), all k-words ranging from length 1 through 6, and codon use. For each *M. smithii* strain, the composition of each gene was compared against (i) the composition of the genome as a whole and (ii) the composition of highly expressed genes. Genes that mapped to the KEGG orthology (KO) groups for ribosomal proteins were used to calculate the highly expressed test set. The gene and control vectors were compared using either the G-test statistic or Pearson correlation. The significance of the results was calculated in two ways; first, the Bonferroni corrected P value was calculated for the G-test; second, because the distribution of compositional counts may violate normality, the method of picking significance thresholds based on the rank order of gene scores of Tsirigos et al. [101] was employed. Because highly expressed genes frequently possess unusual gene compositions, gene transfer was predicted only in cases where the gene did not match the whole-genome model, and the gene also did not match the highly expressed model. Annotated tRNAs and rRNAs were also excluded from the analysis.

Phylogenetic analysis of horizontal gene transfer. Phylogenetic confirmation of gene transfers predicted by compositional means was performed using the RIATA-HGT program of PhyloNet version 1.7 [102]. I obtained all available gene sequences for all KO groups that contained one or more *M. smithii* genes. Annotations for gene family level KEGG assignments were obtained by blasting each protein sequence against version 54 of the KEGG database. The best hit with a KEGG assignment was taken. Multiple assignments were given if the best hit had more than one annotation. Python scripts were used to generate separate FASTA files for each orthology group containing the amino acid sequences for *M. smithii* and KEGG proteins. All sequences for each orthology group were then separately aligned in MUSCLE [102] using maxiters = 4, and gene trees for each group were constructed in FASTTREE [103]. PhyloNet requires that no paralogs be present on protein trees. Therefore, multiple members of a KO present in a single KEGG genome were reduced to a single copy by removing sequences that produced the longest branches on the resulting phylogenetic tree. However, for *M. smithii* genes, I wanted to ensure that the process of paralog resolution did not prevent detection of possible xenologs (extra gene copies introduced by gene transfer). Therefore, all *M. smithii* genes were retained in each gene tree in the analysis. The species tree used consisted of the KEGG 16S rRNA sequences for each lineage in the tree, gathered by BLAST against the E. coli rrsG gene, and alignment in PyNAST. The location of "msi," the M. smithii strain present in KEGG, was taken as the tree position for all *M. smithii*. Because all multiple copies of gene family members were retained in *M. smithii* genomes, it was necessary to introduce an artificial polytomy (with

negligible branch length) into the species tree at the location of *M. smithii* ('msi'), with one tip for each paralog/strain combination. This approach is identical to separately running an analysis of each gene copy, but is computationally more tractable because it avoids re-inferring all transfers not involving *M. smithii* across the rest of the tree many times.

Results and Discussion

To better understand genomic differences among *M. smithii* strains, HGT was detected using both compositional and phylogenetic methods. Compositional HGT detection was performed by examining the typicality of dinucleotides, codons, and k-words of lengths 4 and 6. Because highly expressed genes are known to contain unusual compositions, genes were scored for typicality against both a whole-genome compositional model and a model built using ribosomal proteins [4, 92]. Only genes that were found to be below the significance threshold when compared against both models were annotated as transferred**.** In order to select significance thresholds for gene transfer, genes in each genome were ordered from most to least atypical. As reported previously [101], gene typicality was observed to increase rapidly for the most extreme genes, and then rise only gradually for the rest of the genome (Figure 12a). In this case, thresholds were set at the point where the change between overlapping 30 gene windows was less than 0.1% of the score of the previous window.

Hansen *et al.*, Fig. S7

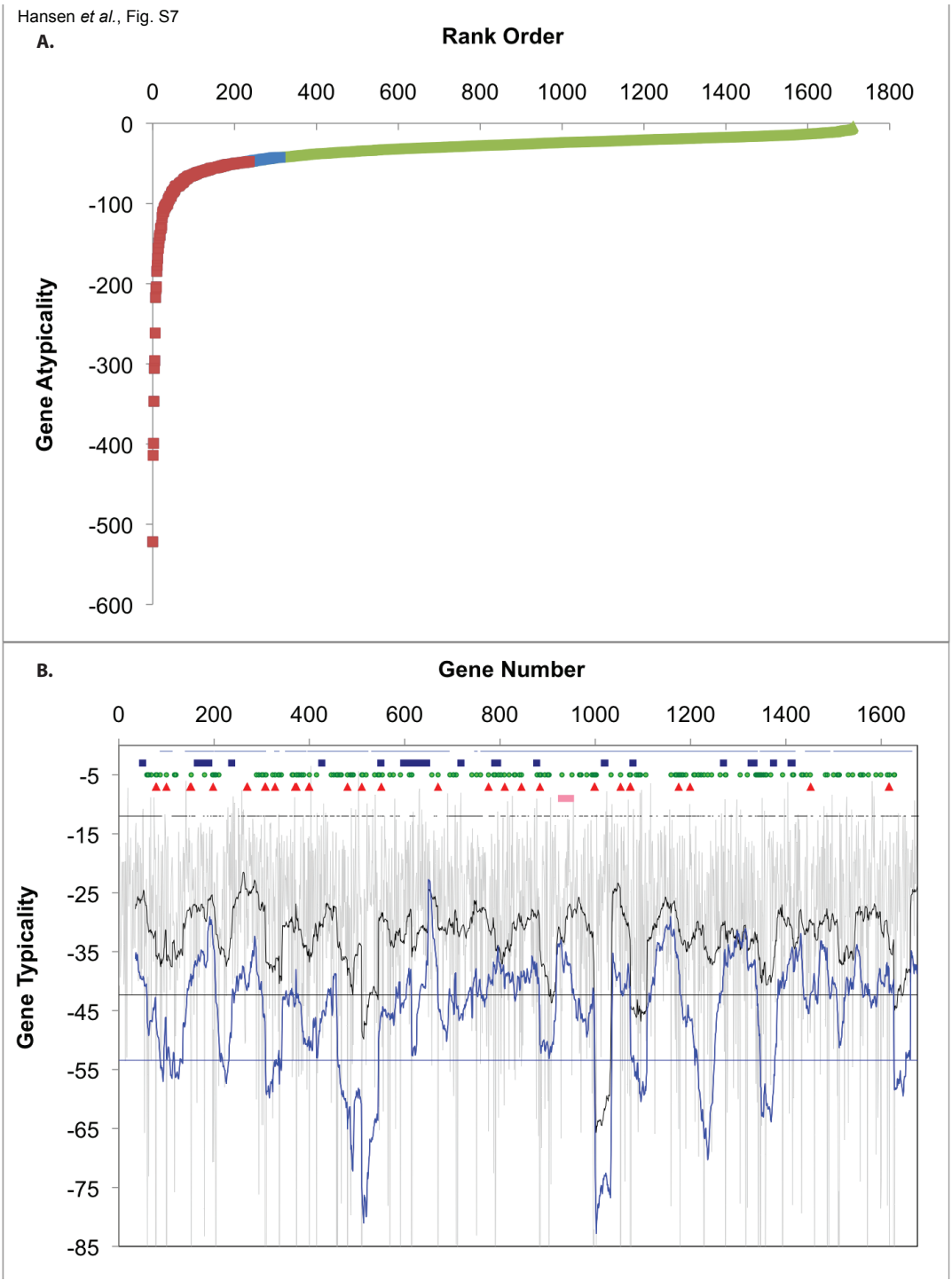


Figure 12. Compositional Analysis of Horizontal Gene Transfer. (A) Threshold for gene atypicality in strain METSMIALI against the whole-genome model. The vertical axis represents the compositional typicality (in this case, of 3-1 dinucleotide usage) of each gene in the genome of the METSMIALI type strain. Scores along the vertical axis represent the G-statistic [made negative so as to represent gene typicality following the convention of Tsirigos *et al.* [101]]. A threshold for the significance of atypical genes has been chosen in two ways: either using a rank order threshold (ref. [101], red points) or by naively assuming a normal distribution and applying the Bonferroni corrected G-test (red plus blue points). In this case, the two methods select similar significance thresholds. (B) Dinucleotide Atypicality in the METSMIALI genome. The colored trendlines indicate differences between gene dinucleotide composition and the composition of either the whole-genome (black line) or ribosomal proteins (blue lines). Each trendline represents a moving average over a 50-gene window. The gray lines show gene typicality for each gene against the whole genome model. In order for a gene to be scored as transferred, the individual gene typicality must be below the significance threshold (horizontal lines) for both comparison sets. Tracks along the top of the graph represent gene annotations; from top to bottom, these are: core genome members (thin blue line), ribosomal proteins (blue squares), horizontally transferred genes (green circles), adhesin-like protein (ALP) genes (red triangles), degenerate prophage (pink bar), and members of the variable genome (thin black line).

Among the compositional measures I analyzed, the proportion of genes defined as horizontally transferred ranged from 3.3% to 10.1% in the dataset as a whole (Table 1**;** see Figure 12b for a graphical representation of the thresholds applied).

Table 1. Compositional evidence of Horizontal Gene Transfer in the *M. smithii* pangenome

Method	Significance Measure	Atypical/Total	Percent
3-1 Dinucleotide	Rank order threshold; G-score	4200/41694	10.1%
3-1 Dinucleotide	Rank order threshold; Pearson correlation	1410/41694	3.3%
Codon Usage	Rank order threshold; G-score	3973/41694	9.5%
Codon Usage	Rank order threshold; Pearson correlation	1675/41694	4.0%
K-words (length 4)	Rank order threshold; G-score	3230/41694	7.7%
K-words (length 4)	Rank order threshold; Pearson correlation	2223/41694	5.3%
K-words (length 6)	Rank order threshold; G-score	2336/41694	5.6%
K-words(length 6)	Rank order threshold; Pearson correlation	3300/41694	7.9%

However, because the absolute number of horizontally transferred genes predicted can depend on the compositional measure chosen, the stringency of the thresholds selected, the amount of time that has passed since the transfer occurred, and the compositional distinctiveness of gene transfer donors ([91]; reviewed in [4]), I do not

focus my analysis on the absolute magnitude of gene transfer in these lineages. Instead, I was primarily interested in differences in the frequency of HGT events for different classes of genes, and how this process has contributed to the evolution and specialization of the characterized *M. smithii* strains.

HGT has contributed to both the core and variable components of the *M. smithii* pan-genome

When using compositional methods of horizontal gene transfer detection, I observe that gene transfer is more frequent in the *M. smithii* variable genome than the core. For example, when examining 3-1 dinucleotide use [92] and using the rank order of G scores as the significance threshold, 5.7% of the core genes in the pan-genome show compositional evidence of transfer, compared with fully 16.4% of the variably represented genes, suggesting an approximately three-fold enrichment of gene transfer in the variable relative to the core components of the pan-genome. However, others have observed that phylogenetic methods tend to detect more ancient transfer events than compositional methods [89]. Consistent with these observations, 73% of the genes for which PhyloNet found evidence of HGT were part of M. smithii's core genome, indicating transfer before the divergence of strains. By contrast, most putative HGT events predicted by compositional methods were part of the variable genome (59.3–68.0% of transfers, depending on the method) (Table 2).

Table 2. Distribution of HGT genes in the *M. smithii* core or variable genome by detection method.

Category*	Variable Genome		Core Genome	
	Genes	%	Genes	%
Codons	2695	67.8%	1278	32.2%
Codons (with KO mappings)	816	46.6%	935	53.4%
Dinuc 3-1	2858	68.0%	1342	32.0%
Dinuc 3-1 (with KO mappings)	756	42.5%	1023	57.5%
K-words order 5	1386	59.3%	950	40.7%
K-words order 5 (with KO mappings)	418	32.2%	879	67.8%
PhyloNet	1333	26.0%	3790	73.4%
PhyloNet and codons	174	54.5%	145	45.5%
PhyloNet and dinuc 3-1	146	45.9%	172	54.1%
PhyloNet and kwords order 5	114	40.7%	166	59.3%
Phage	17	10.9%	139	89.1%

*Categories listed as 'with KO mappings' represent the subset of the pan-genome that could be mapped to KEGG orthology groups.

This difference may be due in part to the requirement of phylogenetic methods for orthologs of the gene under investigation: compositional HGT predictions for the subset of genes that could be mapped to KEGG orthology groups were also biased toward the core genome. Genes with both compositional and phylogenetic evidence of transfer tend to be more evenly split between the core and variable genomes than transfers supported by either type of evidence alone (Table 2). Taken together, these findings suggest that gene transfer has shaped both the core genome of *M. smithii* and differences between strains. External evidence further supports a role for HGT in shaping the core genome of *M. smithii*: 89.1% of genes within prophage (as detected by PhageFinder) are part of the core genome (Table 2).

Functional contribution of horizontally transferred genes to the *M. smithii* pangenome. Due to the plastic nature of many bacterial and archaeal genomes, such genomes are often divided into two components: core genes, which are universally conserved in a given taxon; and variable genes, which may vary between different members of a taxon. Taken together these two components comprise the pangenome.

To test for differences in the functions contributed to the *M. smithii* pangenome by the core genome, variable genome, or horizontally transferred genes, each of these three gene sets were annotated to KEGG pathways (level 2). The *M. smithii* core genome is enriched in genes involved in translation while being

depleted in membrane transporters and unclassified metabolic genes (Bonferroni-corrected G-test for significance; $P < 0.001$). The variable genome is enriched in genes for membrane transporters, glycan biosynthesis and metabolism, and genes whose functions are poorly characterized, while being depleted for genes involved in translation (Bonferroni-corrected G-test; $P < 0.001$). Horizontally transferred genes, regardless of the detection method used, are most divergent from the pan-genome in their functional profile than either the core or variable components of the *M. smithii* pan-genome (assayed by UPGMA clustering of Euclidian distance between KEGG level 2 category counts). This finding suggests that gene transfer has contributed significant functional diversity to *M. smithii*. To understand in more detail the specific categories of genes that have been most frequently transferred, significant HGT results for 3-1 dinucleotide use were pooled across genomes and categorized according to KEGG pathway and KEGG orthology group, weighting genes with multiple pathway annotations on a per gene (rather than per annotation) basis (Table 3). Note that this weighting procedure (which corrects for over- or under-annotated genes by weighting each gene equally) accounts for the fractional counts in Table 3.

As previously observed for genomic islands [24], genes of unknown or poorly characterized function dominated the HGT pool. Among genes with known KEGG level 2 pathway annotations, those in the KEGG category for folate biosynthesis were the most frequently transferred (101.7 normalized annotations). Tetrahydromethanopterin (THMP) methyltransferase genes were the most frequently transferred KEGG orthology (KO) within this group (23 putative HGT

events for the D subunit). THMP methytransferase [104] participates in both the methanogenesis and folate biosynthesis pathways by transferring a methyl group from 5-Methyl-THMP to coenzyme-M. Genes involved in coenzyme-M recycling during methanogenesis were similarly frequently transferred, including methyl-coenzyme M reductase α subunit (EC 2.8.4.1; 23 annotations), and heterodisulfide reductase subunit a (EC 1.8.98.1; 22 annotations). Other frequently exchanged KEGG pathway functions included PST-family polysaccharide transporters (50.5/52.5 normalized annotations were compositionally atypical, representing a 5.3-fold enrichment in the putative HGT pool). Phylogenetic analysis of HGT revealed similar trends. Genes involved in the KEGG folate biosynthesis pathway are the second most frequently transferred functional class (after unclassified metabolic genes). Methanogenesis genes are also among the most abundant transferred functional classes (rank order 22/173 classes). As in the analysis of genes with atypical dinucleotide compositions, phylogenetic HGT detection found transfer in KO groups involved in methyl-coenzyme M recycling, including those for THMP methyltransferase A, B, and C subunits (EC 2.1.1.86), methyl-coenzyme M reductase system component A2, and heterodisulfide reductase (B and D subunits) (EC 1.8.98.1). Although some functional categories such as 'immune system' or 'neurodegenerative diseases' may appear of special interest, such annotations are a frequent side effect of genes that have vertebrate homologs (or are simply misannotated in KEGG). In general many papers do not report these categories, but I include them here 'as is' for completeness.

Table 3. KEGG functional categories of genes with compositional evidence of horizontal gene transfer.

KEGG Pathway	Normalized Compositionally Atypical Genes in pathway*	Percent	All genes in pan-genome in pathway	Percent	Fold Enrichment
Unclassified; Poorly Characterized	214.7	12.1	3067	13.0	0.93
Metabolism; Metabolism of Cofactors and Vitamins	201.5	11.3	2395	10.1	1.12
Unclassified; Cellular Processes and Signaling	197.4	11.1	1031	4.4	2.54
Genetic Information Processing; Replication and Repair	186.9	10.5	1259	5.3	1.97
Unclassified; Genetic Information Processing	142.7	8.0	1918	8.1	0.99
Environmental Information Processing; Membrane Transport	132.9	7.5	1268	5.4	1.39
Unclassified; Metabolism	124.8	7.0	1881	8.0	0.88
Metabolism; Carbohydrate Metabolism	75.2	4.2	1371	5.8	0.73
Metabolism; Nucleotide Metabolism	70.0	3.9	1237	5.2	0.75
Metabolism; Glycan Biosynthesis and Metabolism	61.5	3.5	298	1.3	2.74
Metabolism; Enzyme Families	59.6	3.4	402	1.7	1.97
Metabolism; Amino Acid Metabolism	58.4	3.3	1981	8.4	0.39
Metabolism; Energy Metabolism	56.6	3.2	963	4.1	0.78
Environmental Information Processing; Signaling Molecules and Interaction	52.4	2.9	78	0.3	8.89
Genetic Information Processing; Folding, Sorting and Degradation	24.3	1.4	384	1.6	0.84
Metabolism; Xenobiotics Biodegradation and Metabolism	20.8	1.2	516	2.2	0.53
Metabolism; Metabolism of Other Amino Acids	17.2	1.0	269	1.1	0.85
Cellular Processes; Cell Motility	14.9	0.8	57	0.2	3.50
Human Diseases; Infectious Diseases	10.8	0.6	69	0.3	2.09
Environmental Information Processing; Signal Transduction	10.1	0.6	119	0.5	1.12
Genetic Information Processing; Translation	8.7	0.5	2010	8.5	0.06
Cellular Processes; Transport and Catabolism	7.9	0.4	34	0.1	3.09
Genetic Information Processing; Transcription	7.9	0.4	382	1.6	0.28
Organismal Systems; Immune System	7.4	0.4	23	0.1	4.38
Human Diseases; Neurodegenerative Diseases	6.7	0.4	53	0.2	1.69
Organismal Systems; Excretory System	2.8	0.2	21	0.1	1.77
Metabolism; Biosynthesis of Polyketides and Terpenoids	1.7	0.1	292	1.2	0.08
Organismal Systems; Environmental Adaptation	1.2	0.1	6	0.0	2.80
Organismal Systems; Circulatory System	1.0	0.1	3	0.0	4.80
Metabolism; Lipid Metabolism	1.0	0.1	246	1.0	0.05
Metabolism; Biosynthesis of Other Secondary Metabolites	0.2	0.0	135	0.6	0.02

*Genes shown are atypical in 3-1 dinucleotide usage

Evidence for large-scale horizontal gene transfer of adhesin-like proteins.

In addition to characterizing KEGG functional categories, I analyzed ALP gene transfer given their proposed importance in *M. smithii* niche specialization. Because the vast majority of ALP genes could not be assigned to KEGG orthology groups, only a small subset could be tested for gene transfer by using phylogenetic methods. Of the ALPs that could be assigned to KO groups, 6/49 (12.2%) were classified as being horizontally transferred using phylogenetic techniques. When analyzed compositionally, 5 or 6 of 6 of these ALPs were compositionally atypical in dinucleotide use, codon use, and k-words of length 4 or 6. Remarkably, I found that in the full pool of 854 ALP OGUs (operational gene units), between 52% and 65% show evidence of transfer across a variety of compositional measures, an enrichment of 6.4- to 9.3-fold when normalized to the overall levels of gene transfer predicted by the same methods (Table 4).

Table 4. Compositional evidence for horizontal transfer of *M. smithii* ALP genes.

Method	Significance Measure	Atypical/total	Percent	Fold enrichment*
3-1 Dinucleotide	Rank order threshold; G-score	558/ 853	65%	6.4
Codon Usage	Rank order threshold; G-score	538/853	63%	6.6
K-words (length 4)	Rank order threshold; G-score	525/853	62%	8.1
K-words (length 6)	Rank order threshold; G-score	445/853	52%	9.3

*Fold enrichment is relative to the percentage of HGT predicted by a given compositional measure for the *M. smithii* pangenome as a whole.

ALPs that could be mapped to KO groups were less compositionally atypical than

ALPs as a whole (only 30.6– 36.7% were compositionally annotated as transferred for this subgroup). Despite the observation that these genes are highly expressed in *M. smithii* strains, the ALPs annotated as possessing compositional evidence of transfer do not match the model for ribosomal proteins in their genome, meaning that their expression level alone does not account for their compositional atypicality. Large-scale HGT of ALPs would be consistent with their variability among strains, and suggests that gene transfer may help to tune ALP repertoires in individual *M. smithii* strains to promote adaptation to local conditions. Further investigation is needed to test for the functional consequences of ALP repertoire alteration in *M. smithii* strains.

Conclusions

In this chapter we saw that both the conserved core genome of *Methanobrevibacter smithii*, as well as genome components that vary between strains isolated from diffent families (and family members), have been impacted by horizontal gene tranfer. Genes in pathways involved in methanogenesis appear to have been frequently subject to horizontal gene transfer, suggesting that this key feature of methanogen biology may have been modified via lateral transfer. Finally, adhesin-like proteins in the *M. smithii* pangenome display extremely high levels of compositional atypicality (from both the genome as a whole and highly-expressed ribosomal proteins) across a variety of different measures. This suggests that this class of genes may be frequently exchanged.

Adhesin-like proteins are believed to play a role in cell-cell adhesion, suggesting the hypothesis that variability in these proteins may provide a mechanism for *M. smithii* to control it's niche within the host, or physical aggregation with syntrophic, H_2 –producing bacteria ([3] contains additional evidence from cooccurance analysis that *M. smithii* forms such associations) .

CHAPTER IV

HIGH-THROUGHPUT STUDIES OF MICROBIAL HABITAT ADAPTATION: PRINCIPLES AND PROGRESS

Setting the stage for high-throughput studies of habitat adaptation

As we have seen in Chapters II and III, a combination of phylogenetic and comparative genomic analysis can yield powerful insights into microbial evolution. In this chapter, I review recent high-throughput studies of microbial habitat adaptation, highlighting techniques that have proven useful across several studies, as well as common obstacles that are frequently encountered. The text is derived in part from a submitted first-authored manuscript submitted to *Current Opinion in Microbiology* (see acknowledgements for a list of all co-authors).

We live in a world suffused with microbial life. Universal trees of life[33, 105] constructed by a variety of methods unambiguously show that microbial bacteria, archaea, and eukaryotes constitute the vast majority of life's diversity. These

diverse organisms perform many important ecological functions across a wide range of natural and man-made environments: photosynthesis in the world's oceans[106]; nitrogen fixation and provision of carbohydrates in association with plant roots[107]; even modification of the chemistry of the upper atmosphere by communities in droplets of cloud-water[108]. The bodies of animals are also colonized internally and externally by microorganisms, which play crucial roles in the development[109], homeostasis[110], and behavior[111-113] of their hosts.

How have bacteria, archaea, and microbial eukaryotes adapted to survive and thrive across such a range of lifestyles and habitats? Understanding the relationship between microbial genome sequence and fitness in a given environment is both a fundamental question in evolutionary biology, and a matter of great societal importance. As we seek to gain a predictive understanding of phenomena such as the emergence (or reemergence) of pathogens[114], the impact of human activities from agriculture to the combustion of fossil fuels on ecosystems, or the effects of dietary or medical interventions on human health (e.g. administration of anti- or probiotics), accurate descriptions of the mechanisms by which microorganisms have adapted to environmental changes in the past will provide critical guidance.

Traditionally, questions of microbial habitat adaptation have been addressed by experimental manipulation of microbes in pure culture, or by comparisons of genome sequences. More recently, however, large decreases in the cost of sequencing have allowed such approaches to be complemented by the collection of

unprecedented quantities of 16S rRNA[115], metagenomic[116, 117], transcriptomic[118], and whole-genome data[119]. The 'microbial data deluge' has spurred the development of new computational tools, and has also made possible systematic study of large-scale processes such as habitat adaptation in ways that would have been previously intractable. Here I highlight how the increasing availability of sequence data from diverse environments is allowing researchers to systematically explore questions about the evolution of habitat adaptation in microbial genomes. I emphasize current trends in the use of tools and analytical approaches, highlighting those that have recently been applied to yield novel insights into this question (Table 5), as well as the outstanding methodological challenges that remain to be overcome.

Table 5. Links to software and resources discussed in the text.

Category	Title	Description	Link References	
Ordination	QIIME	Tool for the analysis of community diversity in Python	http://qiime.sourceforge.net/	32
	Vegan	R package containing several ordination methods, along with other tools.	http://cc.oulu.fi/~jarioksa/softhelp/vegan.html	-
	Mothur	Tools for community diversity analysis	http://www.mothur.org/	120
Ancestral State Reconstruction	PAML	A multipurpose and widely used tool for evolutionary analysis after treebuilding, including ancestral state reconstruction.	http://abacus.gene.ucl.ac.uk/software/paml.html	121
	EREM	A tool for ancestral state reconstruction of gene presence/absence	http://carmelab.huji.ac.il/software/EREM/erem.html	122
	Ape	A phylogeny package for R, that includes functions for estimation of ancestral states.	http://ape.mpl.ird.fr/ape_features.html	123
	Mr. Bayes	A classic program for Bayesian phylogenetic inference.	http://mrbayes.csit.fsu.edu/	124
	Mesquite	An extensive graphical suite for phylogenetic analysis.	http://mesquiteproject.org/mesquite_folder/docs/mesquite/whyMesquite.html	-
	BEAST	A tool for Bayesian phylogenetic inference.	http://beast.bio.ed.ac.uk/Main_Page	125
Phylogenetic Comparative Measures	Ade4	Classical multivariate analysis R package, including methods for phylogenetic comparative measures.	http://pbil.univ-lyon1.fr/ADE-4/home.php?lang=eng	126
	Adephylo	R package; Describes phylogenetic signal present in data	http://cran.r-project.org/web/packages/adephylo/index.html	127
	Picante	R package containing phylogenetic comparative methods, as well as ordination techniques	http://picante.r-forge.r-project.org/	128
HGT detection	PhyloNet	A memory-efficient tool for phylogenetic HGT analysis.	http://bioinfo.cs.rice.edu/phylonet/	102
	AnGST	(analyser of gene and species trees)	http://almlab.mit.edu/angst/	129
	DarkHorse	A distribution based HGT detection tool, with a database of precalculated results for many genomes	http://darkhorse.ucsd.edu/	130, 131
	Phangorn	Package for the phylogenetic	http://cran.r-	132

		analysis of horizontal gene transfer	project.org/web/packages/phangorn/index.html	
Metadata curation	MG-RAST	Analysis, comparison, and metadata curation for metagenomic sequences	http://metagenomics.anl.gov/	[133]
	QIIME-DB	A web server for running analysis of community diversity, backed by a large database of well-annotated samples.	http://www.microbio.me/qiime	-
	EMP submission portal	Submission portal for the earth microbiome project	http://www.microbio.me/emp	-
	GOLD	Manually curated metadata for genome and metagenomic sequences; accessible by HTML.	www.genomesonline.org	[134]
Metadata Standards	MIMARKS	The Minimal Information about a MARKer gene Standard	http://www.gensc.org/gc_wiki/index.php/MIMARKS	-

High-throughput studies of microbial habitat adaptation

It is now well established that the distribution of microbial organisms across different environmental conditions is correlated with their phylogeny, both in terms of the beta diversity of microbial communities[39, 60], and the habitat range of individual lineages[135]. Some of the best established types of habitat adaptation include reduced genome size in intracellular endosymbionts[136] , increases in genome size and prevalence of two-component regulators in cosmopolitan organisms[137], increase in acidic amino acids as a response to salinity [138 139 140 141 142], and increased rRNA copy number in fast-growing, highly competitive organisms[143-146].

Genome Reduction

Genome reduction is one of the best-studied examples of genome evolution as a habitat adaptation in microbial organisms. Genomic minimalism is typically associated with organisms living in a host-associated environment, either as endosymbionts or obligate parasites (e.g. [147, 148]), where increasing reliance on the host leads to loss of numerous pathways. The reduced genomes of the insect symbionts *Buchnera* (450 kb) and *Carsonella* (160 kb) have lost many biosynthetic pathways, but retain genes for amino acid biosynthesis, which forms the basis for their relationship with the host [147]. The extent of genomic reduction tends to

increase as the length of the obligate relationship with the host increases, with the greatest reduction seen in the mitochondria and plastid organelles that have been stably incorporated in eukaryotic cells for more than 1 billion years and contain only a handful of genes [149]. Organelles also provide the most extreme example of eukaryotic genome reduction, in this case in the secondary plastids, which were acquired by acquisition of a eukaryotic alga. Two lineages with secondary plastids, cryptophytes and chlorarachniophytes, still retain a relict nucleus of the secondary red or green algal symbiont called a nucleomorph that has undergone extreme genome reduction, and appears to be on a path to complete loss[150].

The genomic trajectory of obligate intracellular parasites has followed a similar reductive path, with extensive loss and/or reduction in biosynthetic pathways that corresponds to an increased reliance on the host[148]. Many eukaryotic lineages have undergone large scale genomic streamlining when they become obligate parasites[151]. The most extreme example is in microsporidia, a lineage of highly reduced fungi that are obligate intracellular parasites of diverse animals. The microsporidian *Enterocytozoon bieneusi*, an enteric pathogen in humans, has even lost the ability to synthesize its own ATP and instead has transporters to import ATP from its host[152]. Genomic reduction has also occurred in highly abundant free-living bacteria (e.g. *Pelagibacter ubique* and *Prochlorococcus*), where selection for rapid reproduction has presumably selected for streamlining of genomic content[153, 154]. These organisms retain most of the biosynthetic pathways, but selection pressure in *Pelagibacter* has resulted in very short intergenic regions and eliminated redundant regions of the genome[153, 154].

Increasingly, research into microbial habitat adaptation is successfully leveraging the availability of publically available genome, marker gene, and metagenome sequence data to contextualize new findings. Specifically, recent studies of microbial co-occurrence[155], habitat adaptation[2, 156], survival strategy[143], and genome evolution[129, 157] have converged on related strategies, and faced common challenges. Based on this recurrent trend in recent high-throughput studies of microbial habitat adaptation, I discuss at length the generalized workflow for comparative analysis presented in Figure 13. I discuss the challenges involved in matching sequenced genomes to habitat assignments, determining which environmental parameters are most likely to be relevant for an analysis, separating the effects of habitat adaptation from those of shared evolutionary history, and detecting horizontal gene transfer (Figure 13).

Challenges in defining environment

In order to understand how microbial genomes change in response to environmental adaptation, we need an operational definition for environment, and a way to relate microbial genomes present in the analysis and the environments to which they are adapted (Figure 13**)**. There are many unresolved debates regarding the definition of environment and the relevant spatial and temporal scales of sampling**.** The gold standard would be the careful selection of a range of environments, followed by the sequencing of large numbers of phylogenetically representative complete genomes directly from those environments. In practice, however, this is not yet attainable on a large scale, although substantial progress is being made in techniques for obtaining genome sequences from single-cells[158 159]. Instead, approaches based on proxy information must be used. Common approaches involve annotating environments based on the original source of cultured organisms**,** surveys of the literature, or database annotations based on one of these approaches**.** Annotating habitat from the source of the isolate is limited both by cultivation bias (the organisms that grow best in culture often represent a non-random subset of environmental diversity[11])**,** and because many organisms, especially those abundant in individual samples, are 'cosmopolitan' and can inhabit a variety of environments[160]. Careful surveys of the literature can be very useful in establishing a broader sense of the set of environments with which a sequenced organism must

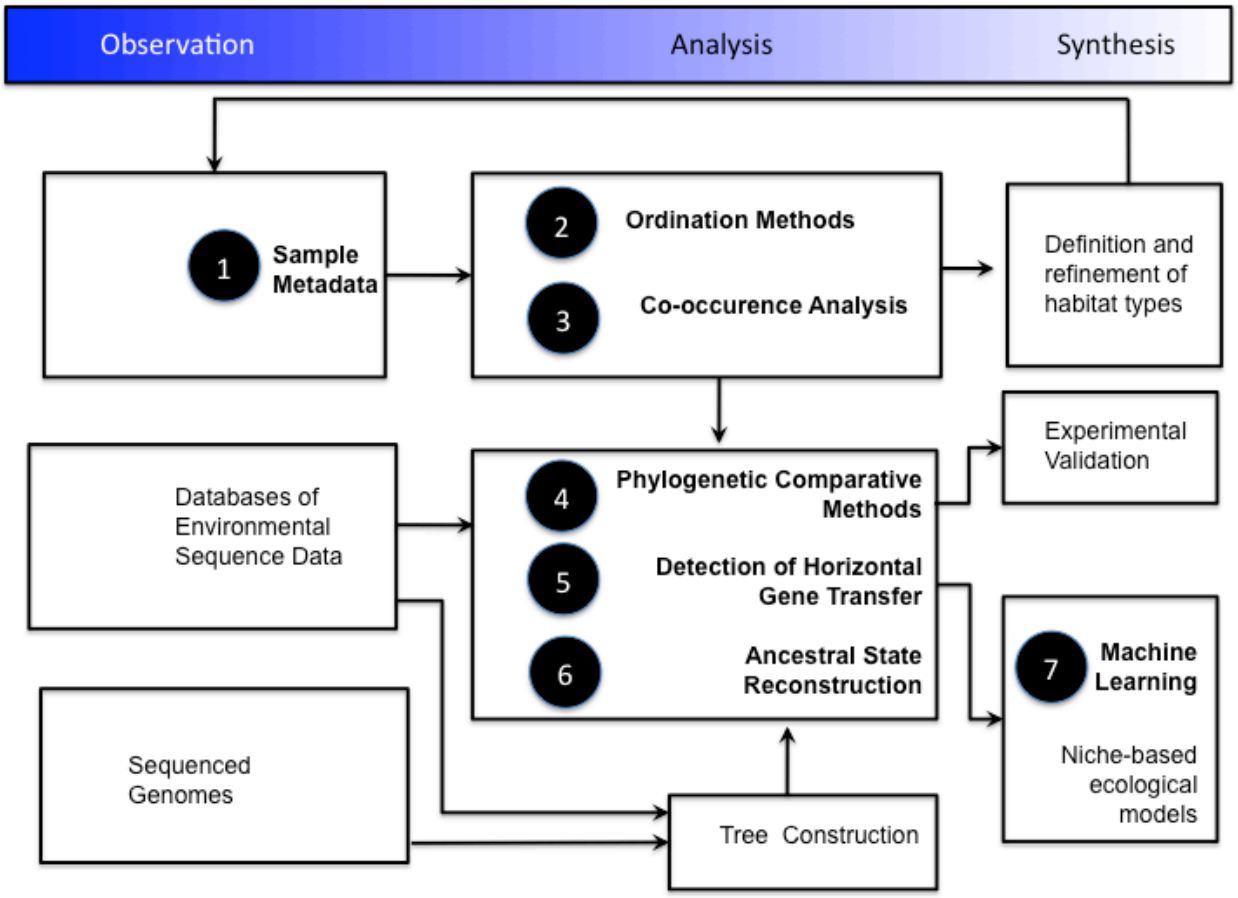


Figure 13. Recurrent themes in the analysis of microbial habitat adaptation. Numbered topics in bold correspond to sections in the text (see main text for additional detail). In order to compare microbes across habitats, it is first necessary to define the environmental factors that structure microbial communities. Insights into this question can be gained by combining sequence data from community surveys (e.g. 16S rRNA or other marker gene sequences) with rich metadata (Topic 1), using ordination techniques (Topic 2). These results can then help to define (and refine) important habitat categories. Interactions between organisms (such as competition or cooperation) can be characterized using co-occurrence analysis (Topic 3). When well-defined and annotated habitat categories (or data on environmental parameters) are available, surveys of microbial communities can be combined with genome sequence data and phylogenetic trees to allow more detailed study of habitat adaptation. Such studies include phylogenetic comparative measures (Topic 4), detection of horizontal gene transfer (Topic 5), and ancestral state reconstruction (Topic 6). Application of these techniques in combination allows for inference of traits involved in habitat adaptation: these traits/habitat associations can then be put into a predictive framework using machine learning techniques (Topic 7) or ecological modeling. Finally, traits predicted to be important for habitat adaptation can be selected for detailed experimental study (for example by mutagenesis followed by competition in microcosms).

contend, but such surveys are laborious and are limited to the lineages actually discussed. Differences in data handling and reporting between studies makes annotation from the literature and from culture collections both challenging and time consuming. An emerging alternative approach is to search community survey data for close relatives of sequenced genomes. Such an approach has the advantage that it can be conducted in a relatively unbiased manner, and can subsequently associate organisms with the environmental samples in which they are found. As databases of 16S rRNA and metagenomic community surveys accumulate, automated methods for surveying the habitat range of microbial taxa (see e.g. [161], [2]) using community surveys will become increasingly effective.

Metadata annotation

The rapid accumulation of studies encompassing thousands of samples and billions of sequences has the potential to allow myriad new insights through comparative analysis. However, in order to maximize this potential, accurate contextual information about the samples (often called "sequence metadata") is an increasingly important consideration (Figure 13,Topic 1). The utility of datasets for comparative analysis is frequently limited by the quality of metadata reported for the sampled environment. Such limitations can be introduced during data collection, data encoding, or data reporting. In the first case, datasets are often limited by reporting of only those physical, chemical or geographic parameters relevant the

particular hypothesis at hand (even if others were collected). A lack of widely adopted standards for encoding the metadata that describes samples also presents significant challenges for comparative analyses. Differences in annotation can range from relatively simple (the use of different names or abbreviations to represent the same body site), to very challenging (differing definitions of environment types). Another limitation occurs during publication: although journals require that sequence data be made publically available, the same requirement has not been enforced for sample metadata.

In order to address these issues, many new sequencing efforts are now adopting the Minimal Information about any (x) Sequence (MIxS) standards, which was proposed by the Genomic Standards Consortium (GSC - http://www.gensc.org/) (Yilmaz P. *et al*., Nat. Biotech., in Press). The MIxS standard encapsulates three metadata compliant data-types, which are the Minimal Information about a (Meta)Genome Sequence (MIMS/MIGS - http://www.gensc.org/gc_wiki/index.php/MIGS/MIMS)[162] and the Minimal Information about a MARKer gene Sequence (MIMARKS - http://www.gensc.org/gc_wiki/index.php/MIMARKS) (Yilmaz P. *et al*., Nat. Biotech., in Press). These standards require researchers to supply their metadata using controlled vocabulary terminology and ontological values, which will greatly benefit those trying to collate and compare across studies. Due to the adoption of such standards, some databases are also starting to require MIxS-compliance during metadata submission. These include the Metagenomics RAST Server (MG-RAST - http://metagenomics.anl.gov/) [133], the Human Microbiome Project (HMP –

http://www.hmpdacc.org/), the Earth Microbiome Project (EMP – http://www.earthmicrobiome.org/) and the QIIME Database (http://www.microbio.me/qiime).

Ordination Methods

When investigating habitat adaptation in microbes, it is crucial to first have a baseline understanding of how microbial communities vary across environmental samples (microbial β diversity), and the main factors that drive such variation (Figure 13, Topic 2**)**. Ordination methods have been widely and fruitfully applied to address these questions. By assessing the microbial composition of each microbial community, researchers can assess the extent to which those communities are partitioned into distinct clusters, or arrayed along a continuous gradient based on environmental factors (see [163] for a survey of ordination methods) .

Ordination analyses performed on microbial community composition data acquired via sequencing of the gene encoding the small subunit ribosomal RNA have been used to distinguish microbial communities, and to identify environmental factors that contribute to both large and small-scale differences between communities. For example, Lozupone *et al.*[60] found a clear split between saline and non-saline environments among non-host associated microbial communities. King *et al.* combined ordination techniques with biogeography to demonstrate the dominant role of pH, plant abundance and snow depth in shaping the microbial communities

found in alpine soil and to build global distribution models for microorganisms in this habitat[164], and Fierer *et al.* used 16S rRNA composition data to show that microbial communities on individuals' hands were far more similar to the communities on their computer keyboards than they were to communities from other individuals' hands[165].

A community-wide perspective on the factors structuring microbial diversity can also be obtained by shotgun metagenomic data. DNA or RNA sequences from random locations on the genomes of many microbes in a community can be assigned to functional (or other) categories, and again ordination methods may be applied to the resulting data. The (dis)agreement between 16S rRNA data and metagenomic data could then be visualized and quantified via Procrustes analysis, which compares the similarity of pairs of ordinations (see [115] for an example of applying this technique to the 5' and 3' paired-end reads of the same rRNA molecules in environmental samples). Such comparisons are one method of determining, at the community level, the degree to which the pool of functional genes in a microbial assemblage is predictable from phylogeny (and thus can detect biologically interesting signatures of competition or functional convergence).

Finally, ordination methods can help to inform high-throughput studies of microbial habitat adaptation by determining which environmental parameters are most important in structuring community diversity (the environmental parameters most important in structuring communities of microbial organisms are not always

intuitively obvious). Objective methods for defining relevant metadata parameters and defining working habitat categories are crucial, because many studies rely heavily on the lifestyle or habitat categories defined in a small number of online databases (primarily NCBI[166] and GOLD[134]) to test comparative genomic hypotheses – thus careful refinement of these categories and addition of more detailed subcategories (based in part on the results of ordination techniques) would yield rapid dividends in comparative analysis.

Application of machine learning techniques

Machine learning techniques hold promise for relating gene functions to habitat distributions (Figure 13, Topic 7). These techniques have been used extensively in taxonomic classification of metagenomic data and many other problems in bioinformatics, but their application to classification and clustering of microbial communities by habitat is relatively new[167]. This emerging approach has been successful for a number of different habitat types. For example, Muegge *et al.* (in press) used a nearest-neighbor approach to demonstrate that phylogenetic characterizations of microbial communities can be used to predict metagenomic profiles of those communities. Werner *et al.* used supervised classifiers to identify a small subset of operational taxonomic units that were highly predictive of the type of bioreactor in brewery wastewater-treatment systems[168]. Supervised classifiers have also recently been applied to source tracking of fecal contamination in water supplies[169].

The primary purpose of supervised machine learning in the context of microbial habitat adaptation is to build predictive models of the differences between habitats. A supervised classifier takes as its input a set of biological samples (training data) characterized by, for example, observations of operational taxonomic units, or counts of gene categories, along with metadata identifying the source habitats of those communities. The output is a model designed to predict the source habitat for novel biological samples not included in the training data, and an estimate of the expected future accuracy of the model. In many cases the classifier will also report a measure of the predictive capability of each of the dependent variables (e.g. gene categories). One of the main advantages of machine learning techniques is that they are designed to discover general trends present in the training data even when the number of dependent variables is much larger than the number of samples, while ignoring idiosyncrasies specific to that training data set (i.e., avoiding overfitting). One exciting direction is that once sufficient genomes linked to environmental samples have been collected, machine learning techniques will be ideal for understanding which genes, regulatory structures, or other properties of the genome are specifically associated with presence in an environment, especially when combined with the phylogenetic methods discussed in the next section.

Phylogenetic comparative methods

Once habitats have been assigned to organisms, relating genome properties to habitats is still challenging. Because all organisms share a common ancestry, each genome sequence cannot be counted as an independent observation when conducting statistical analyses, including machine learning techniques. Instead, the evolutionary history that relates organisms must be taken into account[170] (Figure 13, Topic 4). The importance of this well established, but often ignored, principle is illustrated in Figure 14.

Phylogenetic comparative methods are of particular relevance to microbial ecologists because the organisms selected for genome sequencing are not currently distributed across the tree of life evenly (although efforts are underway to ameliorate this problem[119]). This sequencing bias exacerbates the problems of interpretation introduced when traits are correlated with phylogeny.

Recent investigations of microbial adaptation to the human gut[2], global co-occurrence patterns[155], and genomic changes associated with growth rate[143] have investigated such patterns by plotting relevant traits against phylogenetic distance, and found useful information in both trends that are largely explained by phylogeny (e.g. similarity in GC content[155,143]) and those that also contain signals that cannot be fully explained by phylogeny (e.g. gene content during adaptation to life in the gut[2], gene content, and genome size in co-occurring organisms[155]). Other studies have employed rarefaction, in which data are evened out across categories by discarding members of overrepresented taxa. This technique can provide a useful check on the effects of oversampled taxa, but suffers from the obvious drawback that

it frequently discards a large portion of the data, because n is limited by the least sampled taxon. Nonetheless, the utility

(a) Raw Habitat Data

Node Name	Habitat	Metabolic Genes	Other Genes
A	Habitat 1	2,042	2,957
B	Habitat 1	2,018	2,981
C	Habitat 1	2,021	2,978
D	Habitat 1	2,043	2,956
E	Habitat 1	2,046	2,953
F	Habitat 1	2,016	2,983
G	Habitat 1	2,050	2,949
H	Habitat 1	2,062	2,937
I	Habitat 2	2,102	2,897
J	Habitat 2	2,111	2,888
K	Habitat 2	2,077	2,922
L	Habitat 2	2,080	2,919
M	Habitat 2	2,067	2,932
N	Habitat 2	2,083	2,916
O	Habitat 1	2,032	2,967
P	Habitat 1	2,032	2,967

Summary

Habitat	Metabolic Genes	Other Genes	Average % Metabolic
Habitat 1	20,362	29,628	49.7%
Habitat 2	17,474	12,520	58.3%

Result is significant by G-test for independence (G = 7.886759, p = 0.0024899)

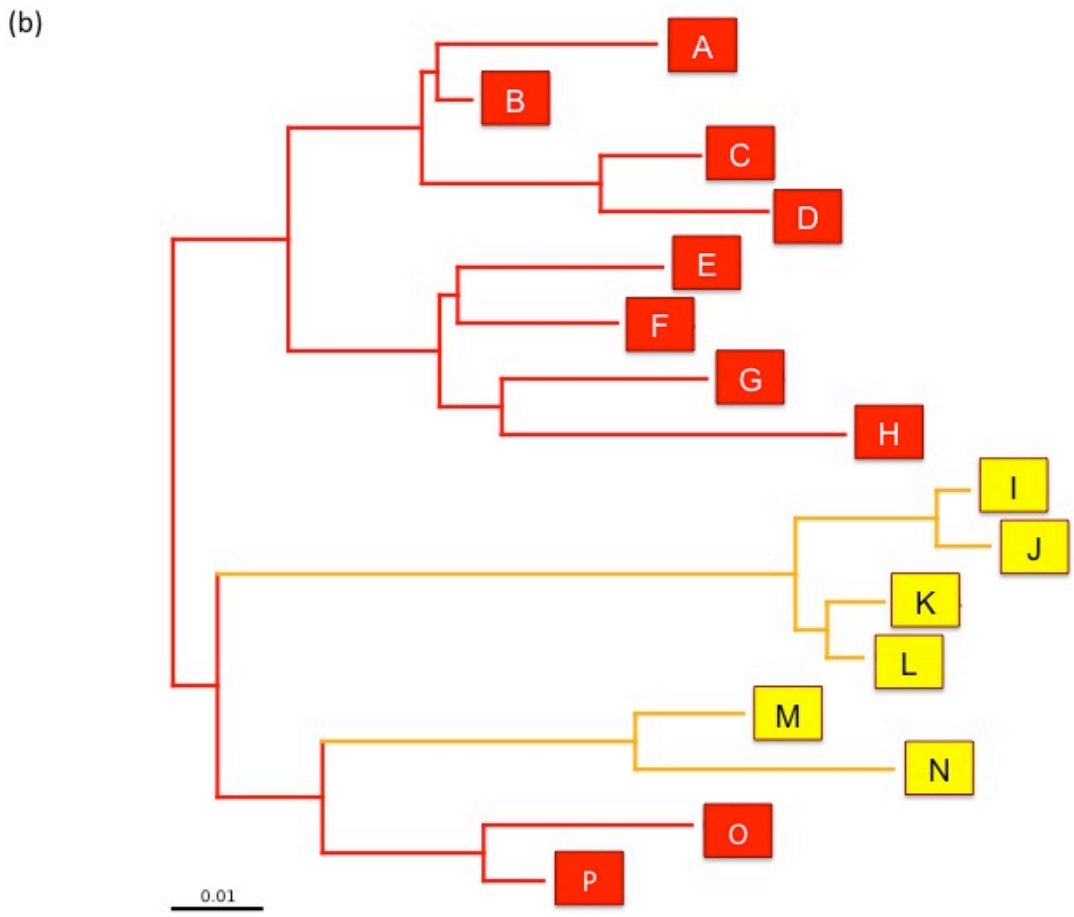


Figure 14. The importance of phylogenetic correction in comparing traits across habitats. Consider the problem faced by an investigator seeking to test whether adaptation to a copiotrophic environment ('Habitat 2') is correlated with acquisition of additional metabolic genes relative to an oligotrophic environment ('Habitat 1'). Given gene presence/absence data derived from whole genome sequences, (a) it may be tempting to use traditional statistical methods without phylogenetic correction to test this hypothesis. For example, naïve assessment of the effect of habitat on gene content using the G-test for independence may lead an investigator to conclude that the increase in representation of metabolic genes between organisms found in Habitat 2 over Habitat 1 (58.53% vs. 49.7%;). In this example, the G-test for independence yields a highly significant p value ($p = 0.00249$) However, examination of a the phylogeny relating the genomes (b) reveals a great deal of phylogenetic structure that is ignored by our statistical test. To illustrate the frequency with which phylogenetically-unaware statistical methods can generate false positive results [127, 128, 171-175] in both qualitative and quantitative data, we simulated the results of applying the G-test to simulated data, similar to that depicted in a) and b) in which *habitat adaptation and metabolic gene evolution are purely independent*. The panel depicts 1000 rounds of simulation on balanced 256 taxon trees, with branch lengths were drawn from an exponential distribution (with mean = 0.05). In each round of simulation, balanced trees were simulated and 5000 binary characters (representing gene presence/absence), plus one habitat character were simulated in a purely neutral fashion (there was no genuine correlation between habitat and gene content), with symmetrical gene gain/loss. Ideally, we would expect no more than 50/1000 (0.05%) false positive rate from a valid statistical test. However, in 38.4% (384/1000) of trials, a G-test of gene content versus habitat falsely reveal a statistically significant result ($p < 0.05$). Thus, application of phylogenetic comparative measures (see **Table 5** for available software) in studies of microbial habitat adaptation is essential.

of relatively unsophisticated methods like regression against phylogenetic distance and rarefaction suggests that inclusion of more formal analyses of phylogenetic signal (for example, phylogenetic independent contrasts[171] and phylogenetic generalized least squares), along with reconstructions of ancestral states could play an important role in future studies of microbial habitat adaptation. The development[171 172 173] and testing [174, 175] of phylogenetic comparative methods for quantitative traits, as well as software packages [127, 128] to make such methods easily accessible, are active areas of research, but many tools exist for estimating these characters without phylogenetic bias (see Table 5) and should be applied in microbial studies.

Ancestral state reconstruction

Reconstruction of ancestral states is a powerful tool to understand molecular and genomic evolution, which is increasingly being applied to the study of microbial habitat adaptation. Ancestral traits for a group of species can be inferred based on a phylogenetic tree, an alignment of the observed states, and a model of evolution of the character under study. By analyzing a character in a group of extant species, the most probable state the character had in the common ancestor of these species can be determined, thus identifying changes that have occurred since divergence. The ancestral sequence can be estimated by one of several methods, such as

parsimony[176], maximum likelihood[177, 178], or Bayesian inference[179-181] (selected tools for performing ancestral state reconstruction are listed in Table 5). For relatively recent evolutionary events, it is sometimes possible to infer probable gene sequences at ancestral nodes. The estimated sequence can then be synthesized, cloned into a vector that is transfected into a cell, and the expressed protein can subsequently be purified in order to study its properties. Based on this process, new insights into the evolution of dim-light vision [182] and steroid receptors [183] have been gained. In addition to gene sequence, ancestral state reconstruction has been applied to infer traits, such as mitochondrial metabolism [184] and the content of genomes[185, 186], or even to infer characteristics of ancestral environments[187]. In the future, it seems likely that integrated studies of genomic evolution including both ancestral state reconstruction of genome contents, sequence-based analyses of selective pressure (e.g. via K_a/K_s ratios)[188], tests of the order of trait divergence [189] and detection of horizontal gene transfer could yield new insights into the evolution of microbial habitat adaptation.

Relating co-occurrence patterns to bacterial genomes

One way to understand potential interactions between organisms that may impact environmental distribution is through the application of co-occurrence analysis (Figure 13, Topic 3). For instance, species that support each other's growth, such as in syntrophic relationships where one organism produces metabolites that

are consumed by the other, would be expected to positively co-occur across samples. In contrast, species that competitively exclude each other (e.g. because of similar metabolic requirements) might negatively co-occur. Co-occurrence patterns, however, are confounded because both positive and negative associations can also be driven by environmental preferences[155, 190].

Combining co-occurrence studies with comparative genomics can clarify the biological properties that drive associations among microbes[155]. As an example, Chaffron *et al.* performed a global analysis of co-occurrence patterns using 16S rRNA surveys representing 3000 distinct sampling events for which sequence data was deposited in GenBank. They then assessed the genomic properties of the subset of OTUs for which close relatives had genome sequences. Although some of the positive associations in the 16S rRNA OTU network reflected known or suspected syntrophic associations, such as a consortium involved in the anaerobic oxidation of methane, the general trends suggested that the major factor driving positive associations was shared environmental preference. Positively co-occurring OTUs were more phylogenetically related than random OTU pairs, extending to lineages that diverged up to 10% at the 16S rRNA level (these would typically be placed in different taxonomic families). Interestingly, positively co-occurring OTUs had more similar genome size, GC content, and relative coverage of KEGG functional pathways than random OTU pairs. The high similarity in GC content could be entirely explained by phylogeny, but the similarities in genome size and KEGG functional pathway coverage were higher than phylogeny could explain. Thus inhabiting the same environment may drive convergence of genome size and

metabolic potential in divergent microbes[155].

Horizontal Gene Transfer

Ongoing studies have continued to document the important roles played by horizontal gene transfer (HGT) in microbial habitat adaptation (**Fig 1 Topic 5**). Although perhaps the greatest emphasis in studies of horizontal gene transfer has been in bacteria and archaea, HGT is also an important mechanism for habitat adaptation in microbial eukaryotes. For example, soil fungi have acquired genes to break down the glucuronides found in vertebrate urine into a usable carbon source, ciliates from the rumen of cows and sheep have acquired ~150 gene families of bacterial origin to break down the down cellulosic plant material, and pathogenic fungi have acquired virulence genes via HGT(reviewed in [191]).

HGT can be detected by several methods which can be generally classified into phylogenetic methods (primarily comparison of gene trees with a 'species tree' or one another), deviations in nucleotide, codon, or amino acid composition, or by finding specific genes or sequences associated with DNA mobility (e.g. transposons, phage or integron integrases, etc.) (see [4] for a review). Although there is ongoing controversy[192] [193] about the total extent of horizontal gene transfer, and the implications of HGT for microbial (especially bacterial and archaeal) phylogeny[194], it is increasingly clear both (i) that HGT has played a major role in bacterial evolution, and (ii) that trees of the universal or nearly-universal genes give the same overall

phylogenetic pattern on average[193], implying that the extent of HGT is not so great that measures of vertical inheritance, such as 16S rRNA phylogenies, are meaningless. Several recent studies of horizontal gene transfer have therefore focused on separating the relative contribution of Horizontal Gene Transfer (by conjugation, phage transduction, transformation, etc.[4]) and vertical descent (including gene loss, duplication, evolution of new gene families, and sequence divergence) to the evolution of gene content.

Schliep and colleagues[157] used information embedded in the set (or 'forest') of gene trees from 100 bacteria and archaea to identify sections of gene trees that were not consistent with vertical descent, but did correspond to lifestyle ('anaerobe') or habitat ('soil') features as derived from NCBI annotations. This analysis yielded sets of gene families that could be better explained by lifestyle or habitat annotations than by taxonomy (~19% of gene families analyzed for hyperthermophiles) as well as networks of gene exchange amongst taxa and clusters of genes that were gained or lost in association with lifestyle.

David and Alm[129] used AnGST, a model that tests for gene duplication, gene loss, and horizontal gene transfer within a single framework, to reconstruct the evolutionary history of 3,983 gene families. The results implied a rapid 'Archaeal Expansion' 3.33-2.65 billion years ago in which the number of gene families expanded by ~26% during a period of rapid diversification. By examining the timing of the expansion, and finding that the functional categories of genes occurring

during this event were primarily associated with redox and electron transfer (O_2 binding, Fe binding, and Fe-S binding were the most enriched categories), David and Alm were able to connect this expansion to the 'great oxygenation event': a dramatic biotically-mediated event in Earth's history, in which the production of oxygen by photosynthesis began to exceed buffering capacity and thus raise O_2 levels in the atmosphere and ocean.

Algorithms that include a unified model of gene evolution hold great promise for the study of habitat adaptation in microbial genomes (see Table 5 for links to the AnGST and Phangorn packages used in these analyses). The separation of genome evolution into specific vertical or horizontal components, and relating patterns in each to changes in habitat or lifestyle are also promising avenues for future research.

Source/sink dynamics

Attempts to map the habitat range of an organism using (metagenomic or marker gene) community surveys is that the presence of a microbe in an assemblage is not proof that the organism is adapted for life there. If a productive (source) and an unproductive (sink) environment are linked by high rates of migration, even relatively abundant organisms in the unproductive environment can be maintained primarily by migration from the source, rather than reproduction in the sink[195]. Such source/sink dynamics have been extensively documented in the ecology of

micro- and macroscopic organisms[114, 195, 196], and are likely to play important roles in many microbial communities**.** For example, microbial assemblages from the human gut may contain transient populations of microorganisms associated with ingested food or the mouth community, in addition to the indigenous community. The complexities presented by source/sink dynamics are compounded by the prevalence of dormancy in microbial populations[197], which can increase the ability of microbes to emigrate to, and persist in, marginal habitats. Currently available techniques for minimizing the effect of source/sink dynamics when annotating habitat range from community surveys include requiring the presence of an OTU across multiple samples, considering the relative presence of an organism in a habitat as a proportion of its total abundance across all environments, and experimental comparison of rRNA and rDNA ratios to test for metabolism in the sample can indicate the presence of alive and actively transcribing organisms as opposed to just their DNA. One additional recent approach to this problem involves new algorithms for tracking recent migration from a source environment (Knights *et al*, submitted). This approach can also detect laboratory contamination, which can lead to inappropriate conclusions about cosmopolitanism (see [160] and references contained therein). However, accurate techniques for inferring microbial habitat adaptation (fitness in a particular habitat, rather than merely presence) from community surveys remain a topic where further development is needed.

Conclusion

The increasing availability of large-scale 16S rRNA and metagenomic community surveys, in combination with whole-genome sequences, provides novel opportunities to conduct large-scale studies relating the survival strategies of microbial organisms to their genomic features. Using the structure of the tree of life will be essential in establishing baseline predictions for trait conservation given phylogeny, and thereby distinguishing novel adaptations to a particular habitat from traits preserved solely due to shared evolutionary history. Given this phylogenetic baseline, large collections of community surveys with backing metadata can be used to detect genomic variations associated with life in a range of environmental conditions. Statistical tools are now available for investigating adaptation along ecological gradients, detecting horizontal gene transfer, reconstructing the evolutionary history of genes involved in environmental adaptation, and inferring positive and negative correlations in species abundance. A major challenge for future studies will be designing and testing accessible, high-throughput pipelines that combine these tools to gain biological insight and generate testable hypotheses from the extremely large-scale sequence collection efforts currently underway.

CHAPTER V.

CONCLUSION

Evolutionary models incorporating phylogeny, genome evolution, and environment will aid in the interpretation of 16S rRNA community surveys

Studies of microbial ecology are increasingly leveraging high-throughput sequencing technologies to gain a wider window into the microbial world. However, the immense complexity and variety of microbial communities implies that such tools (despite the enormous amount of data that they generate) will still provide only very partial information about microbial diversity and ecology. Moreover, even as the cost of sequencing falls, our ability to detect organisms (using 16S rRNA sequences, or other marker genes) will continue to outstrip our ability to collect complete genome sequences. Thus, methods for extrapolating what we know (or don't know) about an organism given the sample in which it was observed and its position on a phylogenetic tree are likely to be of great importance to genomic, metagenomic, and marker gene studies.

Chapter II provides a descriptive examination of the relationship between phylogenetic distance, environmental adaptation (to the human gut), lifestyle (pathogen vs. non-pathogen) and gene content. Moving forward, one key challenge will be to integrate the findings from this analysis, and similar descriptive efforts (e.g. [45, 155]) into predictive models. Crucially, such models will need to take into

account the uncertainty introduced by horizontal gene transfer and the (often very significant) evolutionary distance that separates an uncultured microorganism and its closest cultured or fully sequenced relative. In many cases, we will not be able to make confident prediction about the physiology of an uncultured organism. However, being able to identify these cases should help both to guide interpretation, and suggest organisms that would be fruitful targets for laboratory cultivation and/or genome sequencing.

In the case of bacteria inhabiting the human gut, I observed ([2] and Chapter II) a greater level of shared gene content conservation than expected given phylogeny and genome size. This finding confirmed the earlier hypothesis that ecological differences might explain outliers to the regression of gene content on phylogenetic distance [45]. It has subsequently been independently confirmed [155] that phylogenetically diverse, co-occuring bacteria across a range of environments share more genes than one would expect given phylogeny alone (when compared against random bacterial genomes).

It is not yet clear to what extent gene content conservation in gut micro-organisms that I observed were due to horizontal gene transfer. I observed a dramatic enhancement of the effects of habitat on gene content in the plasmids carried by gut bacteria, suggesting that gene transfer may account for a great deal of the effect of habitat on gene content conservation. However, combined analyses of phylogeny, gene content, and horizontal gene transfer across a broad range of

environments will be needed to test whether this effect is a universal feature of bacterial habitat adaptation.

Compositional, phylogenetic, and mobile-element based methods for the detection of horizontal gene transfer yield complementary insights into microbial genome evolution.

Typically analysis of horizontal gene transfer has employed a single detection technique. However, because all currently available HGT detection algorithms have fairly high false postive and false negative rates , application of multiple, complementary techniques can more detailed insights. Chapter II provided an argument for utilizing a combination of techniques when analyzing horizontal gene transfer, and reviewed many of the sequence features of mobile elements that could be incorporated into such combined analyses. Chapter III describes the application of a combination of compositional and phylogenetic methods (plus the detection of prophage) to study the evolution of *Methanobrevibacter smithii*. This analysis yielded several insights that would not have been available from the application of a single class of techniques alone. These insights included: (i) many differences in the HGT estimates of compositional and phylogenetic methods could be resolved by applying compositional methods to the restricted subset of genes amenable to phylogentic HGT analysis, (ii) many classes of horizontally transferred genes (e.g. ALPs, methanogenesis genes) could be detected by a variety of methods, greatly

strengthening my confidence in the prediction (iii) many prophage genes, despite clear sequence feature-based evidence of transfer, had ameliorated in composition [91] sufficiently during their residence in *M. smithii* genomes that evidence of transfer could no longer be detected. Based on this experience, applying multiple methods of HGT detection seems advisable.

Tools that facilitate the application of multiple HGT detection methods to the same data set (ideally in combination with functional annotation) would greatly simplify this process and broaden its application. CodonExplorer [29, 31] provides one model for an interface that allows data to be gathered and analyzed using multiple compositional HGT detection methods. In that case, the pre-calculating and caching raw compositional data ahead of time allows users to see results more rapidly than if this information had to be recalculated for each analysis. Similar strategies could be applied to gene trees (for phylogenetic HGT methods) and mobile element annotations in sequenced genomes (indeed many databases of mobile elements already exist, as I describe in [4]). Uniting these disparate approaches into an easy to use and extensible tool remains a challenge for future research and software development.

Integrating studies of horizontal gene transfer with analyses of microbial ecology

Many initial studies of horizontal gene transfer focused on estimating the global extent of the phenomenon [72, 99, 100], and its implications for phylogenetic inference [73,

[79]. Increasingly, however, it has become possible to integrate such studies with analyses of microbial habitat adaptation or community dynamics (see Chapter IV for more extensive discussion).

Although there remain methodological challenges to overcome, single-cell genomics (see [198-202]) appears to be one extremely promising avenue for such integrated research As discussed in Chapter I, although existing shotgun metagenomic approaches allow for comparison of the genomic features of entire microbial communities, the mixture of all organisms in the community into a single pool of sequences greatly complicates studies of ecological interactions or gene transfer within communities. Single-cell genomics, if proven to be sufficiently cheap and accurate, has the potential to resolve many of these issues by providing snapshots of many phylogenetically representative genomes from a community. By preserving the association between genes and organisms, inference of horizontal gene transfer and other aspects of the evolution of bacterial genomes (e.g. gene duplications, gene deletions, sequence inversions, phage integration, etc.) is still possible. Such techniques may allow, for example, the direct observation of the spread of antibiotic resistance genes from lineage to lineage in the intestinal tracts of individuals treated with antibiotics if applied to fecal samples collected in time series during treatment. Regardless of the success of this particular technique, it seems likely that approaches integrating analysis of bacterial genome evolution, community dynamics, and physiology will yield new insights into the complex and fascinating microbial world.

REFERENCES

1. Zaneveld, J., *et al*. (2008) Host-bacterial coevolution and the search for new drug targets. *Curr Opin Chem Biol* 12, 109-114
2. Zaneveld, J.R., *et al*. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res* 38, 3869-3879
3. Hansen, E.E., *et al*. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4599-4606
4. Zaneveld, J.R., *et al*. (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* 154, 1-15
5. Gill, S.R., *et al*. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355-1359
6. Moran, N.A. (2006) Symbiosis. *Curr Biol* 16, R866-871
7. Charleston, M.A., and Perkins, S.L. (2006) Traversing the tangle: algorithms and applications for cophylogenetic studies. *J Biomed Inform* 39, 62-71
8. Stevens, J. (2004) Computational aspects of host-parasite phylogenies. *Brief Bioinform* 5, 339-349
9. Osawa, R., *et al*. (1993) Microbiologial Studies of the Intestinal Microflora of the Koala, *Phascolarctos cinereus*. 2. Pap, a Special Maternal Feces Consumed by Juvenile Koalas. *Australian Journal of Zoology* 41, 611-620
10. Dethlefsen, L., *et al*. (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449, 811-818
11. Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71, 8228-8235
12. Ochman, H., and Moran, N.A. (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096-1099
13. Wu, D., *et al*. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol* 4, e188
14. Hosokawa, T., *et al*. (2007) Obligate symbiont involved in pest status of host insect. *Proc Biol Sci* 274, 1979-1984
15. Xu, J., *et al*. (2007) Evolution of Symbiotic Bacteria in the Distal Human Intestine. *PLoS Biol* 5, e156
16. Sonnenburg, J.L., *et al*. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* 307, 1955-1959
17. Samuel, B.S., *et al*. (2007) Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. *Proc Natl Acad Sci U S A* 104, 10643-10648
18. Turnbaugh, P.J., *et al*. (2007) The human microbiome project. *Nature* 449, 804-810
19. Kurokawa, K., *et al*. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 14, 169-181
20. Hansen-Wester, I., *et al*. (2002) Analyses of the evolutionary distribution of Salmonella translocated effectors. *Infect Immun* 70, 1619-1622
21. Hensel, M. (2004) Evolution of pathogenicity islands of Salmonella enterica. *Int J Med Microbiol* 294, 95-102
22. Lujan, S.A., *et al*. (2007) Disrupting antibiotic resistance propagation by inhibiting the conjugative DNA relaxase. *Proc Natl Acad Sci U S A* 104, 12282-12287

23. Dahlgren, M.K., *et al*. (2007) Design, Synthesis, and Multivariate Quantitative Structure-Activity Relationship of Salicylanilides-Potent Inhibitors of Type III Secretion in Yersinia. *J Med Chem* 50, 6177-6188
24. Hsiao, W.W., *et al*. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet* 1, e62
25. Klumpp, J., and Fuchs, T.M. (2007) Identification of novel genes in genomic islands that contribute to Salmonella typhimurium replication in macrophages. *Microbiology* 153, 1207-1220
26. Liu, Z., *et al*. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35, e120
27. Turnbaugh, P.J., *et al*. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031
28. Tringe, S.G., *et al*. (2005) Comparative metagenomics of microbial communities. *Science* 308, 554-557
29. Zaneveld, J., *et al*. (2009) CodonExplorer: an interactive online database for the analysis of codon usage and sequence composition. *Methods Mol Biol* 537, 207-232
30. Kuczynski, J., *et al*. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* 11, 210
31. Hamady, M., *et al*. (2009) CodonExplorer: an online tool for analyzing codon usage and sequence composition, scaling from genes to genomes. *Bioinformatics* 25, 1331-1332
32. Caporaso, J.G., *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7, 335-336
33. Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* 276, 734-740
34. Iwabe, N., *et al*. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A* 86, 9355-9359
35. Woese, C.R. (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* 97, 8392-8396
36. Woese, C.R. (1987) Bacterial evolution. *Microbiol Rev* 51, 221-271
37. Olsen, G.J., and Woese, C.R. (1993) Ribosomal RNA: a key to phylogeny. *FASEB J* 7, 113-123
38. Doolittle, W.F., and Brown, J.R. (1994) Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci U S A* 91, 6721-6728
39. Ley, R.E., *et al*. (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 6, 776-788
40. Ley, R.E., *et al*. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022-1023
41. Turnbaugh, P.J., *et al*. (2008) A core gut microbiome in obese and lean twins. *Nature*
42. Frank, D.N., *et al*. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 104, 13780-13785
43. Dethlefsen, L., *et al*. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 6, e280
44. Li, M., *et al*. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A* 105, 2117-2122
45. Konstantinidis, K.T., and Tiedje, J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 10, 504-509
46. Konstantinidis, K.T., and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102, 2567-2572
47. Welch, R.A., *et al*. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A* 99, 17020-17024
48. Gressmann, H., *et al*. (2005) Gain and loss of multiple genes during the evolution of Helicobacter pylori. *PLoS Genet* 1, e43
49. Sreevatsan, S., *et al*. (1997) Restricted structural gene polymorphism in the Mycobacterium

tuberculosis complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A* 94, 9869-9874
50. Achtman, M., *et al*. (2004) Microevolution and history of the plague bacillus, Yersinia pestis. *Proc Natl Acad Sci U S A* 101, 17837-17842
51. Moran, N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108, 583-586
52. Andersson, S.G., and Kurland, C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol* 6, 263-268
53. Sallstrom, B., and Andersson, S.G. (2005) Genome reduction in the alpha-Proteobacteria. *Curr Opin Microbiol* 8, 579-585
54. Fukuchi, S., *et al*. (2003) Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol* 327, 347-357
55. Paul, S., *et al*. (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* 9, R70
56. Hutchinson, G.E. (1959) Homage to Santa Rosalia, or Why are there so many kinds of animals? *Am. Nat*. 93, 145-149
57. Sokurenko, E.V., *et al*. (1998) Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. *Proc Natl Acad Sci U S A* 95, 8922-8926
58. Sokurenko, E.V., *et al*. (2004) Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in Escherichia coli. *Mol Biol Evol* 21, 1373-1383
59. Liolios, K., *et al*. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 36, D475-479
60. Knight, R., *et al*. (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol* 8, R171
61. Altschul, S.F., *et al*. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410
62. Coenye, T., and Vandamme, P. (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett* 228, 45-49
63. DeSantis, T.Z., Jr., *et al*. (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* 34, W394-399
64. Lane, D.J. (1991) 23S/16S rRNA Sequencing. In *Nucleic Acid Techniques in Bacterial Systematics* (Stackebrandt, E., and Goodfellow, M., eds), Wiley
65. DeSantis, T.Z., *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72, 5069-5072
66. Sheneman, L., *et al*. (2006) Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* 22, 2823-2824
67. Liu, Z., *et al*. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 36, e120
68. Backhed, F., *et al*. (2005) Host-bacterial mutualism in the human intestine. *Science* 307, 1915-1920
69. Tamames, J. (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol* 2, RESEARCH0020
70. Thomas, C.M., and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3, 711-721
71. Jaspers, E., and Overmann, J. (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol* 70, 4831-4839
72. Nakamura, Y., *et al*. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 36, 760-766
73. Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124-2129
74. Green, J.L., *et al*. (2008) Microbial biogeography: from taxonomy to traits. *Science* 320, 1039-1043

75. Turnbaugh, P.J., *et al*. (2009) A core gut microbiome in obese and lean twins. *Nature* 457, 480-484
76. Costello, E.K., *et al*. (2009) Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*
77. Field, D., *et al*. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26, 541-547
78. Samuel, B.S., and Gordon, J.I. (2006) A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proc Natl Acad Sci U S A* 103, 10011-10016
79. Syvanen, M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* 28, 237-261
80. Page, R.D., and Charleston, M.A. (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol* 7, 231-240
81. Nelson, K.E., *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399, 323-329
82. Lander, E.S., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860-921
83. Jordan, I.K., *et al*. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* 11, 555-565
84. Mirkin, B.G., *et al*. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3, 2
85. Farahi, K., *et al*. (2003) RED-T: utilizing the Ratios of Evolutionary Distances for determination of alternative phylogenetic events. *Bioinformatics* 19, 2152-2154
86. Kechris, K.J., *et al*. (2006) Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. *Proc Natl Acad Sci U S A* 103, 9584-9589
87. Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* 201, 187-191
88. Kinsella, R.J., and McInerney, J.O. (2003) Eukaryotic genes in Mycobacterium tuberculosis? Possible alternative explanations. *Trends Genet* 19, 687-689
89. Ragan, M.A., *et al*. (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol* 14, 4-8
90. Sharp, P.M., and Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15, 1281-1295
91. Lawrence, J.G., and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44, 383-397
92. Karlin, S., *et al*. (1998) Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol* 29, 1341-1355
93. Lawrence, J.G., and Ochman, H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A* 95, 9413-9417
94. Hooper, S.D., and Berg, O.G. (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol* 54, 365-375
95. Vernikos, G.S., and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* 22, 2196-2203
96. Groisman, E.A., *et al*. (1993) Molecular, functional, and evolutionary analysis of sequences specific to Salmonella. *Proc Natl Acad Sci U S A* 90, 1033-1037
97. Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205, 309-316
98. Faith, J.J., and Pollock, D.D. (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165, 735-745
99. Ge, F., *et al*. (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol* 3, e316

100. Choi, I.G., and Kim, S.H. (2007) Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A* 104, 4489-4494
101. Tsirigos, A., and Rigoutsos, I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* 33, 922-933
102. Than, C., *et al*. (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9, 322
103. Price, M.N., *et al*. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5, e9490
104. Sauer, F.D. (1986) Tetrahydromethanopterin methyltransferase, a component of the methane synthesizing complex of Methanobacterium thermoautotrophicum. *Biochem Biophys Res Commun* 136, 542-547
105. Ciccarelli, F.D., *et al*. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283-1287
106. Johnson, Z.I., *et al*. (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* 311, 1737-1740
107. Lugtenberg, B., and Kamilova, F. (2009) Plant-growth-promoting rhizobacteria. *Annu Rev Microbiol* 63, 541-556
108. Womack, A.M., *et al*. Biodiversity and biogeography of the atmosphere. *Philos Trans R Soc Lond B Biol Sci* 365, 3645-3653
109. Cheesman, S.E., *et al*. Epithelial cell proliferation in the developing zebrafish intestine is regulated by the Wnt pathway and microbial signaling via Myd88. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4570-4577
110. Samuel, B.S., *et al*. (2008) Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. *Proc Natl Acad Sci U S A* 105, 16767-16772
111. Sharon, G., *et al*. Commensal bacteria play a role in mating preference of Drosophila melanogaster. *Proc Natl Acad Sci U S A* 107, 20051-20056
112. Andersen, S.B., *et al*. (2009) The life of a dead ant: the expression of an adaptive extended phenotype. *Am Nat* 174, 424-433
113. Vyas, A., *et al*. (2007) Behavioral changes induced by Toxoplasma infection of rodents are highly specific to aversion of cat odors. *Proc Natl Acad Sci U S A* 104, 6442-6447
114. Sokurenko, E.V., *et al*. (2006) Source-sink dynamics of virulence evolution. *Nat Rev Microbiol* 4, 548-555
115. Caporaso, J.G., *et al*. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4516-4522
116. Qin, J., *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65
117. Peterson, J., *et al*. (2009) The NIH Human Microbiome Project. *Genome Res* 19, 2317-2323
118. Stewart, F.J., *et al*. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol* 12, R26
119. Wu, D., *et al*. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056-1060
120. Schloss, P.D., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* 75, 7537-7541
121. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24, 1586-1591
122. Carmel, L., *et al*. EREM: Parameter Estimation and Ancestral Reconstruction by Expectation-Maximization Algorithm for a Probabilistic Model of Genomic Binary Characters Evolution. *Adv Bioinformatics*, 167408
123. Paradis, E., *et al*. (2004) APE: Analyses of Phylogenetics and Evolution in R language.

Bioinformatics 20, 289-290
124. Ronquist, F., and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572-1574
125. Drummond, A.J., and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7, 214
126. Dufour, S.D.a.A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 22, 1-20
127. Jombart, T.*, et al*. adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* 26, 1907-1909
128. Kembel, S.W.*, et al*. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463-1464
129. David, L.A., and Alm, E.J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469, 93-96
130. Podell, S., and Gaasterland, T. (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* 8, R16
131. Podell, S.*, et al*. (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics* 9, 419
132. Schliep, K.P. phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592-593
133. Meyer, F.*, et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386
134. Liolios, K.*, et al*. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38, D346-354
135. von Mering, C.*, et al*. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315, 1126-1130
136. Moran, N.A.*, et al*. (2008) Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* 42, 165-190
137. Konstantinidis, K.T., and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* 101, 3160-3165
138. Reistad, R. (1970) On the composition and nature of the bulk protein of extremely halophilic bacteria. *Arch Mikrobiol* 71, 353-360
139. Lanyi, J.K. (1974) Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriol Rev* 38, 272-290
140. Oren, A. (1999) Bioenergetic aspects of halophilism. *Microbiol Mol Biol Rev* 63, 334-348
141. Kunin, V.*, et al*. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* 4, 198
142. Rhodes, M.E.*, et al*. Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environ Microbiol* 12, 2613-2623
143. Vieira-Silva, S., and Rocha, E.P. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* 6, e1000808
144. Klappenbach, J.A.*, et al*. (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* 66, 1328-1333
145. Klappenbach, J.A.*, et al*. (2001) rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res* 29, 181-184
146. Stevenson, B.S., and Schmidt, T.M. (2004) Life history implications of rRNA gene copy number in Escherichia coli. *Appl Environ Microbiol* 70, 6670-6677
147. Moran, N.A.*, et al*. (2008) Genomics and Evolution of Heritable Bacterial Symbionts. *Annual Review of Genetics* 42, 165-190
148. Pallen, M.J., and Wren, B.W. (2007) Bacterial pathogenomics. *Nature* 449, 835-842
149. Gray, M.W. (1999) Evolution of organellar genomes. *Current Opinion in Genetics & Development* 9, 678-687.
150. Archibald, J.M., and Lane, C.E. (2009) Going, going, not quite gone: nucleomorphs as a case

study in nuclear genome reduction. *J Hered* 100, 582-590
151. Keeling, P.J., and Slamovits, C.H. (2005) Causes and effects of nuclear genome reduction. *Curr Opin Genet Dev* 15, 601-608
152. Keeling, P.J., *et al*. The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. *Genome Biol Evol* 2, 304-309
153. Giovannoni, S.J., *et al*. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242-1245
154. Partensky, F., and Garczarek, L. (2010) Prochlorococcus: Advantages and Limits of Minimalism. *Annual Review of Marine Science* 2, 305-331
155. Chaffron, S., *et al*. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* 20, 947-959
156. Merhej, V., *et al*. (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* 4, 13
157. Schliep, K., *et al*. Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol Biol Evol* 28, 1393-1405
158. Ishoey, T., *et al*. (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol* 11, 198-204
159. Coleman, M.L., and Chisholm, S.W. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci U S A* 107, 18634-18639
160. Nemergut, D.R., *et al*. Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* 13, 135-144
161. Lozupone, C.A., *et al*. (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci U S A* 105, 15076-15081
162. Kottmann, R., *et al*. (2008) A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 12, 115-121
163. Ramette, A. (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* 62, 142-160
164. Rousk, J., *et al*. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J* 4, 1340-1351
165. Fierer, N., *et al*. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A* 107, 6477-6481
166. Benson, D.A., *et al*. GenBank. *Nucleic Acids Res* 39, D32-37
167. Knights, D., *et al*. Supervised classification of human microbiota. *FEMS Microbiol Rev* 35, 343-359
168. Werner, J.J., *et al*. Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc Natl Acad Sci U S A* 108, 4158-4163
169. Smith, A., *et al*. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Res* 44, 4067-4076
170. Harvey, P.H.a.P., Mark D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press
171. Felsenstein, J. (1985) Phylogenies and the Comparative Method. *The American Naturalist* 125, 1-15
172. Blomberg, S.P., *et al*. (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717-745
173. Jombart, T., *et al*. Putting phylogeny into the analysis of biological traits: a methodological approach. *J Theor Biol* 264, 693-701
174. Laurin, M. Assessment of the relative merits of a few methods to detect evolutionary trends. *Syst Biol* 59, 689-704
175. Freckleton, R.P., *et al*. (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* 160, 712-726
176. Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 83-92

177. Yang, Z., *et al*. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641-1650
178. Koshi, J.M., and Goldstein, R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol* 42, 313-320
179. Yang, Z., and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* 14, 717-724
180. Huelsenbeck, J.P., and Bollback, J.P. (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol* 50, 351-366
181. Pagel, M., *et al*. (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst Biol* 53, 673-684
182. Yokoyama, S., *et al*. (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci U S A* 105, 13480-13485
183. Thornton, J.W., *et al*. (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301, 1714-1717
184. Gabaldon, T., and Huynen, M.A. (2003) Reconstruction of the proto-mitochondrial metabolism. *Science* 301, 609
185. Ma, J., *et al*. (2006) Reconstructing contiguous regions of an ancestral genome. *Genome research* 16, 1557-1565
186. Paten, B., *et al*. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research* 18, 1829-1843
187. Gaucher, E.A., *et al*. (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425, 285-288
188. Marri, P.R., *et al*. (2006) Gene gain and gene loss in streptococcus: is it driven by habitat? *Mol Biol Evol* 23, 2379-2391
189. Ackerly, D.D., *et al*. (2006) Niche evolution and adaptive radiation: testing the order of trait divergence. *Ecology* 87, S50-61
190. Horner-Devine, M.C., *et al*. (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* 88, 1345-1353
191. Andersson, J.O. (2009) Gene transfer and diversification of microbial eukaryotes. *Annu Rev Microbiol* 63, 177-193
192. Galtier, N., and Daubin, V. (2008) Dealing with incongruence in phylogenomic analyses. *Philos Trans R Soc Lond B Biol Sci* 363, 4023-4029
193. Koonin, E.V., *et al*. Comparison of Phylogenetic Trees and Search for a Central Trend in the "Forest of Life". *J Comput Biol*
194. Andam, C.P., *et al*. Biased gene transfer mimics patterns created through shared ancestry. *Proc Natl Acad Sci U S A* 107, 10679-10684
195. Kawecki, T.J. (2000) Adaptation to marginal habitats: contrasting influence of the dispersal rate on the fate of alleles with small and large effects. *Proc Biol Sci* 267, 1315-1320
196. Venail, P.A., *et al*. Dispersal scales up the biodiversity-productivity relationship in an experimental source-sink metacommunity. *Proc Biol Sci* 277, 2339-2345
197. Jones, S.E., and Lennon, J.T. Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci U S A* 107, 5881-5886
198. Ochman, H. (2007) Single-cell genomics. *Environ Microbiol* 9, 7
199. Siegl, A., *et al*. Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *ISME J* 5, 61-70
200. Walker, A., and Parkhill, J. (2008) Single-cell genomics. *Nat Rev Microbiol* 6, 176-177
201. Yoon, H.S., *et al*. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 332, 714-717
202. Zhao, R. (2005) From single cell gene-based diagnostics to diagnostic genomics: current applications and future perspectives. *Clin Lab Sci* 18, 254-262