LIFE HISTORIES OF BACTERIA: GENOMIC FOUNDATIONS AND ECOLOGICAL IMPLICATIONS

By

Benjamin R. Roller

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics – Doctor of Philosophy
Ecology, Evolutionary Biology and Behavior – Dual Major

2015

ABSTRACT

LIFE HISTORIES OF BACTERIA: GENOMIC FOUNDATIONS AND ECOLOGICAL IMPLICATIONS

By

Benjamin R. Roller

Life history tradeoffs are of great interest to biologists because they are central to biodiversity theory and have the power to explain the outcomes of competition. One particular tradeoff has been a frequent target of study, the relationship between rapid and efficient reproduction. Researchers have pursued evidence for the existence of this tradeoff in many study systems and the literature surrounding this topic is extensive. While theoretical studies have indicated that a rate-efficiency tradeoff should play an important role in the evolution of bacterial populations, experiments have frequently found conflicting evidence for its existence and influence on bacterial evolution. In this dissertation I explore the physiological and ecological conditions favoring rapid and efficient bacterial growth.

Microbial ecologists have long observed that the richness of cultivation medium leads to the growth of different types of bacteria. Copiotrophic bacteria have a higher relative fitness under conditions of resource abundance, while oligotrophic bacteria are favored when resources are scarce. The central topic of my dissertation research is to explore if copiotrophic bacteria employ rapid growth life history tactics and if oligotrophic bacteria employ efficient growth life history tactics. It has been noted that rapidly growing bacteria tend to possess multiple copies of the ribosomal RNA operon (*rrn*) in their genomes, while oligotrophic bacteria typically encode few *rrn* copies. I examined if *rrn* copy number was related to rapid and efficient growth tactics using physiological experiments and comparative genomics.

The major findings from my research suggest that copiotrophs tend to utilize rapid growth tactics, while oligotrophs utilize efficient growth tactics. This evidence is consistent with a rate-efficiency tradeoff underlying divergence on this life history axis. I also demonstrate that *rrn* copy number is quantitative predictor of life history tactics and that it explains a large fraction of variation in the genome content of diverse bacterial species. Finally, I have explored particular features of bacterial genomes which play a role in life history adaptation.

It is increasingly recognized that bacteria directly influence the health of our planet. However, the scale of bacterial diversity is immense and there is much more to learn before we can manage bacterial communities to improve wellbeing on Earth. Applying life history theory to bacteria holds promise for improving our understanding of the ecological and evolutionary forces acting on bacterial populations and communities.

ACKNOWLEDGMENTS

My graduate research would not have been possible without the support of many people and institutions. I am grateful to Tom Schmidt who has been a terrific mentor and PhD adviser. Not only has Tom provided intellectual and scientific guidance, but his approach to research and academic life has also been a great example to learn from. Tom's working group has gone through many changes during my time, but the inquisitive spirit and enlightened atmosphere have been a constant source of scientific inspiration. The Schmidt lab has been a fun group of colleagues and friends to spend time with over the last seven years.

I would also like to acknowledge the financial and institutional support that has helped me perform my research. The MSU graduate school, MMG department, EEBB program, Kellogg Biological Station, and U.S. Department of Energy Science Graduate Fellowship have all provided funding and the opportunity to grow as a scientist. My graduate committee has provided the guidance I needed, despite their dispersal away from East Lansing. I appreciate the extra effort Drs. Rob Britton, Jay Lennon, and Rich Lenski have contributed to help me during my graduate career.

Most importantly, I would like to acknowledge my family. My parents, brothers, and in-laws have provided great encouragement over the last several years and a welcome respite from science over holidays and vacations. My wife, Alexa, has been the most supportive and encouraging partner anyone could hope for. She has helped me to become a better person and a better scientist. I have the privilege of working in a job I love, but the most exciting part of my work day is when it is over and I get to come home to her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Life history theory provides an integrated perspective on bacterial reproduction

Introduction

Among all domains of life, Bacteria are conspicuous for their rapid reproduction. They possess maximum reproduction rates much faster than Archaea and Eukaryota (Neidhardt, 1999; Kempes *et al.*, 2012). For example, a one μm^3 bacterium with infinite resources and a doubling time of 20 minutes could generate a volume of offspring larger than the Earth in 48 hours (Russell & Cook, 1995). However, rapid reproduction rates are not universal among bacteria. A wide variety of maximum recorded population growth rates have been observed among diverse bacterial species (Vieira-Silva & Rocha, 2010), indicating substantial variation in their reproductive habits. A bacterium's fitness, *i.e.* lifetime reproductive success, is partly determined by its maximal population growth rate, but rapid reproduction is transient and infrequent for bacteria in natural environments (Hoehler & Jørgensen, 2013; Schmidt & Konopka, 2009). Exponential reproduction can only be supported for short periods of time so fitness is also influenced by other traits in combination with the environment (Vasi *et al.*, 1994). A bacterium's overall pattern of survival and reproduction across multiple environments is its life history. My thesis explores patterns in reproductive variation across the bacterial tree of life through the lens of life history evolution.

Environmental microbiologists have frequently classified bacteria into groups based on their reproductive phenotypes (Kuznetsov *et al.*, 1979). One commonly used scheme relates bacterial reproduction to a spectrum of nutritional preference from copiotrophy to oligotrophy.

Copiotrophic bacteria are most competitive when resources are abundant, while oligotrophic bacteria are favored when resource availability is chronically low (Koch, 2001). In Chapters 1 and 2 of my thesis, I argue that copiotrophy and oligotrophy are intrinsically linked to the life history tactics of rapid or efficient reproduction, respectively. Bacteria achieve these tactics using sets of coadapted traits shaped by natural selection to cope with the contrasting environmental pressures imposed by resource availability. The number of rRNA operon (*rrn*) copies encoded in a bacterial genome is a proxy of the organism's life history and Chapter 3 details a redesign of the *rrn*DB, a database of *rrn* copy numbers, which was updated and linked to annotated genome features of the Kyoto Encyclopedia of Genes and Genomes (KEGG). I provide evidence in Chapter 4 that a number of life history traits are consistently correlated with the number of *rrn* copies encoded in a bacterium's genome. I use this information in Chapters 4 and 5 to demonstrate that life history evolution influences genome content across the bacterial tree of life and examine the genome features driving this pattern.

This introductory chapter is composed of three parts. First, I examine how ideas from classical life history evolution apply to bacteria and demonstrate the conceptual link between life history, Hutchinson's definition of an ecological niche, and my central hypothesis. I then summarize the evidence demonstrating an intrinsic link between rapid bacterial reproduction, protein synthesis, and *rrn* copy number. Finally, I explore the physiological costs of rapid bacterial growth and the evolutionary implications of this expensive life history tactic.

Life history evolution, Hutchinson's niche, and the Bacteria

Classical life history theory views variation in reproduction and survival phenotypes as the outcome of natural selection acting to optimize the investment of limited resources. One

central claim of this theory is that only two types of information are needed to explain life history variation: how fitness determining traits interact with the environment and how they interact with each other (Stearns, 2000). However, applying classical life history theory to bacteria is not straightforward. Eukaryotic organisms have been the focus of classical life history theory but the most important life history traits of these organisms, such as age/size at sexual maturity and reproductive lifespan, are not applicable to bacteria which typically reproduce asexually and are rarely reproductively immature. In order to apply life history theory to bacteria, we first must understand the factors driving bacterial fitness. Two important determinants of bacterial fitness are reproductive rate – the number of progeny produced per unit time, and reproductive efficiency – the number of progeny produced per unit of resource consumed (Maclean, 2008; Bachmann *et al.*, 2013). Rapid and efficient reproduction phenotypes are influenced by numerous underlying metabolic and molecular traits (Flamholz *et al.*, 2013; Bachmann *et al.*, 2013; Vasi *et al.*, 1994), and are also selected for under the contrasting conditions of nutritional richness and scarcity (Pfeiffer *et al.*, 2001). Therefore, rapid and efficient reproduction are life history tactics (Stearns, 1976) – sets of coadapted traits shaped by natural selection to cope with particular environmental pressures. These life history tactics are differentially beneficial to copiotrophs and oligotrophs.

The central hypothesis of this thesis is that *rrn* copy number is a quantitative proxy of a bacterium's place on the life history spectrum from oligotrophy to copiotrophy. If true, metrics of rapid reproduction should be positively correlated with *rrn* copy number, while metrics of efficient reproduction should be negatively correlated with *rrn* copy number. Additionally, traits which influence nutritional habits and that are linked to either the rate or efficiency of a bacterium's reproduction should be correlated with *rrn* copy number among bacteria.

Throughout this thesis I will present evidence from comparative studies of many bacteria that supports these hypotheses, ultimately describing how traits that influence bacterial fitness interact with each other and important environmental variables.

A visual summary of my thesis findings on bacterial life history is presented in Figure 1.1. This figure illustrates how fitness varies for three hypothetical bacterial species within their Hutchinsonian niche – the multidimensional environmental conditions which allow for persistence. The extent of a species' niche is depicted by the boundaries of each irregular shape, while each species' life history is the pattern of fitness, indicated by color density, within its niche. This visualization of life history, which is certainly a simplification of the true multidimensional niche, shows the three axes of environmental variation. Resource availability is depicted as a key axis of life history variation because both the rate and efficiency of bacterial growth are correlated with this variable within and between species (Roller & Schmidt, 2015). While all three hypothetical species in Figure 1.1 have a pattern of fitness that increases with higher resource availability, each species possesses a maximum fitness at a different place along the resource availability spectrum. I demonstrate in Chapter 4 that *rrn* copy number is a quantitative proxy of an organism's placement along this axis. This is illustrated in Figure 1.1 by the parallel relationship between the *rrn* copy number plane and the nutrient availability dimension of niche space. This idea is based on a collection of physiological, genetic, and ecological evidence suggesting that high protein synthesis capacity and *rrn* copy number are inherent features of rapid reproduction.

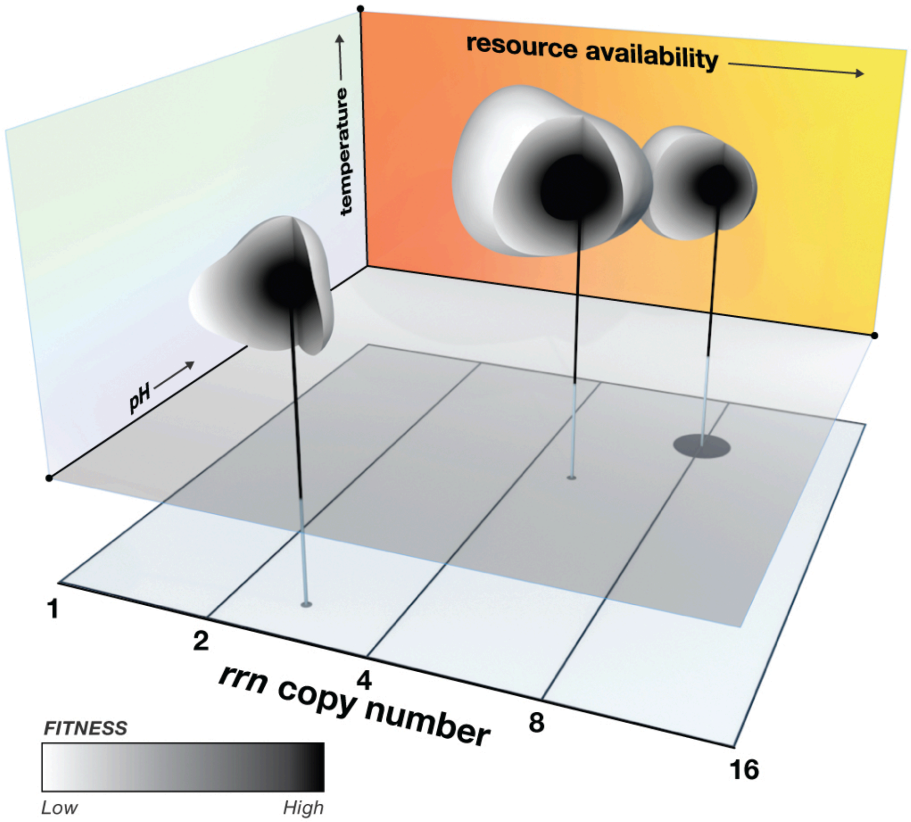


Figure 1.1: Life history mapped onto niche space. Conceptual model of three bacterial species' life histories mapped onto three axes of their multidimensional niches. *rrn* copy number predicts a major axis of fitness variation among species.

Physiology of bacterial reproduction

Reproduction requires a coordination of many biosynthetic processes. Bacteria alter their biomass composition and size depending on their reproduction rate (Schaechter *et al.*, 1958). During unrestricted reproduction a bacterium sees an essentially unchanging environment with all nutrients available in excess of biosynthetic demand. The macromolecular composition of a bacterium in unrestricted reproduction is constant and all major biomass constituents are produced at the same rate. The physiological condition where all macromolecule biosynthesis occurs at the same rate is defined as balanced growth, and it can also be achieved when continuous cultivation in a chemostat reaches a steady-state. Balanced growth is an essential

experimental technique for making repeatable measurements of physiological phenomena in conditions of nutrient limitation and nutrient excess (Neidhardt *et al.*, 1990). For the remainder of this thesis growth will be used synonymously with reproduction, unless otherwise specified, as they are equivalent under balanced growth conditions.

Despite the fleeting occurrence of maximal reproduction rates in nature, cultivating bacteria under nutrient replete conditions has led to a greater understanding of the physiology of reproductive variation. Jacques Monod's quantitative approach to bacterial growth was seminal and demonstrated that reproduction rate depended on the concentration of a limiting nutrient for a given medium (Monod, 1949). It was later shown that bacterial reproduction rate can also be manipulated by altering the chemical composition of growth media (Schaechter *et al.*, 1958). Regardless of which of these two methods is used, bacteria reproducing at the same steady-state reproduction rate and temperature have a similar biomass composition and overall physiology (Neidhardt *et al.*, 1990). These findings clearly demonstrate that the physiological state of a growing bacterium is dictated by the quantity and quality of nutrients in the environment.

This baseline knowledge was used to determine the physiological changes necessary for a population to transition from slow to rapid growth. When a bacterium in balanced growth is shifted to a medium that supports a faster reproduction rate, a repeatable series of events happen at the cellular level. RNA synthesis increases immediately upon transfer to the nutrient-rich medium, followed a few minutes later by an increase in protein synthesis rate. However, DNA synthesis and cell division continued at the pre-shift rates for a considerable amount of time before eventually increasing to the same rate as all other major biosynthetic reactions (Kjeldgaard *et al.*, 1958). Intriguingly, the rate of protein synthesis per unit RNA was nearly constant at all growth rates measured in these studies (Schaechter *et al.*, 1958).

These results were some of the first to indicate that the protein synthesis apparatus was not operating at a faster rate in rapidly growing cells. Instead, rapidly growing bacteria made more ribosomes to increase the effective protein synthesis rate per cell. Later studies with more precise analytical methods detailed how the protein chain elongation rate and the quantity of ribosomes changed in rapidly growing cells. When the growth rate of an *Escherichia coli* strain is increased from 0.6 to 2.5 doublings per hour the protein chain elongation rate increases only from 13 to 20 amino acid residues per second, while the ribosome content per cell volume increases by a factor of three from 10,800 to 32,000 per μm^3, assuming a change in cell volume from 0.63μm^3 to 2.25μm^3 (Bremer & Dennis, 1996; Donachie and Robinson, 1987). An increased ribosome concentration with higher growth rates has also been observed in oligotrophic bacteria. The number of ribosomes per cell volume in the model oligotrophic bacterium *Sphingopyxis alaskensis* RB2256 increased by a factor of ten, from 4,000 to 40,000 per μm^3, when shifting from a starved, non-growing state to a growth rate of 0.2 doublings per hour (Fegatella *et al.*, 1998). While the highest measured ribosome concentration appears similar between *E. coli* and *S. alaskensis* strains, their maximum growth rates are much different, suggesting that the ribosomes of *S. alaskensis* RB2256 may be operating at a slower chain elongation rate (Fegatella *et al.*, 1998). A comparative study of these two strains, along with eight soil isolates, showed that bacteria which grow more quickly upon nutrient amendment have protein synthesis machinery which operates at a faster rate (Dethlefsen & Schmidt, 2007).

The combined evidence suggests that when any bacterium shifts from slower to faster growth rates it must increase the fraction of its biomass devoted to protein synthesis (Neidhardt *et al.*, 1990). However, the absolute ribosome content is very different for the small cells of *S. alaskensis,* which possess 2,000 ribosomes at maximal reproduction rates, while larger *E. coli*

cell must synthesize 72,000 ribosomes at its maximum growth rate. *E. coli* copes with the biosynthetic burden of high rRNA and tRNA demand by increasing transcription of stable RNA. Twenty-four percent of all active RNA polymerase is synthesizing stable RNA when *E. coli* grows at 0.6 doublings per hour, but this increases to 79% at 2.5 doublings per hour (Bremer & Dennis, 1996). Transcription of the *rrn* operon is under high demand during rapid growth, and it is interesting to note that bacteria with faster maximum recorded growth rates tend to possess a greater number of *rrn* operon copies in their genomes (Vieira-Silva & Rocha, 2010). The *rrn* operons of faster growing bacteria are also more asymmetrically distributed along the chromosome and nearer to the origin of replication than in bacteria with slower growth rates (Couturier & Rocha, 2006). This benefits rapidly growing cells, which possess multiple replication forks, by further boosting the effective number of *rrn* genes per cell (Bremer & Dennis, 1996; Couturier & Rocha, 2006). This suggests that rapid growth selects for increasing genomic *rrn* copy number over evolutionary time.

These observations do not demonstrate a causal relationship between *rrn* copy number and rapid growth. However, the following genetic studies show how rapid growth and *rrn* copy number are related to each other. *E. coli* possesses 7 *rrn* operons in its genome and the functional equivalence among *rrn* copies is generally assumed (Condon *et al.*, 1992; 1995). Genetic inactivation and deletion experiments have shown that if *E. coli* loses even a single *rrn* copy it has an increased lag time when transitioning from slow to rapid growth (Condon *et al.*, 1995; Stevenson & Schmidt, 2004). Inactivating multiple *rrn* copies in combination showed that *E. coli* can lose 1 or 2 *rrn* copies with no detectable change in unconstrained growth rates, but losing 3 or more functional *rrn* copies led to significantly slower unconstrained growth rates (Condon *et al.*, 1995). An *rrn* deletion experiment showed even more extreme phenotypes when the *E. coli*

strain used had been subject to 10,000 generations of experimental evolution and had evolved a faster maximal growth rate and shorter lag time (Vasi *et al.*, 1994). Single *rrn* copy deletion strains had significantly decreased growth rates relative to the parental (evolved) strain, while deleting 2 *rrn* operons further decreased growth rate. Additionally, the magnitude of the growth rate defect in the double knockout increased non-additively when compared to each individual knockout on rich medium (Stevenson & Schmidt, 2004). These genetic studies provide direct evidence that *rrn* multiplicity is adaptive for rapidly growing bacteria, allowing them to achieve fast growth rates and short lag times.

Rapid growth is directly linked to high *rrn* copy number, but can high *rrn* copy number be used to identify rapidly growing bacteria? Experiments using soil bacterial communities have tested if bacteria responding rapidly to nutrient amendments also possess high *rrn* copy number. In the first experiment a soil bacterial community was amended with nutrients on solid growth medium and monitored over hundreds of hours to obtain colony formation curves. Bacteria which formed visible biomass over the first 48 hours of the experiment had a higher mean copy number (5.5 *rrn*) than bacteria which first became visible only after more than 150 hours of incubation (1.4 *rrn*) (Klappenbach *et al.*, 2000). A follow-up experiment tested if a resource pulse to the soil – the herbicide 2,4-D which can be used as a carbon and energy resource – would shift community structure towards bacteria with high *rrn* copy number. The 2,4-D degrading bacterial community in the control microcosm was dominated by a species with 2 *rrn* copies while the amended microcosms were dominated by species with 5 or more *rrn* copies (Klappenbach *et al.*, 2000). A separate study of soil bacterial community succession after rice-paddy flooding indicated that the early-successional communities were dominated by high *rrn* bacteria which formed colonies more quickly in the laboratory, suggesting they are capable of

outcompeting low *rrn* bacteria when environmental conditions quickly change (Shrestha *et al.*, 2007). Finally, a study comparing hundreds of bacterial species identified *rrn* copy number as a strong predictor of maximum recorded growth rates ($R^2 = 0.41$) (Vieira-Silva & Rocha, 2010). Taken together, these physiological, genetic, and ecological studies provide evidence linking rapid bacterial growth to the possession of many *rrn* copies in a genome.

Costs and implications of rapid bacterial growth

Bacteria capable of rapid growth must invest heavily in their protein synthesis capacity but this investment comes with significant costs. The well-studied biochemistry of *E. coli* allows a calculation of the ATP budget of biosynthesis under a variety of nutritional conditions. Protein polymerization is by far the single largest biosynthetic expense of a growing cell, accounting for more than 50% of all ATP spent during growth on glucose with inorganic salts. This energetic cost increases to more than 60% of ATP when amino acids and nucleic acids are provided (Stouthamer, 1973). Rapidly growing bacteria have a high cellular ribosome content, and this up-regulated molecular machine is also expensive to run. In order to fuel rapid growth, bacterial cells must produce large amounts of energy.

Although many heterotrophic bacteria respire when growing slowly on limiting concentrations of sugars, a transition to faster growth at higher sugar concentrations is often correlated with detectable fermentation co-occurring with respiration, a process known as overflow metabolism (Neijssel & Tempest, 1975; Molenaar *et al.*, 2009). This suggests there are constraints on energy production inherent to bacterial catabolism. One such constraint is that efficient (high ATP yield) pathways of glucose catabolism require more enzymatic protein than less efficient pathways, effectively a higher overhead cost for the cell (Flamholz *et al.*, 2013).

More generally, it has been shown that the thermodynamics of heterotrophic metabolism alone lead to an inevitable tradeoff in the rate and yield of ATP production (Pfeiffer & Bonhoeffer, 2002). These studies suggest that rapid growth may be fundamentally incompatible with efficient growth. Evolutionary simulations demonstrate that thermodynamic tradeoffs are sufficient to select for rapidly growing heterotrophic organisms which use respiro-fermentation in homogenous, high resource environments. This study also showed that an efficient, obligately-respiring organism was more competitive at low resource concentrations and when spatial heterogeneity increased (Pfeiffer *et al.*, 2001). A recent study demonstrated that this theoretical study is relevant to evolutionary pressures faced by bacteria in the laboratory. Mutants of *Lactococcus lactis* with an improve ATP yield and a growth rate defect could be selected for using the spatial structure provided by a water-in-oil emulsion (Bachmann *et al.*, 2013). These provocative results beg the question, have these same pressures led to observable rate-efficiency tradeoffs in non-mutagenized populations of evolving bacteria?

The experimental evolution of 12 *E. coli* populations in batch culture provides an excellent test for a rate-efficiency tradeoff because it has selected for faster unconstrained growth rates and decreased lag times (Vasi *et al.*, 1994). While no correlation was observed among the average rate and efficiency of the 12 populations, analyses within populations indicated the presence of a rate-efficiency tradeoff (Novak *et al.*, 2006). It is unclear from this experiment why the tradeoff was seen within populations, but not evident when comparing across all populations. Many potential mechanisms were offered to explain this result: multiple selection pressures, historical contingency leading to population-specific evolutionary constraints, or the experiment not running long enough for the tradeoff to be observed. If this final explanation is correct, it suggests a comparative approach may provide complementary evidence for rate-efficiency

tradeoffs. My thesis is based upon such a comparative study among bacteria with life histories expected to be favored by rapid or efficient growth tactics. Evidence in Chapter 4 is consistent with a rate-efficiency tradeoff underlying the spectrum of life histories from copiotrophy to oligotrophy.

Summary

Bacteria have been evolving for billions of years under a wide-variety of environmental conditions. As a result, extant bacteria exhibit a large variety of maximum specific growth rates (Vieira-Silva & Rocha, 2010). This suggests many evolutionary pressures constrain bacteria from competing in the arena of extremely rapid growth. I propose that an innate growth rate-efficiency tradeoff may underlie bacterial life histories, with abundant resources leading to selection for rapid growth in copiotrophic bacteria. In contrast, chronic poor resource availability in heterogeneous environments has led to selection for efficient growth for oligotrophic bacteria. Any given bacterium has likely been subject to both of these extreme selective pressures at some point over its evolutionary history, but I argue that the number of genomic *rrn* copies of this bacterium is a good proxy for its current adaptations to resource availability.

REFERENCES

REFERENCES

Bachmann H, Fischlechner M, Rabbers I, Barfa N, Branco Dos Santos F, Molenaar D, *et al.* (2013). Availability of public goods shapes the evolution of competing metabolic strategies. *Proc Natl Acad Sci USA* **110**:14302–14307.

Bremer H, Dennis PP. (1996). Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In:*Escherichia coli and Salmonella*, Neidhardt, FC (ed) Vol. 2, ASM Press: Washington, D.C., pp. 1553–1569.

Condon C, Liveris D, Squires C, Schwartz I, Squires CL. (1995). rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. *Journal of Bacteriology* **177**:4152–4156.

Condon C, Philips J, Fu ZY, Squires C, Squires CL. (1992). Comparison of the expression of the seven ribosomal RNA operons in Escherichia coli. *EMBO J* **11**:4175–4185.

Couturier E, Rocha EPC. (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol* **59**:1506–1518.

Dethlefsen L, Schmidt TM. (2007). Performance of the translational apparatus varies with the ecological strategies of bacteria. *Journal of Bacteriology* **189**:3237–3245.

Donachie WD, Robinson AC. (1987). Cell division: Parameter values and the process. In: *Escherichia coli and Salmonella*, Neidhardt, FC (ed) Vol. 2, ASM Press: Washington, D.C., pp. 1578-1593.

Fegatella F, Lim J, Kjelleberg S, Cavicchioli R. (1998). Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. *Appl Environ Microbiol* **64**:4433–4438.

Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R. (2013). Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc Natl Acad Sci USA* **110**:10039–10044.

Hoehler TM, Jørgensen BB. (2013). Microbial life under extreme energy limitation. *Nat Rev Micro* **11**:83–94.

Kempes CP, Dutkiewicz S, Follows MJ. (2012). Growth, metabolic partitioning, and the size of microorganisms. *Proc Natl Acad Sci USA* **109**:495–500.

Kjeldgaard N, Maaloe O, Schaechter M. (1958). The transition between different physiological states during balanced growth of Salmonella typhimurium. *Microbiology* **19**:607.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**:1328–1333.

Koch A. (2001). Oligotrophs versus copiotrophs. *Bioessays* **23**:657–661.

Kuznetsov S, Dubinina G, Lapteva N. (1979). Biology of oligotrophic bacteria. *Annual Reviews in Microbiology* **33**:377–387.

Maclean RC. (2008). The tragedy of the commons in microbial populations: insights from theoretical, comparative and experimental studies. *Heredity* **100**:471–477.

Molenaar D, Berlo RV, Ridder D de, Teusink B. (2009). Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* **5**:1–10.

Monod J. (1949). The growth of bacterial cultures. *Annual Reviews in Microbiology* **3**:371–394.

Neidhardt FC. (1999). Bacterial Growth: Constant Obsession withdN/dt. *Journal of Bacteriology* **181**:7405–7408.

Neidhardt FC, Ingraham JL, Schaechter M. (1990). Physiology of the Bacterial Cell: A Molecular Approach. Sinauer Associates: Sunderland, MA.

Neijssel O, Tempest D. (1975). The regulation of carbohydrate metabolism in Klebsiella aerogenes NCTC 418 organisms, growing in chemostat culture. *Archives of Microbiology* **106**:251–258.

Novak M, Pfeiffer T, Lenski R, Sauer U, Bonhoeffer S. (2006). Experimental tests for an evolutionary trade-off between growth rate and yield in E. coli. *American Naturalist* **168**:242–251.

Pfeiffer T, Bonhoeffer S. (2002). Evolutionary consequences of tradeoffs between yield and rate of ATP production. *Zeitschrift für Physikalische Chemie* **216**:51.

Pfeiffer T, Schuster S, Bonhoeffer S. (2001). Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**:504–507.

Roller BRK, Schmidt TM. (2015). The physiology and ecological implications of efficient growth. *ISME J* **9**:1481–1487.

Russell JB, Cook GM. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* **59**:48–62.

Schaechter M, Maaloe O, Kjeldgaard N. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced growth of Salmonella typhimurium. *Microbiology* **19**:592.

Schmidt TM, Konopka AE. (2009). Physiological and Ecological Adaptations of Slow-Growing, Heterotrophic Microbes and Consequences for Cultivation. In:*Microbiology Monographs*, Microbiology Monographs Vol. 10, Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 101–120.

Shrestha PM, Noll M, Liesack W. (2007). Phylogenetic identity, growth-response time and rRNA operon copy number of soil bacteria indicate different stages of community succession. *Environmental Microbiology* **9**:2464–2474.

Stearns SC. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften* **87**:476–486.

Stearns SC. (1976). Life-History Tactics: A Review of the Ideas. *The Quarterly Review of Biology* **51**:3–47.

Stevenson BS, Schmidt TM. (2004). Life history implications of rRNA gene copy number in Escherichia coli. *Appl Environ Microbiol* **70**:6670.

Stouthamer AH. (1973). A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* **39**:545–565.

Vasi F, Travisano M, Lenski RE. (1994). Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. *American Naturalist* 432–456.

Vieira-Silva S, Rocha EPC. (2010). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet* **6**:e1000808.

CHAPTER 2

The physiology and ecological implications of efficient growth

This research was first published in the article: Roller, BRK and TM Schmidt. The physiology and ecological implications of efficient growth. The ISME Journal (2015) 9, 1481-1487; doi:10.1038/ismej.2014.235; published online 9 January 2015.

Abstract

The natural habitats of microbes are typically spatially structured with limited resources, so opportunities for unconstrained, balanced growth are rare. In these habitats, selection should favor microbes that are able to use resources most efficiently, that is, microbes that produce the most progeny per unit of resource consumed. On the basis of this assertion, we propose that selection for efficiency is a primary driver of the composition of microbial communities. In this article, we review how the quality and quantity of resources influence the efficiency of heterotrophic growth. A conceptual model proposing innate differences in growth efficiency between oligotrophic and copiotrophic microbes is also provided. We conclude that elucidation of the mechanisms underlying efficient growth will enhance our understanding of the selective pressures shaping microbes and will improve our capacity to manage microbial communities effectively.

Introduction

The conceptual foundation of microbial physiology was built on studies of microbes during balanced growth in homogenous cultures. Schaechter *et al.* (1958) first established that bacteria adjust their macromolecular composition to match growth rate. The rate of a population's exponential growth is an integrated signal composed of the chemical and physical features of an environment. As stated elegantly by Neidhardt (1999), '…when growth is the ultimate interest, one cannot long delve into single enzymes and genes, or even individual pathways and mechanisms, without at some point returning to the whole cell and asking about the coordinated operation of processes.' The growth rate of a microbe—the number of progeny produced per unit time—is an important component of fitness in most environments and

therefore a pivotal life history trait. Any trait that impacts fitness by altering a microbe's reproduction or survival is termed a life history trait, and these include responses to varying resource availability, population density and other extracellular factors (Vasi *et al.,* 1994). The collection of life history traits define a microbe's life history—the overall pattern of reproduction and survival.

A less obvious, but equally important life history trait is the efficiency of microbial growth—the number of progeny produced per unit of resource consumed. Like growth rate, growth efficiency integrates a microbe's physiology, ecology and evolutionary history. The efficiency of growth for any given microbe depends on multiple environmental and population-specific factors, including the free energy available from a resource (Linton and Stephenson, 1978), pathways for resource utilization (Flamholz *et al.,* 2013), the availability of precursors for biomass synthesis (Stouthamer, 1973) and the fraction of available energy devoted to maintenance functions instead of growth (Hoehler and Jørgensen, 2013). We do not yet know the collection of specific genetic determinants that underlie growth efficiency, but as described below, it is obvious that efficiency is a life history trait and is under selection in most environments. Although our ultimate goal is to understand all elements of a microbes fitness in concert, including both reproduction and survival components, our primary focus in this work is on two life history traits that impact reproduction—the rate and efficiency of population growth.

Growth efficiency is important from an evolutionary perspective and has repercussions for understanding how ecosystems function. Carbon use efficiency (CUE), the amount of carbon incorporated into biomass per carbon resource consumed, is one way to measure growth efficiency and is a proxy for the number of progeny produced per unit resource. It also provides a quantitative measure of the impact that microbes have on nutrient cycling in an ecosystem. In

animals' digestive tracts, microbes impact many essential processes for the holoorganism, especially the nutritional value extracted from their diets (McFall-Ngai *et al.,* 2013) where microbial CUE is likely a critical variable. Heterotrophic microbes are major contributors to the global carbon cycle (Cho and Azam, 1988; Singh *et al.,* 2010)—respiring 60 gigatonnes of terrestrial organic matter to carbon dioxide (CO_2) each year, roughly six times more than annual anthropogenic emissions (Trivedi *et al.,* 2013)—yet we are just beginning to understand the efficiency of carbon use by heterotrophic microbes and its impact on ecosystem carbon cycling (Manzoni *et al.,* 2012; Lee and Schmidt, 2014). Exploring variations in CUE among diverse microbes will improve our knowledge of how microbial communities impact carbon flux, from the small scale of host–microbiome interactions to largescale annual CO_2 flux from an ecosystem.

We address two primary questions in this perspective: which environmental characteristics favor efficiency and what is the extent of plasticity in growth efficiency of individual microbes? Key findings are illustrated using aerobic heterotrophs, but should also apply to fermentative microbes and those that respire any of an array of terminal electron acceptors other than oxygen (O_2). In regards to the terminology used to describe efficiency, ecologists often measure growth efficiency in carbon (C) units, that is, moles of C incorporated into biomass per mole of C consumed, and use the terms CUE, microbial growth efficiency and bacterial growth efficiency interchangeably to describe this measure. Microbial physiologists and engineers more often describe efficiency in terms of yield. Yields are expressed in units that are not as easily compared across microbial populations or growth conditions, for example, biomass per gram of resource, per mole ATP or per mole of electrons. We have elected to use CUE as a

measure of efficiency in addition to progeny per resource measurements. CUE varies between 0 and 1 and provides an intuitive comparison across organisms and resources.

When is efficient growth favored?

An intriguing study of how spatial heterogeneity and varying resource availability influences selection on growth rate and growth efficiency was conducted using mathematical simulations of heterotrophic microbes. In these simulations, two types of 'organisms' competed across gradients of spatial structure and resource flux. One was a rapidly growing, inefficient, respiro-fermentative organism. The other was an efficient, but slow growing, obligately respiring organism. Efficient growth was favored over rapid growth when the flux of resources was low and spatial heterogeneity was high. As the flux of resources increased and the environment became more homogeneous, the rapidly growing organism was favored (Pfeiffer *et al.,* 2001). Potential tradeoffs between growth rate and growth efficiency have also been evaluated experimentally with genetically modified yeast strains. The competing strains were isogenic except for a single mutation that made one strain capable of using only the more efficient process of respiration but slowed growth rate. The other strain gained energy primarily through the less efficient process of fermentation and grew more rapidly (Maclean and Gudelj, 2006). When these strains competed in a homogenous, continuous culture, the rapid growing organism was more fit and outcompeted the efficient organism. Altering only the temporal availability of resources by using batch culture or in combination with spatial structure by using a metapopulation of batch cultures, allowed for the coexistence of rapid and efficient strains. Taken together, these studies indicate that a few key factors—low resource concentrations,

spatial heterogeneity and temporal resource dynamics—can increase the fitness of efficient strains.

Typical laboratory cultivation differs from the conditions microorganisms experience in their natural environments, where spatial heterogeneity is pervasive (Stocker, 2012). The lack of spatial or temporal structure in typical laboratory cultivation causes resources to be a global commodity shared by the entire experimental population. Selection therefore favors rapid growth in both batch and continuous cultivation evolution experiments (Dykhuizen and Hartl, 1981; Vasi *et al.,* 1994). Spatial heterogeneity can be accomplished in a laboratory setting by using an oil emulsion that compartmentalizes individual microbes. In one study, a population of randomly mutagenized *Lactococcus lactis* was serially propagated in an oil emulsion. This led to a rate efficiency tradeoff between isolated clones. Isolates with increased growth efficiency relative to the parental strain of *L. lactis* were observed, yet they typically grew slower than inefficient clones in the population. The clone with the greatest yield and slowest growth rate had a large increase in relative abundance throughout the 28 days of propagation, demonstrating selection favoring efficient growth (Bachmann *et al.,* 2013).

Spatial heterogeneity, low resource concentrations and temporal resource dynamics can all favor efficient growth because each of these factors influence the scale of competition, effectively privatizing resources to individuals and shifting the cost of inefficient resource use from the community to the individual. Populations founded by inefficient organisms will be smaller than those founded by efficient organisms, given the same amount of resource for each population (Pfeiffer *et al.,* 2001). Competition between organisms capable of achieving varied population sizes in spatially structured environments can lead to counterintuitive results. Using genetically modified strains of *Escherichia coli*, Chuang *et al.* (2009) showed that selection

could favor a strain that produced a larger final population size in a spatially structured metapopulation, even when it was at a growth rate disadvantage, a phenomenon that has been described statistically as the Yule–Simpson effect. This phenomenon demonstrates that considering the environmental context in which selection acts on microorganisms is critical.

Two key variables, resource availability and the free energy content of these resources, have large impacts on the physiology and growth efficiency of microbial populations and deserve a more detailed discussion and analysis. The framework presented for how these variables impact growth efficiency will then be used to develop a model of growth efficiency that distinguishes two distinct life histories.

Growth efficiency varies with resource availability

Heterotrophic bacteria utilize organic compounds for two primary purposes: as a source of energy (extracted through catabolism) and as a source of carbon molecules to build biomass (anabolic reactions). The fractionation of carbon between catabolism and anabolism varies within and between organisms. One factor that modulates the fractionation of carbon within an organism is growth rate. At submaximal growth rates, bacteria uncouple anabolism from catabolism (Tempest and Neijssel, 1984) and a larger fraction of the cell's energy budget is devoted to maintenance functions rather than biomass synthesis (Tempest and Neijssel, 1984; Russell and Cook, 1995). The decreased proportion of biosynthesis in the energy budget at submaximal growth rates can be measured by tracking the fate of carbon to biomass or CO_2.

Previously published measures of the fractionation of carbon between catabolism and anabolism are often reported as biomass yields using mass units. Because yields are resource dependent, comparing yields across different resources requires converting the data to a common

currency. We convert growth yields to CUE—moles of carbon incorporated into biomass per total moles of carbon consumed (Equation 1a). During non-fermentative growth, the total carbon utilized by heterotrophs is equivalent to carbon incorporated into biomass production plus carbon respired, and CUE can be expressed as in Equation 1b. Oxygen consumption and dry biomass measurements can also be used to calculate growth efficiency, as in Equation 1c, when there is a consistent carbon content in the biomass of the organism and the respiratory quotient (RQ, the ratio of CO_2 produced per O_2 consumed) reflects complete oxidation of the substrate. We calculated CUE, using Equations 1b and c, and the steady-state concentration of the limiting resource in a series of chemostat experiments with strains of *Klebsiella aerogenes* (Herbert, 1976; Neijssel and Tempest, 1976). The assumptions that the biomass has a constant carbon content and the carbon source is completely oxidized are reasonable for this organism during carbon-limited growth (Neijssel and Tempest, 1975; Herbert, 1976), so we are able to quantify the impact of resource availability on CUE.

Equation 1

(1a) (1b) (1c)

$$\mathrm{CUE} = \frac{\text{Biomass C (mol)}}{\text{Total C Utilized (mol)}} = \frac{\text{Biomass C}}{\text{Biomass C + Respired C}} = \frac{(\text{dry biomass} \times \%\text{C in biomass})}{(\text{dry biomass} \times \%\text{C in biomass}) + (\text{RQ} \times \text{O}_2\ \text{Utilized})}$$

To calculate the concentration of the limiting resource (R) we used Equation 2, where R is a function of the dilution rate (D) and physiological properties of the organism: the maximum specific growth rate (μ_{max}) and the concentration of the limiting resource that supports the organism's growth at half of the maximal rate (K_s).

Equation 2

$$R = \frac{D \times K_S}{\mu_{max} - D}$$

During glycerol-limited growth of *K. aerogenes* in a chemostat culture, a larger fraction of the total carbon consumed is assimilated into biomass as growth rate increases and a smaller proportion is required for respiration (Figure 2.1a). All carbon is accounted for at each steady-state growth rate of *K. aerogenes* (biomass C + respired C), indicating that partial oxidation products of glycerol are not accumulating in the medium.

Converting the carbon metabolism data in Figure 2.1a into CUE (Equation 1) and plotting this against the steady-state limiting resource concentration (Equation 2) provides insight into the relationship between these parameters (Figure 2.1b). Growth efficiency increases as the limiting resource concentration is raised during glycerol-limited growth, until reaching a plateau of maximum efficiency at higher resource concentrations. The same relationship between growth efficiency and resource concentration is apparent regardless of whether growth is phosphate, sulfate or glucose limited, assuming complete carbon source oxidation (Figure 2.1c). K_s values for glucose (Neijssel and Tempest, 1975) and sulfate (Owens and Legan, 1987) were taken from the literature on *K. aerogenes*, whereas the values for phosphate (Owens and Legan, 1987) and glycerol (Neijssel and Tempest, 1975; Owens and Legan, 1987) are derived from *E. coli.* Although this historical data may not perfectly reflect the K_s values realized during the original experiment, the same relationship with CUE is observed even when manually altering K_s values within a larger range of values of closely related organisms from the literature (data not shown). In addition, in the sulfate- and phosphate-limited cultures, the lowest concentrations of limiting

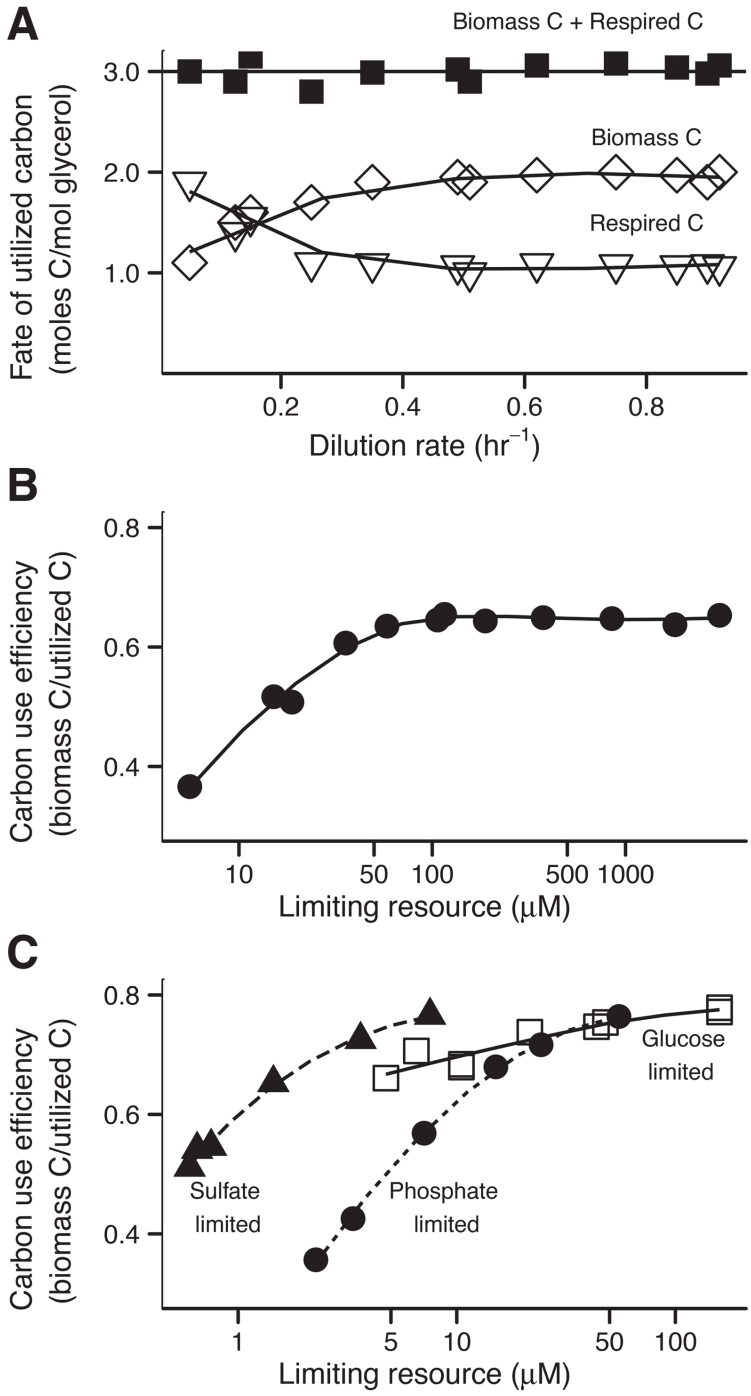


Figure 2.1: Efficiency varies with resource concentration. The influence of limiting resources on the CUE of *K. aerogenes* NCTC 418 in chemostat cultures. (a) Allocation of carbon in a glycerol-limited chemostat culture as a function of dilution rate (Herbert, 1976). (b) Variation in CUE related to the steady-state glycerol concentration (calculated from Herbert, 1976). (c) Relationship between CUE and the steady-state limiting resource concentration in glucose-, phosphate- or sulfate-limited conditions (calculated from Neijssel and Tempest, 1976). Curve fitting in all panels was generated using a locally weighted regression algorithm (LOESS) to help visualize trends.

resource do not always lead to complete carbon source oxidation (Neijssel and Tempest, 1975). This is based on the measurements at a single dilution rate and will not alter our interpretations and conclusions because it will lead to a lower CUE than was calculated at the lowest resource concentrations.

These results imply that carbon metabolism becomes more efficient as growth becomes less nutritionally constrained, with an organism's maximal growth efficiency reached near unconstrained, balanced growth. As discussed below, maximal CUE is also specific to the organic source being metabolized and likely specific for the entire physical and chemical environment.

Growth efficiency varies with resource quality

Growth efficiency is also dependent on the amount of energy captured during the oxidation of different organic compounds. To demonstrate the magnitude of changes in CUE due to the energy content of different carbon sources, we gathered the data from batch cultivation experiments in which 10 different species of bacteria were grown in minimal media with different organic compounds serving as the sole carbon and energy source (Linton and Stephenson, 1978). We calculated CUE assuming biomass had a constant carbon content for all organisms (Simon and Azam, 1989) and plotted against the heat of combustion per carbon atom in the organic compound supporting growth (Figure 2.2). Bacteria growing on resources with small amounts of free energy per carbon atom must use energy to reduce the carbon to the oxidation state of their biomass. This increased demand for energy, in the form of reducing equivalents, decreases overall CUE on low energy resources. When the energy content of the carbon in the resource and biomass (calculated from Cordier *et al.,* 1987) is similar, the

efficiency of growth stops increasing. This phenomenon has been reported for microbes in soils, as well as in pure culture (Manzoni *et al.,* 2012).

Batch culture allows for unconstrained, balanced growth conditions and, we argue, a microbe's maximal efficiency for that environment. Yet, we still see that growth efficiency depends on the energy content of the carbon resource, even when the resource is provided in quantities that far exceed biosynthetic demand. Despite the increased variability introduced by comparing 10 different bacteria in this analysis, a strong relationship between efficiency and energy content of the carbon resource is observed.

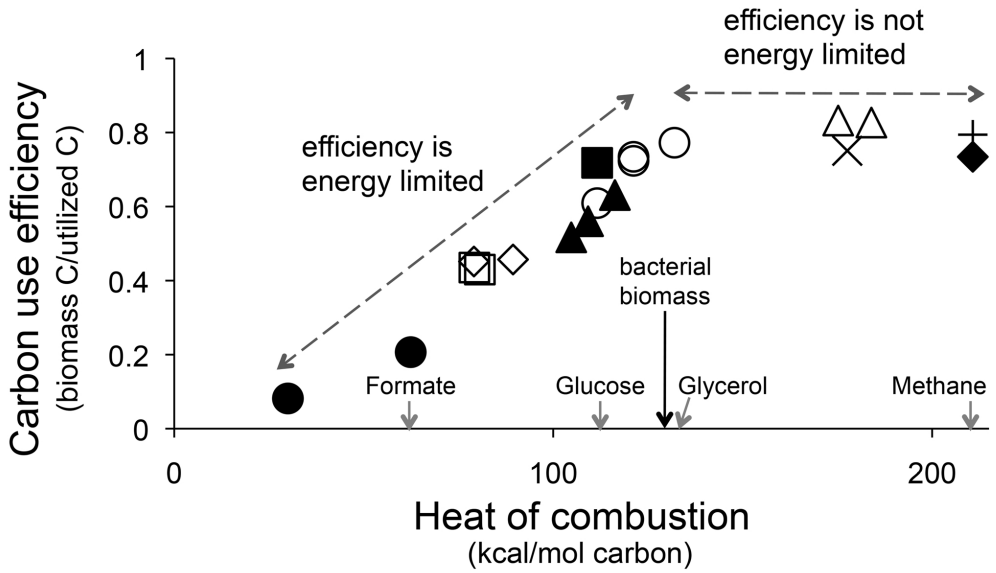


Figure 2.2: Efficiency varies with resource quality. CUE of 10 bacterial species related to the energy content of the organic compound supporting growth (calculated from Linton and Stephenson, 1978). Distinct symbols represent different bacterial species. The heat of combustion of representative organic compounds and an average for bacterial biomass (*E. coli* (130.2) and *M. methylotrophus* (132.5), calculated from Cordier *et al.,* 1987) are presented on the *x* axis.

Life history and growth efficiency

One of the best-known distinctions of microbial life histories is the copiotroph–oligotroph dichotomy. Copiotrophic microbes are selected for rapid growth when resources are abundant, whereas oligotrophic microbes have adaptations for growth in persistently resource-poor environments (Koch, 2001). This framework has many similarities to the concepts underlying *r/K* selection theory in macroecology, where the selective pressures on organisms are a function of resource availability or population density. Unlike *r/K* selection, the copiotroph–oligotroph dichotomy does not require differential patterns of survival or persistence.

There is a striking parallel between the conditions where oligotrophs thrive and the conditions that select for efficient organisms. Many ecosystems that have considerable oligotroph membership, such as the open ocean (Vergin *et al.,* 2013) and bulk soil (Fierer *et al.,* 2007), contain habitats with low resource availability and spatial structure that should select for efficient microbial growth. We propose a conceptual model that outlines how growth efficiency varies between an archetypical copiotroph and oligotroph as a function of resource concentration and quality (Figure 2.3).

The relationship between growth efficiency and resource concentration in our proposed model follows the same general pattern for both the oligotroph and copiotroph. The capacity for growth is indicated by the extent of the growth efficiency function, which terminates when growth is no longer supported. These two life histories have distinguishing features in the relationship between efficiency and concentration of the limiting resource in their environment.

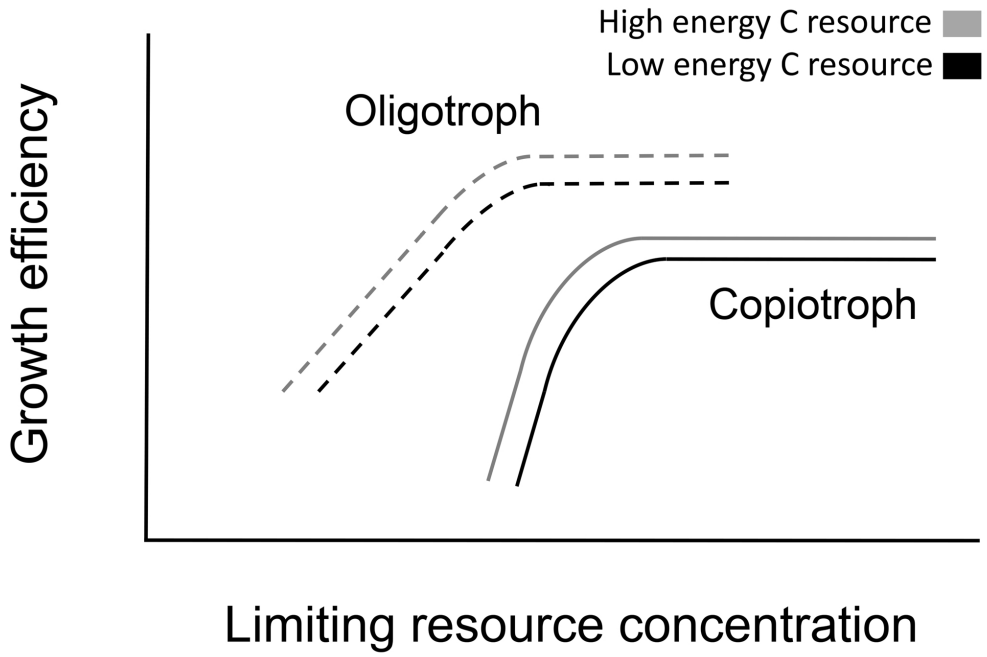


Figure 2.3: Conceptual model of life history and efficiency. Proposed model of growth efficiency for distinct bacterial life histories. The efficiency of copiotrophic (solid lines) and oligotrophic (dashed lines) bacteria should be compared on resources with the same energy content (indicated by line color) and at the same limiting resource concentration.

The proposed model contains the following elements and hypotheses:

1. As has been proposed previously (Zhao *et al.,* 2013), oligotrophs are superior competitors for resources at low resource concentrations. This is visualized as the oligotroph having the capacity to grow at a much lower concentration of a limiting resource than the copiotroph.
2. On the basis of the evidence from Figure 2.1, growth efficiency increases for both copiotrophs and oligotrophs as the limiting resource concentration increases up to a threshold where maximum efficiency is achieved. As a smaller proportion of carbon

metabolism is directed towards maintenance energy, efficiency increases until it reaches a maximum near balanced growth.

3. We hypothesize that maintenance energy is lower for oligotrophs compared with copiotrophs. There are two consequences of this in our model: the minimum growth efficiency is higher for the oligotroph and the rate of increase in growth efficiency is slower. This extends the range of resource concentrations supporting the oligotroph's growth. Eight cellular functions have been described as the components of maintenance energy (van Bodegom, 2007) and oligotrophs have been shown to minimize costs associated with three of these functions—protection from oxygen stress, cell motility, and the synthesis and turnover of macromolecules. A large clade of marine oligotrophs have lost the capacity to synthesize oxygen stress protectants when they are freely available in their environment (Morris *et al.,* 2012) and many described oligotrophs are also non-motile (Lauro *et al*., 2009; Stocker, 2012). Additionally, genome streamlining is common in oligotrophs (Giovannoni *et al.,* 2014), which may be an adaptation to decrease the amount of resources invested in macromolecule synthesis and turnover. Taken individually, any one of these traits is not exclusive to, or universally present in, oligotrophs. However, there is a tendency towards minimizing maintenance energy costs in oligotrophs and more work must be done to evaluate this hypothesis.
4. On the basis of the evidence from Figure 2.2, growth efficiency increases for both copiotroph and oligotroph as the energy content of the resource is increased. This is displayed in our model in Figure 2.3, as gray versus black lines.
5. We propose that the maximal growth efficiency of oligotrophs is higher than copiotrophs and that it is reached at a lower concentration of limiting resource. This is supported by

evidence from one of the few comparisons of an oligotroph, *Sphingopyxis alaskensis*, and a copiotroph, *Vibrio angustum*, under identical conditions (Cavicchioli *et al.,* 2003). The oligotroph had a greater population density at all dilution rates, and thus resource concentrations, measured in the chemostat.

6. It has been postulated that oligotrophs grow less well, or not at all, in resource-rich environments (Koch, 2001). This is captured by the termination of the oligotroph's growth efficiency functions at a lower resource concentration than the copiotroph's.

Although many of our hypotheses are built upon observations from chemostat cultures, insufficient physiological data are available for generating hypotheses at extremely low resource concentrations—corresponding to very slow or non-growing states. Technical limitations of cultivation technology are largely responsible for this lack of data, but the physiology of extreme resource starvation, where reproduction and survival processes co-occur, likely has a large role in determining microbial fitness in natural environments. It is tempting to speculate that if oligotrophs are more efficient in all physiological states, they would have an increased carrying capacity, the parameter *K* in *r/K* selection, relative to copiotrophs in the same conditions. However, there are not enough data to support a universal prediction of oligotrophs possessing increased carrying capacity, persistence or other *K*-selected traits that are not directly linked to growth. We believe it is important to make predictions about growth physiology during extreme starvation, but more data are needed to understand the interplay between growth and persistence in near non-growth conditions for all life histories.

We hope these hypotheses will stimulate critical discussion of the many potential mechanisms underlying the growth efficiency of microbial populations. The physiology of an

individual microbe encompasses thousands of individual reactions and growth efficiency integrates these reactions into an emergent phenotype. In addition, growth efficiency directly interacts with both ecological and evolutionary processes in microbial communities. All natural microbial environments contain spatial structure, resource limitation or temporal resource dynamics. Therefore, all natural microbial environments, from relatively stable syntrophically-associated subsurface communities to dynamic host–microbe systems, must impart some selective pressure for efficient growth on their microbial assemblages. Although the consequences of these selective pressures for ecosystem functioning are unclear, any attempts to manage microbial communities must recognize evolutionary pressures favoring efficient growth are most likely present in natural microbial systems. Just as past microbiologists have used growth rate to better understand the coordination of cellular processes necessary for reproduction, modern microbiologists have the opportunity to use growth efficiency to unify our understanding of the physiological, ecological and evolutionary processes shaping microbial communities.

Acknowledgements

We would like to acknowledge Arvind Venkataraman, Byron Smith, Clive Waldron, Alex Schmidt and Zarraz Lee for valuable feedback throughout the writing process. This work was supported in part by the Department of Energy Office of Science Graduate Fellowship Program (DOE SCGF), made possible in part by the American Recovery and Reinvestment Act of 2009, administered by ORISEORAU under contract no. DE-AC05-06OR23100; the National Science Foundation's Long-Term Ecological Research Program through grant no. DEB 1027253 and the National Institutes of Health (GM0099549).

REFERENCES

REFERENCES

Bachmann H, Fischlechner M, Rabbers I, Barfa N, Branco Dos Santos F, Molenaar D *et al.* (2013). Availability of public goods shapes the evolution of competing metabolic strategies. *Proc Natl Acad Sci USA* **110**: 14302–14307.

Cavicchioli R, Ostrowski M, Fegatella F, Goodchild A, Guixa-Boixereu N. (2003). Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of *Sphingopyxis alaskensis* (formerly *Sphingomonas alaskensis)*. *Microbl Ecol* **45**: 203–217.

Cho BC, Azam F. (1988). Major role of bacteria in biogeochemical fluxes in the ocean's interior. *Nature* **332**: 441–443.

Chuang J, Rivoire O, Leibler S. (2009). Simpson's paradox in a synthetic microbial system. *Science* **323**: 272–275.

Cordier J, Butsch B, Birou B, Stockar U. (1987). The relationship between elemental composition and heat of combustion of microbial biomass. *Appl Microbiol Biotechnol* **25**: 305–312.

Dykhuizen D, Hartl D. (1981). Evolution of competitive ability in *Escherichia coli*. *Evolution 581–594.*

Fierer N, Bradford M, Jackson R. (2007). Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354–1364.

Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R. (2013). Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc Natl Acad Sci USA* **110:** 10039–10044.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* **8**: 1553–1565.

Herbert D. (1976). Stoicheiometric aspects of microbial growth. In Dean A, Ellwood DC, Evans C, Melling J (eds) *Continuous Culture 6: Applications and New Fields*. Ellis Horwood: Chichester, UK, pp 1–30.

Hoehler TM, Jørgensen BB. (2013). Microbial life under extreme energy limitation. *Nat Rev Microbiol* **11**: 83–94.

Koch A. (2001). Oligotrophs versus copiotrophs. *Bioessays* **23**: 657–661.

Lauro F, McDougald D, Thomas T, Williams T, Egan S, Rice S *et al*. (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106**: 15527–15533.

Lee ZM, Schmidt TM. (2014). Bacterial growth efficiency varies in soils under different land management practices. *Soil Biol Biochem* **69:** 282–290.

Linton J, Stephenson R. (1978). A preliminary study on growth yields in relation to the carbon and energy content of various organic growth substrates. *FEMS Microbiol Lett* **3:** 95–98.

Maclean RC, Gudelj I. (2006). Resource competition and social conflict in experimental populations of yeast. *Nature* **441**: 498–501.

Manzoni SS, Taylor PP, Richter AA, Porporato AA, Agren GIG. (2012). Environmental and stoichiometric controls on microbial carbon-use efficiency in soils. *New Phytol* **196**: 79–91.

McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Losˇo T, Douglas AE *et al.* (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci USA* **110:** 3229–3236.

Morris JJ, Lenski RE, Zinser ER. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**: e00036–12.

Neidhardt FC. (1999). Bacterial growth: constant obsession with dN/dt. *J Bacteriol* **181**: 7405–7408.

Neijssel O, Tempest D. (1976). Bioenergetic aspects of aerobic growth of *Klebsiella aerogenes* NCTC 418 in carbon-limited and carbon-sufficient chemostat culture. *Arch Microbiol* **107**: 215–221.

Neijssel O, Tempest D. (1975). The regulation of carbohydrate metabolism in *Klebsiella aerogenes* NCTC 418 organisms, growing in chemostat culture. *Arch Microbiol* **106:** 251–258.

Owens J, Legan J. (1987). Determination of the Monod substrate saturation constant for microbial growth. *FEMS Microbiol Lett* **46:** 419–432.

Pfeiffer T, Schuster S, Bonhoeffer S. (2001). Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**: 504–507.

Russell JB, Cook GM. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* **59**: 48–62.

Schaechter M, Maaloe O, Kjeldgaard N. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonella typhimurium.* *Microbiology* **19:** 592.

Simon M, Azam F. (1989). Protein content and protein synthesis rates of planktonic marine bacteria. *Mar Ecol Prog Ser* **51:** 201–213.

Singh BK, Bardgett RD, Smith P, Reay DS. (2010). Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nat Rev Microbiol* **8:** 779–790.

Stocker R. (2012). Marine microbes see a sea of gradients. *Science* **338:** 628–633.

Stouthamer AH. (1973). A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* **39:** 545–565.

Tempest D, Neijssel O. (1984). The status of YATP and maintenance energy as biologically interpretable phenomena. *Ann Rev Microbiol* **38:** 459–513.

Trivedi P, Anderson IC, Singh BK. (2013). Microbial modulators of soil carbon storage: integrating genomic and metabolic knowledge for global prediction. *Trends Microbiol* **21:** 641–651.

van Bodegom P. (2007). Microbial maintenance: a critical review on its quantification. *Microb Ecol* **53:** 513–523.

Vasi F, Travisano M, Lenski RE. (1994). Long-term experimental evolution in *Escherichia coli.* II. Changes in life-history traits during adaptation to a seasonal environment. *Am Nat 432–456.*

Vergin KL, Done B, Carlson CA, Giovannoni SJ. (2013). Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. *Aquat Microb Ecol* **71:** 1–13.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013). Abundant SAR11 viruses in the ocean. *Nature* **494**: 357–360.

CHAPTER 3

*rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development

This research was first published in the article: Stoddard, SF, BJ Smith, R Hein, BRK Roller and TM Schmidt. *rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research (2015) 43, D593-598; doi:10.1093/nar/gku1201; published online 20 November 2014.

This publication was the collaboration of many people. TM Schmidt, BRK Roller, and SF Stoddard envisioned the strategy for automating updates to the database and linking the public database to an internal database of annotated genomic information. SF Stoddard and RH Hein provided technical expertise to build and maintain database. BRK Roller, SF Stoddard, BJ Smith, and TM Schmidt created and implemented quality control tests. All authors wrote and revised the paper.

Abstract

Microbiologists utilize ribosomal RNA genes as molecular markers of taxonomy in surveys of microbial communities. rRNA genes are often co-located as part of an *rrn* operon, and multiple copies of this operon are present in genomes across the microbial tree of life. *rrn* copy number variability provides valuable insight into microbial life history, but introduces systematic bias when measuring community composition in molecular surveys. Here we present an update to the ribosomal RNA operon copy number database (*rrn*DB), a publicly available, curated resource for copy number information for bacteria and archaea. The redesigned *rrn*DB (http://rrndb.umms.med.umich.edu/) brings a substantial increase in the number of genomes described, improved curation, mapping of genomes to both NCBI and RDP taxonomies, and refined tools for querying and analyzing these data. With these changes, the *rrn*DB is better positioned to remain a comprehensive resource under the torrent of microbial genome sequencing. The enhanced *rrn*DB will contribute to the analysis of molecular surveys and to research linking genomic characteristics to life history.

Introduction

In bacteria and archaea, the ribosomal RNA operon (*rrn*) typically codes for the 16S, 23S and 5S rRNAs. Together with a suite of proteins, these form ribosomes—the molecular machines responsible for catalyzing the mRNA-dependent polymerization of amino acids into protein. Unlike most bacterial and archaeal genes, the rRNA operon is frequently found in multiple copies, from 1–15 in bacteria and 1–4 in archaea (Klappenbach *et al.,* 2001). It has been suggested that *rrn* copy number is an index of microbial life histories, wherein rapid growth in response to favorable conditions and high translational power (copiotrophic life

history traits) are positively correlated with *rrn* copy number (Klappenbach *et al.,* 2000; Dethlefsen and Schmidt, 2007), and oligotrophic organisms tend to have low copy number (Eichorst *et al.,* 2007; Cavicchioli *et al.,* 2003). Due to the central importance of ribosomal RNAs in the formation of peptide bonds (Schuwirth *et al.,* 2005), rRNA genes share regions of highly conserved sequence that are interspersed with more variable regions. These characteristics make the 16S gene a useful phylogenetic marker, key to our modern understanding of the evolutionary relationships among microbes.

The abundance of sequence data and knowledge about secondary structure has made the 16S gene the most popular target for culture-independent, sequence-based methods in microbiology. With the rapidly shrinking cost of sequencing, whole community 16S surveys have become a core tool in microbial ecology. Curated databases of aligned 16S sequences, including SILVA (Quast *et al.,* 2013), the Ribosomal Database Project (RDP) (Cole *et al.,* 2014) and Greengenes (DeSantis *et al.,* 2006), have been developed to facilitate analysis of sequence data. Analysis pipelines usually produce estimates of per-taxon relative abundances based on the number of copies of 16S genes recovered in a sequence library.

Unfortunately, given the variable per-genome copy number of the 16S gene, a frequently recovered sequence may represent a high copy number taxon of lesser abundance, or a low copy number taxon of higher abundance. Inferences based on relative abundance of 16S genes may therefore not be representative of true community structure (Kembel *et al.,* 2012). This can be an important source of systematic bias in 16S surveys, along with differential DNA extraction and polymerase chain reaction amplification (Pinto and Raskin 2012; Yuan *et al.,* 2012; Morgan *et al.,* 2010). Given knowledge of 16S gene copy number, molecular surveys can be corrected to remove this bias.

By mapping recovered sequences to available microbial genomes based on similarity in the 16S gene, the 16S copy number of the organism that contributed the sequence can be estimated and survey data adjusted accordingly. This general approach has been implemented in several software packages, including CopyRighter (Angly *et al.,* 2014), pplacer and the picante R package (Kembel *et al.,* 2012), and incorporated into PICRUSt (Langille *et al.,* 2013). The accuracy of these methods depends on a reference database of known 16S copy numbers mapped to a taxonomy or phylogeny. The *rrn*DB is a carefully curated, publicly available resource for copy number information, which can be easily integrated into existing correction methods.

Here we introduce an updated version of the *rrn*DB providing 16S copy number information derived from a new data source, a new website with expanded features, and mechanisms for maintaining concurrency with new genomes as they are published. At manuscript submission the database included 2635 bacterial records representing 1383 species, and 175 archaeal records representing 148 species. We foresee the new *rrn*DB contributing to improved copy number correction in metagenomic surveys. Further, the changes create a more robust platform for continued development as a resource supporting functional studies involving *rrn* copy number and life history strategies of bacteria and archaea. The new *rrn*DB is available on the WWW at the URL http://rrndb.umms.med.umich.edu/

Database description

Major improvements to the website and database include: expanded organism taxonomies to include both the National Center for Biotechnology Information (NCBI) and

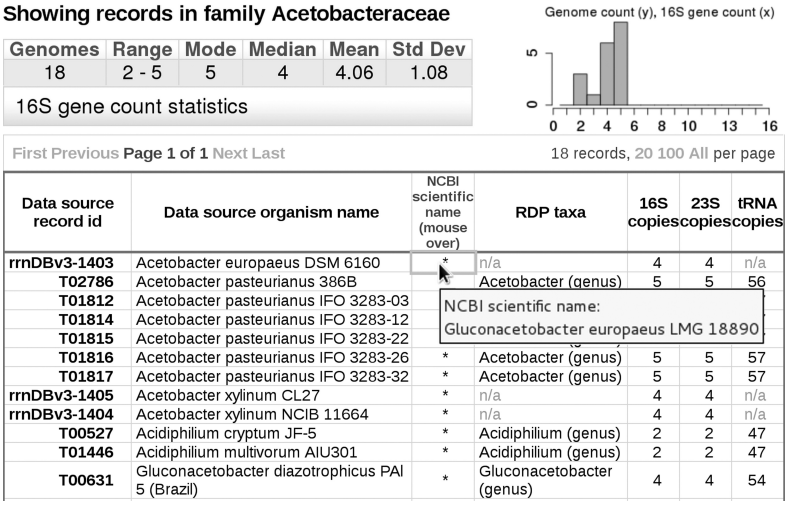
the RDP systems, new statistical summaries for 16S copy number in search results, downloadable copy number data optimized for use in copy number corrections in 16S molecular surveys, a new download area for sharing database contents, improved searching of records enabled by the availability of additional metadata, and additional links to related external resources.

Most records in the database are derived from annotations of published, completed genome sequences and these include estimates of both 16S gene and 23S gene copy number. About 8% of current records are based on data from experimental methods other than genome sequencing and are referred to as organism-based records. Some organism-based records may include data for either 16S or 23S gene copy number, but not both, depending on the experimental methods used. Counts of tRNA genes are present in most genome-based records, but the tRNA data are not quality controlled or curated by the *rrn*DB team. Data about 5S rRNA genes and internally transcribed spacers (ITS) are not present in the *rrn*DB starting with version 4.0.0.

Users can retrieve database entries by two different kinds of text searching, or by browsing a taxonomic hierarchy. 'Search Record Annotations' scans *rrn*DB record fields such as evidence, notes or references for a user-entered search phrase, and also supports retrieval of records by their 16S copy number. 'Search Taxonomy' is a taxonomic name scan, with substring matching, that takes advantage of rich metadata that are available as a result of having integrated the NCBI taxonomy database into *rrn*DB. This mode of searching can retrieve records using obsolete taxonomy names, synonyms, misspellings and others that may be found in the literature, including culture collection strain accessions. Substring searching of RDP taxonomy names is also available. 'Browse Taxonomy' is a way to retrieve records using

pop-up selection lists that can be populated with taxonomic names from either the NCBI or the RDP systems.

Having NCBI and RDP taxonomies both in the system serves the different objectives and starting information that users may have when approaching the website. NCBI taxonomy is ubiquitous in many data sources and is a principal way that different resources are tied together. RDP taxonomy, being more rooted in phylogeny, is used to classify 16S sequences in molecular surveys. Search results are returned on a separate web page in table format (Figure 3.1) , one record per row, where each row is identified by a 'Data source record id' in the first



Showing records in family Acetobacteraceae

Genomes	Range	Mode	Median	Mean	Std Dev
18	2 - 5	5	4	4.06	1.08

16S gene count statistics

First Previous **Page 1 of 1** Next Last — 18 records, 20 100 All per page

Data source record id	Data source organism name	NCBI scientific name (mouse over)	RDP taxa	16S copies	23S copies	tRNA copies
rrnDBv3-1403	Acetobacter europaeus DSM 6160	*	n/a	4	4	n/a
T02786	Acetobacter pasteurianus 386B		Acetobacter (genus)	5	5	56
T01812	Acetobacter pasteurianus IFO 3283-03					
T01814	Acetobacter pasteurianus IFO 3283-12					
T01815	Acetobacter pasteurianus IFO 3283-22					
T01816	Acetobacter pasteurianus IFO 3283-26	*	Acetobacter (genus)	5	5	57
T01817	Acetobacter pasteurianus IFO 3283-32	*	Acetobacter (genus)	5	5	57
rrnDBv3-1405	Acetobacter xylinum CL27	*	n/a	4	4	n/a
rrnDBv3-1404	Acetobacter xylinum NCIB 11664	*	n/a	4	4	n/a
T00527	Acidiphilium cryptum JF-5	*	Acidiphilium (genus)	2	2	47
T01446	Acidiphilium multivorum AIU301	*	Acidiphilium (genus)	2	2	47
T00631	Gluconacetobacter diazotrophicus PAl 5 (Brazil)	*	Gluconacetobacter (genus)	4	4	54

Figure 3.1: Features of *rrn*DB. Screen shot of a 'Browse Taxonomy' search result for the family *Acetobacteraceae* using NCBI taxonomy. Statistics for 16S gene counts of all 18 records are shown in the upper-left table. The distribution of 16S counts among the records is shown in the histogram to the right. Summary data for the individual records are shown in the larger table below. Record ids that are prefixed with '*rrn*DBv3-' were sourced from *rrn*DB v3.1.227. The other record ids are KEGG accessions. Data source organism names have been given higher visibility than NCBI names because they more often include strain designations. Viewing an NCBI name requires a mouse-hover over the table cell as shown for record *rrn*DBv3-1403. RDP taxonomy displayed in this table is limited to genus assignment. Each data source record id is hyperlinked to its corresponding record-detail web page. The records can be reordered by clicking on most column headers.

column. We have adopted the 'T number' accessions of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.,* 2014) as the data source record id for genome-based records, while for organism-based records a permutation of the 'strain id' of the earlier *rrn*DB database is used. Each record is associated with an organism name originating from the data source. Another column is populated with the most recent organism name from the NCBI taxonomy database, and when assigned, the RDP genus is also shown. The organism names from NCBI and the data source are not always identical because data source names tend to lag behind changes in NCBI taxonomy. The copy numbers for 16S, 23S and tRNA genes of each record are in the table, which can be sorted by clicking on the header of most columns.

Each search generates a statistical summary of 16S gene counts for the retrieved records, and these are presented above the main table (Figure 3.1). Statistics include the record count for the result set, the minimum and maximum 16S copy numbers, and the mode, median, mean and standard deviation. A histogram with 15 bars (for copy number of 1–15) showing the relative distribution of 16S copy numbers in the search result, quickly communicates information about the search population. The histogram can be especially illuminating in certain cases, such as the 225 records of the family *Enterobacteriaceae* (Figure 3.2). This family shows a broad range of 1 to 9 for the minimum and maximum *rrn* copy numbers. The histogram reveals the distribution to be bimodal with peaks in the lower and middle regions of the copy number range. Sorting the result table on the '16S copies' column and observation of organism names would reveal the low-copy-number peak to comprise insect-symbiotic organisms exclusively.

A detailed web page report about any record in a search result can be accessed by clicking on the data source record id of the corresponding table row. The detail reports include additional NCBI and RDP taxonomy information, the type of evidence supporting the rRNA gene counts, curator notes and hyperlinks to external KEGG, NCBI BioProject and NCBI taxonomy web pages. The linked-to external pages provide access to gene and genome sequences and annotations for users to wish to dig deeper. For organism-based records, we provide reference citations with links to NCBI PubMed entries.

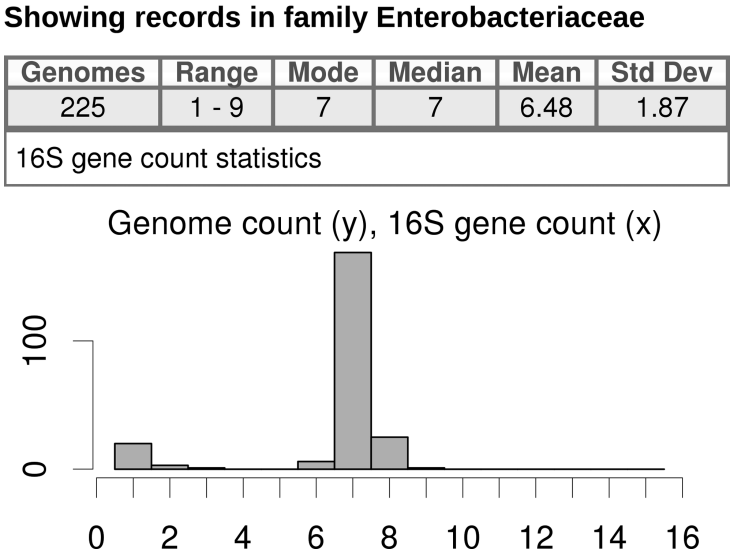


Figure 3.2: *rrn* variation in the *Enterobacteriaceae*. Screen shot showing the statistics and histogram portions of 225 records retrieved by the taxonomy browser for the family *Enterobacteriaceae*. The role of the histogram in clarifying search result statistics is apparent in this example. Although this figure does not show the individual records table like in Figure 3.1, it would be apparent from the organism names that insect-symbiotic bacteria comprise the low-16S cluster.

For the purpose of adjusting organism abundance in molecular surveys, the mean 16S copy number for a taxon can be misleading if calculated from all genomes due to over-representation of some species. One way to correct for this potential source of bias is to calculate the mean of a taxon from the means of its sub-taxa. We have calculated these 'pan-

taxa statistics' for all taxa, from genus to domain level, specifically to support copy number correction. The statistics are available for both RDP and NCBI taxonomies. The RDP development team has extended RDP Classifier to support adjustment of the relative abundance of each taxon. The newer version of the Classifier was trained with the *rrn*DB pan-taxa statistics and is available from RDP (http://rdp.cme.msu.edu/) and the RDP repository on SourceForge (http://sourceforge.net/projects/rdp-classifier/). The 'Estimate' feature of the *rrn*DB website is an on-line interface to the RDP Classifier, including 16S copy number adjustment of taxon abundance for user-uploaded 16S sequence files.

The website includes an 'About *rrn*DB' web page describing the database, a 'Manual' web page describing how to use the various features, and a contact email address for users to ask questions, suggest improvements or alert the curators about problems. A 'Downloads' web page provides access to tab-delimited tables of versioned *rrn*DB data as well as the pan-taxa tables. All of the software resources used in the project are freely available under open-source licenses and have strong community support.

Data sources

Genome-based records in the *rrn*DB are ultimately derived from the NCBI RefSeq collection. The specific data files that we process are acquired from KEGG and carry additional annotation created by KEGG. In particular the *rrn*DB makes use of KEGG '*K* numbers', which apply consistent labeling to orthologous genes across multiple genomes to compute the 16S and 23S rRNA and tRNA gene counts of genomes. The use of *K* numbers to count rRNA gene copy numbers traverses problems caused by inconsistent labeling and annotation errors in sources upstream of KEGG.

KEGG source data are accessed by us through paid academic subscription to the KEGG data via their FTP site (http://www.kegg.jp/kegg/download/). We expect to bring new and updated genomes into the *rrn*DB with increased frequency using the KEGG data source. The amount of KEGG data that are necessary to share in order to run the *rrn*DB website is negligible and well within the terms of the KEGG academic license; therefore, all data made available through the *rrn*DB are presented without restriction for non-commercial use.

The NCBI taxonomy for Bacteria and Archaea is fully integrated into the *rrn*DB so as to support the taxonomy browsing and searching functions of the website. The integrated taxonomy also supports the computation of statistics from the *rrn*DB records aggregated at any node of the NCBI taxonomic tree. The NCBI taxonomy data will be updated together with each update of KEGG genomes. NCBI taxonomy data are freely available at the NCBI FTP site (http://www.be-md.ncbi.nlm.nih.gov/taxonomy/).

Records of the *rrn*DB are also mapped to the taxonomy system used by the Ribosomal Database Project (Cole *et al.,* 2014). Genomes of the *rrn*DB are mapped to RDP taxonomy using the RDP Classifier tool (Wang *et al.,* 2007), where each 16S rRNA gene sequence that is classified at a genus bootstrap score of 0.8 or more contributes to the genome's RDP taxonomy. We have been able to map ~94% of the genome-based records to one or more RDP genera. A genome can map to multiple RDP taxonomies if the genome has multiple 16S genes and a degree of sequence dissimilarity among them. The only genome having been assigned dual RDP taxonomy in *rrn*DB v4.2.2 is that for *Thermoanaerobacterium saccharolyticum* DSM 571 (KEGG T01299), which mapped to the genus *Thermohydrogenium*

as well as to itself. Divergence of 16S sequences within *Thermoanaerobacterium* strains has been described before (Větrovský and Baldrian, 2013).

The *rrn*DB holds 216 organism-based records that use *rrn* copy number estimates from various empirical methods (Lee *et al.,* 2009). Twenty-five of these records have 23S gene counts but not 16S counts, and for the purpose of computing 16S copy number statistics we presume that their 16S and 23 gene copy numbers are equal.

Data curation

Maintaining genome-based resources involves a trade-off between human curation of records, which is laborious but leads to improved data quality, and machine processing of records, which has higher throughput but can compromise data quality. When updating the *rrn*DB, a series of automated quality control (QC) tests are applied to identify genomes that may have problems in annotations that can affect the *rrn*DB. Problematic genomes are held back until the annotations are corrected at their source, or until the genomes can be manually curated. At present our QC pipeline probably retains some genomes that should be allowed through; however, given the increasing number and phylogenetic breadth of published genomes, conservative curation is preferable for most analyses. Our QC pipeline will improve over time as we examine held-back genomes and adjust the QC rules. In addition, our QC strategy does not eliminate human curation, though it does reduce it dramatically compared to earlier versions of the *rrn*DB.

The initial QC tests identify genomes that are missing some annotations, or in some cases all annotations, for 16S or 23S rRNA genes. The tests count the number of genes that are assigned the *K* numbers K01977 (16S rRNA) and K01980 (23S rRNA). It is at this stage

that the 16S and 23S counts that enter the *rrn*DB are also computed for genomes that pass QC. For a genome to pass, it must have at least one 16S gene and one 23S gene that is annotated, and the count of annotated 16S and 23S genes must be equal. To increase confidence in 16S counts at this stage, we perform the tests using two different KEGG data source files that should give identical counts. A genome is held back until the next update if the redundant data sources do not agree.

Nine percent of genomes that have entered our QC pipeline have been held back by the above tests; however, more than half of those passed within four months later, during a subsequent update from a new KEGG release. The condition that 16S and 23S gene counts must be equal is admittedly a blunt tool. Cases of rRNA operons missing the 16S rRNA gene, which would cause unequal 16S and 23S counts, have been demonstrated in some bacteria (Schwartz *et al.,* 1992). Again, our QC pipeline will become more refined as we examine the individual cases of genomes that are held back.

Further annotation-based testing is designed to detect genomes containing duplicate annotations for what is essentially the same 16S gene. As of this writing the duplication test has discovered two genomes where a 16S or 23S gene had been annotated twice, but with a 1- to 8-base offset between the endpoint coordinates of the duplicates.

Sequence-based quality control steps examine the 16S rRNA gene sequences of all genomes for evidence suggesting that any of them may not be a valid 16S gene sequence. This is done by aligning the putative 16S gene test sequences to the SILVA SSU reference set using the SINA Aligner (Pruesse *et al.,* 2012). Any gap in the multiple alignment that is present in every test sequence is removed, then an estimated phylogeny is constructed using FastTree (Price *et al.,* 2010). The midpoint-rooted tree has revealed DNA sequences showing

unexpectedly long, deep branches suggesting potential annotation problems with those sequences. BLAST similarity searches of the suspect sequences against the NCBI number database are then conducted. Sequences showing low-scoring 16S hits, or only hits to non-16S genes, are taken as justification to hold the genome for examination by a curator. Nine genomes have been held back by the sequence-based criteria.

A final QC test looks for genomes having 16S copy number counts that are outside of the usual range displayed by other genomes of its species. Any species group that shows a difference of three or more between the lowest and highest 16S copy number, is manually examined for genomes that are candidates for having annotation errors affecting 16S gene counts. We have used the database to assess 16S copy-number variability in single-species aggregates of records (Figure 3.3) . For 301 species that are represented by at least two records, 77% are invariant within the species for 16S copy number. An additional 16% of species vary by only one 16S copy. Only 3% of species vary by more than two 16S copies and the maximum variability was five (one species represented by two genomes). Seven genomes have been held back by these criteria.

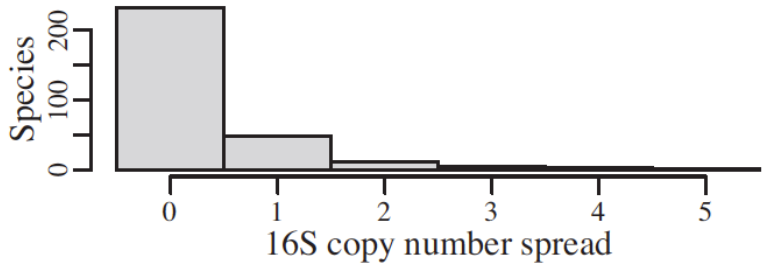


Figure 3.3: *rrn* variation within bacterial species. Histogram showing 16S copy number variability in 301 species aggregates of the *rrn*DB records. Only species that are represented by at least two records are counted in this display. Fully 77% of the species show zero variance in 16S gene copy number count among the comprising records. Sixteen percent of the species vary by only one copy, and only 3% of species show a copy number spread of three or more.

Future development

One goal of research in the Schmidt lab group has been to understand the physiological and evolutionary implications of *rrn* redundancy. That goal has spurred the development of internal resources that have found their way into every major revision of the *rrn*DB since its introduction in 2001. Most recently we have begun to integrate the higher-order functional ontologies of the KEGG database into our research database systems. A goal for development of the *rrn*DB is to extend that access to the integrated functional and copy number data to the broader community. To a large extent, the creation of that capacity was the reason why we chose KEGG as a data source for the new *rrn*DB.

Acknowledgment

The authors thank Jim Cole, Qiang Wang and Benli Chen of the Ribosomal Database Project for helpful discussions and for incorporating *rrn*DB 16S copy number data into RDP Classifier.

Funding

National Institutes of Health [M0099549 to T.M.S.]; National Science Foundation's Long-Term Ecological Research Program [DEB 1027253 to T.M.S]; Department of Energy Office of Science Graduate Fellowship Program [DOE SCGF to B.R.K.R., in part] by the American Recovery and Reinvestment Act of 2009, administered by ORISE-ORAU [DE-AC05–06OR23100]. Funding for open access charge: The National Science Foundation's Long-Term Ecological Research Program [DEB 1027253 to T.M.S.].

REFERENCES

REFERENCES

Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, HugenholtzP, Tyson GW. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**: 1–13.

Cavicchioli R, Ostrowski M, Fegatella F, Goodchild A, Guixa-Boixereu N. (2003). Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of *Sphingopyxis alaskensis* (formerly *Sphingomonas alaskensis*). *Microb Ecol* **45**: 203–217.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**: D633–D642.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dethlefsen L, Schmidt TM. (2007). Performance of the translational apparatus varies with the ecological strategies of bacteria. *J Bacteriol* **189**: 3237–3245.

Eichorst SA, Breznak JA, Schmidt TM. (2007). Isolation and characterization of soil bacteria that define *Terriglobus* gen. nov., in the phylum Acidobacteria. *Appl Environ Microbiol* **73**: 2708–2717.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**: D199–D205.

Kembel SW, Wu M, Eisen JA, Green JL. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* **8**: e1002743.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**: 1328–1333.

Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. (2001). rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* **29**: 181–184.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–821.

Lee ZM, Bussema C III, Schmidt TM. (2009). *rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res* **37**: D489–D493.
Morgan JL, Darling AE, Eisen JA. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* **5**: e10209.

Pinto AJ, Raskin L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* **7**: e43093.

Price MN, Dehal PS, Arkin AP. (2010). FastTree2—approximately maximum—likelihood trees for large alignments. *PLoS One* **5**: e9490.

Pruesse E, Peplies J, Glöckner FO. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JHD. (2005). Structures of the bacterial ribosome at 3.5 A resolution. *Science* **310**: 827–834.

Schwartz JJ, Gazumyan A, Schwartz I. (1992). rRNA gene organization in the Lyme disease spirochete, *Borrelia burgdorferi*. *J Bacteriol* **174**: 3757–3765.

Větrovský T, Baldrian P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* **8**: e57923.

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

Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**: e33865.

CHAPTER 4

A spectrum of bacterial life history strategies are predicted by rRNA operon copy number

Abstract

The potential for rapid reproduction is a hallmark of microbial life, but microbes in nature must also survive and compete when growth is constrained by environmental conditions. A microbe's fitness varies across environments, resulting in a characteristic pattern of survival and reproduction – its life history. Attributes which influence this fitness pattern are life history traits. Despite the value of understanding factors that determine microbial fitness in nature, a systematic framework for classifying and predicting microbial life histories has not been available. Here we show that variation in the number of ribosomal RNA (*rrn*) operons in bacterial genomes reflects a spectrum of life histories, effectively collapsing multiple traits onto a single axis. Using phylogenetically informed analyses, we establish that maximum growth rate doubles when the number of *rrn* operons doubles, and that the carbon use efficiency of heterotrophic bacteria is inversely related to *rrn* copy number. Among 1,167 sequenced species, bacteria with few *rrn* operons have streamlined genomes while bacteria with many *rrn* operons are more likely to encode chemotaxis and this was not due to shared evolutionary history. Although *rrn* copy number is also correlated with PTS transporter abundance (positively) and autotrophy (negatively), these relationships could have arisen due to chance evolutionary events. We also demonstrate that orthologous gene composition among these bacteria covaries with *rrn* copy number, revealing a genome-wide signature of bacterial life histories. Linking *rrn* copy number to bacterial life histories enables ecological predictions not only from sequenced

genomes (Stoddard *et al.*, 2014), but also from surveys of bacterial 16S rRNA-encoding genes via inference of *rrn* copy number (Angly *et al.*, 2014; Kembel *et al.*, 2012). In particular, the correlation between carbon use efficiency and *rrn* copy number predicts that decomposition by oligotroph-dominated communities will leave more carbon in an ecosystem compared to copiotroph-dominated communities in a 'common-garden' experiment. Relationships between *rrn* copy number and life histories also provide a basis for predicting and monitoring changes in microbial community composition in response to perturbations of resource availability.

Introduction

Microbes have a tremendous impact on the biology and geochemistry of our planet (Singh *et al.*, 2010), yet we have a paltry understanding of the environmental factors that sculpt their genomes and shape the composition of complex microbial communities in nature. Recent studies are beginning to provide new ecological and evolutionary insights for enigmatic bacterial species by combining genomics with laboratory cultivation (Giovannoni *et al.*, 2014; Sorokin *et al.*, 2012; H Koch *et al.*, 2014). While new technologies promise to advance our knowledge of the microbial world, conceptual challenges are hindering our understanding of the ecological factors influencing microbial fitness (Prosser, 2015). A framework linking the physiology, ecology, and evolution of diverse microbes is a necessary step on the path to understanding the causes and consequences of microbial fitness variation in nature (Ackermann, 2015).

Biologists have developed life history theory to explain how ecological pressures and evolutionary forces act in concert to produce reproductive variation among species (Stearns, 2000). In this study I apply life history theory to bacteria by synthesizing evidence that nutrient availability is a key ecological variable leading to adaptations in growth physiology with large

fitness consequences. A bacterium's life history is its pattern of fitness variation across the environments in which it can survive, *i.e.*, its niche. One of the primary environmental factors that influences the reproductive success of bacteria is resource availability: natural environments are spatially structured at the scale of individual microbes (Stocker, 2012) and in this heterogeneous milieu abundant resources select for rapidly growing bacteria while low resource availability selects for efficient resource utilization (Pfeiffer *et al.*, 2001). Rapid and efficient growth are complex adaptations that integrate multiple underlying metabolic pathways, cellular characteristics, and physiological processes. Therefore, rapid and efficient growth can be considered life history tactics: sets of coadapted traits shaped by natural selection to cope with the environmental pressure of resource availability.

Microbiologists have long recognized that resource concentration is a key variable when cultivating environmental microbes because it can select for distinct types of organisms (Kuznetsov *et al.*, 1979; Schut *et al.*, 1993). Bacteria which are favored during resource abundance are classified as copiotrophs. They are contrasted by oligotrophs, which have a higher relative fitness during chronic resource scarcity. One axis of bacterial life history variation is the spectrum between copiotrophy and oligotrophy, which describes how fitness changes along a resource gradient. I propose that one can predict any given bacterium's place on this life history spectrum using the number of ribosomal RNA operons (*rrn*) present in its genome.

Multiple lines of evidence suggest bacteria possessing many *rrn* copies utilize a rapid growth life history tactic. Bacteria with many *rrn* copies grow more quickly upon increased resource availability (Klappenbach *et al.*, 2000), synthesize protein at a faster rate (Dethlefsen & Schmidt, 2007) and have faster maximal population growth rates (Vieira-Silva & Rocha, 2010) than organisms with low *rrn* copy number. Efficient growth – the number of progeny produced

per unit resource consumed – is a trait that is commonly attributed to oligotrophic bacteria (Fierer *et al.*, 2007; Giovannoni *et al.*, 2014), but to our knowledge no study has tested this assertion among a diverse collection of oligotrophic strains. Two environmental conditions which select for efficient bacterial growth—spatial structure and low resource concentrations—are synonymous with oligotrophic environments (Roller & Schmidt, 2015), supporting the idea that efficient growth is adaptive for oligotrophic bacteria. In addition, many oligotrophic organisms are known to have a low *rrn* copy number (Fegatella *et al.*, 1998; Eichorst *et al.*, 2007; Grote *et al.*, 2012).

In this study I explore if efficient growth is related to the *rrn* copy number of a diverse collection of bacteria. The carbon use efficiency (CUE) of heterotrophic bacteria was used as a metric of the efficient life history tactic, and CUE is known to vary based on resource availability and resource quality (Roller & Schmidt, 2015). Therefore, measurements of CUE were performed in a common-garden experimental design to control these potentially confounding variables. Additionally, I examined the scale of the relationship between *rrn* copy number and maximal growth rate, because a previous report of this relationship included only a non-parametric correlation coefficient (Vieira-Silva & Rocha, 2010).

Building upon the results relating *rrn* copy number to rapid and efficient life history tactics, I go on to explore if *rrn* copy number is a quantitative proxy of the life history spectrum from copiotrophy to oligotrophy. To do so, I examined if postulated life history traits related to nutrition are accurately predicted by *rrn* operon copy number among the genomes of 1,167 unique bacterial species. These genome inferred traits include chemotactic motility, genome streamlining, thiamine biosynthesis, autotrophy, and phosphoenopyruvate:carbohydrate phosphotransferase system (PTS) transporter richness. Moving beyond any particular life history

trait, I examined if *rrn* copy number has a pervasive influence on the orthologous gene content of bacterial genomes.

This study utilizes a comparative biology approach to quantify how life history traits have co-evolved with *rrn* copy number among thousands of bacterial species. A well-known issue in quantifying relationships among traits using comparative data is that species are not independent data points, as is often assumed in statistical methods. Instead, species have shared ancestry and thus we expect some correlated evolution of traits to occur by chance rather than for adaptive reasons (Felsenstein, 1985; Blomberg & Garland, 2002; Blomberg *et al.*, 2003; Revell, 2009). Therefore, I implement phylogenetically informed statistical analyses to control for the effect of shared evolutionary history on all comparisons among species throughout this study. Applying life history theory to bacteria can provide new insight into the evolutionary forces shaping bacterial genome content. Linking bacterial life history to *rrn* copy number has many practical benefits, especially the ability to generate ecological predictions at a variety of scales from individual strains with complete genomes (Stoddard *et al.*, 2014) to complex communities via inference from molecular survey data (Angly *et al.*, 2014; Kembel *et al.*, 2012).

Materials and Methods

Bacterial strains, media and growth conditions for efficiency experiments

Detailed descriptions of the bacterial strains used are provided in Table 4.1. The medium used for all growth experiments was vitamin and salts base (VSB) supplemented with 10mM sodium succinate as the sole carbon and energy source. VSB medium contains the following per liter: 2.5 g KCl, 0.1 g KH_2PO_4, 0.125 g NH_4Cl, 0.075 g $CaCl_2 \bullet 2H_2O$, 0.31 g $MgCl_2 \bullet 6H_2O$, 0.5 g NaCl, 10mM morpholinepropanesulfonic acid (MOPS), 1.6ml of 1.25M Na_2SO_4 stock solution,

1ml SL-10 trace element solution, 1ml 1000x Vitamin mix, 50 μg cyanocobalamin (vitamin B_{12}), and 100μg Thiamine HCl (vitamin B_1). SL-10 trace elements contains per liter: 10ml of 7.7N HCl, 1.5 g $FeCl_2 \bullet 4H_2O$, 0.19 g $CoCl_2 \bullet 6H_2O$, 0.1 g $MnCl_2 \bullet 4H_2O$, 0.07g $ZnCl_2$, 0.036 g $Na_2MoO_4 \bullet 2H_2O$, 0.024 g $NiCl_2 \bullet 6H_2O$, 0.006 g H_3BO_3, 0.002 g $CuCl_2 \bullet 2H_2O$. 1000x Vitamin mix contains per 100ml: 10mM sodium phosphate buffer, pH 7.1, 4 mg 4-aminobenzoic acid, 1 mg D(+) biotin, 10 mg nicotinic acid, 5 mg D-pantothenic acid hemicalcium salt, 15 mg pyridoxine hydrochloride. *Vibrio natriegens* ATCC 14048 cultures were supplemented with NaCl solution resulting in a 0.02g/ml final concentration.

Strain	Taxonomy (phylum; class; family)	*rrn* copy number	Strain & isolation information
Vibrio natriegens ATCC 14048	Proteobacteria; γ-proteobacteria; *Vibrionaceae*[a]	13	(Eagon, 1962)
Bacillus subtilis Marburg ATCC 6051	Firmicutes; Bacilli; *Bacillaceae*[a]	10	(Conn, 1930)
Escherichia coli K12 MG1655	Proteobacteria; γ-proteobacteria; *Enterobacteriaceae*[a]	7	(Datta *et al.*, 2006)
HF3	Proteobacteria; γ-proteobacteria; *Pseudomonadaceae*[b]	4	(Dethlefsen & Schmidt, 2007; Gorlach *et al.*, 1994)
EC5	Actinobacteria; Actinobacteria; *Micrococcaceae*[b]	4	(Dethlefsen & Schmidt, 2007; Klappenbach *et al.*, 2000)
PX3.14	Proteobacteria; α-proteobacteria; *Rhodospirillaceae*[b]	2	(Dethlefsen & Schmidt, 2007)
Sphingopyxis alaskensis RB2256	Proteobacteria; α-proteobacteria; *Sphingomonadaceae*[b]	1	(Dethlefsen & Schmidt, 2007; Schut *et al.*, 1993; 1997)
TAA166	Acidobacteria; Acidobaceriia; *Acidobacteriaceae*[c]	1	(Stevenson *et al.*, 2004; Eichorst *et al.*, 2007)

Table 4.1: Bacterial strains used in this study. Taxonomy determined by the following sources: [a]NCBI taxonomy, [b](Dethlefsen & Schmidt, 2007), [c](Eichorst *et al.*, 2007).

All growth experiments utilized VSB succinate medium in a 25°C incubator shaking at either 100 rpm (TAA166) or 200 rpm (all other strains). Optical density was measured over time on the Spec20D+ at wavelengths of either 600nm (TAA166) or 420nm (all other strains), which

maximizes measurement sensitivity at high and low cell densities, respectively. Cells were recovered from freezer stock in batch culture and allowed to reach unconstrained and balanced growth by multiple transfers to fresh medium during exponential growth. Unconstrained and balanced growth was empirically determined for each strain when the growth rate no longer improved upon transfer to fresh medium, typically after a dilution of at least 1,000 fold from exponentially growing freezer stock recovery culture.

Maximum recorded growth rate determination

The maximum recorded growth rates for a diverse collection of 176 bacteria with known *rrn* copy number (Vieira-Silva & Rocha, 2010), and for the 8 strains measured for efficiency in this study, were gathered from the literature (Eagon, 1962; Vieira-Silva & Rocha, 2010; Dethlefsen & Schmidt, 2007). The growth of strain TAA166 in this study exceeded the maximum recorded growth rate from the literature (Eichorst *et al.*, 2007), so the value from this study was utilized.

Protein yield and carbon use efficiency measurements

All protein yield and growth efficiency measurements were obtained in cultures with an optical density at least two doublings prior to departing from unconstrained and balanced growth. These optical density values were determined based on prior experiments.

Protein yield was measured using ^{3}H-leucine incorporation along with oxygen (O_2) consumption. Leucine was used because its is one of the least variable amino acids in protein on a mol% basis (Simon & Azam, 1989). The amount of radiolabeled leucine to add was optimized on the most rapidly growing strains and 250nCi ^{3}H-leucine (S.A. 0.5Ci/mol) was added to 30ml

growing cultures of all strains. ^{3}H-leucine incorporation was measured at multiple time points to ensure leucine was not depleted over the course of the experimental measurement. ^{3}H-leucine incorporation was converted into protein production and carbon (C) units with widely used conversion factors— mol% leucine in protein (7.3), intracellular isotope dilution (1.71), protein:C dry weight ratio (0.86)—from a marine microbial community (Simon & Azam, 1989). O_2 consumption (nmol O_2 min^{-1}) was measured with the Unisense microrespiration system. Small volumes from cultures in unconstrained, balanced growth conditions were subcultured for at least 3 separate measurements over short time intervals (<10 minutes). Mean specific O_2 consumption (nmol O_2 $cell^{-1}$ min^{-1}) was calculated by normalizing the O_2 consumption rate by the biomass present in the culture, averaging over at least 3 separate measurements, and multiplying this rate by the integral of the growth equation for the culture during ^{3}H-leucine incorporation. O_2 consumption was converted into carbon dioxide (CO_2) production by assuming a respiratory quotient of 8/7, which represents complete oxidation of the carbon source and that all measurable oxygen consumption can be attributed to respiration. The inferred CO_2 respiration and biomass C production – from O_2 consumption and ^{3}H-leucine measurements – was used as one indirect estimate of carbon use efficiency.

A direct measurement of carbon use efficiency was performed by tracking the fate of ^{14}C-succinate into biomass and carbon dioxide (CO_2). ^{14}C-succinate was added to 30ml cultures growing in 500ml flasks with excess non-labeled succinate during unconstrained, balanced growth. Cultures were sealed after radiochemical addition to trap $^{14}CO_2$ in the culture flask. Sealed cultures were incubated for 1 generation or less after ^{14}C addition and terminated by addition of trichloroacetic acid (5%), which also released dissolved CO_2 into gas phase. Culture headspace was flushed with N_2 gas for 2 hours into a series of 3 gas traps (1:1,

phenethylamine:methanol), which trap gaseous CO_2 in their liquid phase. Gas trap contents and were transferred in triplicate to scintillation cocktail (Biosafe-II) and radioactivity present in the samples was quantified using a scintillation counter (Beckman Coulter).

All radiolabeled biomass from protein yield and carbon use efficiency experiments was precipitated using trichloroacetic acid (TCA, 5% final volume), centrifuged at 11,000g for 10 minutes and washed with ethanol (80%), resolubilized using NaOH (1M), suspended in scintillation cocktail (Biosafe-II) and the radiolabel was quantified in a scintillation counter (Beckman Coulter).

Translational power analysis

Previously published results quantifying translational power for ten bacterial strains (Dethlefsen & Schmidt, 2007) was re-analyzed in this study. Translational power was calculated by normalizing volumetric protein content (fg/fl) by RNA content (fg/fl) and specific growth rate (hr^{-1}).

Comparative genomic and phylogenetic analyses

The May 2014 version of the Kyoto encyclopedia of genes and genomes (KEGG) database was downloaded and used to construct a curated dataset of sequenced bacterial genomes. Symbiotic, commensal, and parasitic bacteria with degraded genomes were excluded if these genomes were associated with signatures of genetic drift (Giovannoni *et al.*, 2014), *e.g.*, high pseudogene counts, elevated rates of non-synonymous substitutions, or expansion of noncoding genetic elements. Bacteria were also excluded if their *rrn* copy number could not be accurately estimated from KEGG data (Stoddard *et al.*, 2014). All genomes were then subjected

to manual curation where a single representative genome was chosen for each unique bacterial species. Representative genomes were selected using the following hierarchical criteria: 1) genome of Type strain of species available, 2) the central tendency of *rrn* operon copy number distribution for the species is accurately reflected by the genome, and 3) greatest number of annotated orthologous genes in KEGG orthology system are present in the genome. This resulted in 1,167 genomes that passed all described criteria. The presence of every asserted ortholog (K0) and module (M0) for each of 1,167 genomes in the curated dataset was extracted from the KEGG database, as well as the estimated *rrn* copy number and genome size (Stoddard *et al.*, 2014; Kanehisa *et al.*, 2013).

A bacterial genome was scored as possessing chemotactic motility if the ortholog for the genes cheA, cheY, fliM, and fliN were all present. These genes encode orthologs of the following proteins: chemotaxis histidine kinase (CheA), chemotaxis response regulator which binds the flagellar motor (CheY), and flagellar motor switch proteins that are bound by CheY (FliM & FliN). Chemotactic systems are diverse, even in model bacteria, and this definition was used to ensure that the genomes of four known and diverse chemotactic organisms, *Rhodobacter sphaeroides*, *Escherichia coli*, *Bacillus subtilis*, and *Rhizobium leguminosarum* bv. *viciae*, were scored as possessing chemotaxis (Miller *et al.*, 2009; Porter *et al.*, 2011).

A bacterial genome was scored as being autotrophic by a combination of manual curation and genome annotation. The first step used KEGG annotation to identify genomes that possessed at least one complete KEGG module for one of four autotrophic pathways: Calvin cycle (M00165), the reductive TCA cycle (M00173), 3-Hydroxypropionate cycle (M00376), or Wood-Ljungdahl pathway (M00377). Organisms possessing the Wood-Ljungdahl pathway were then manually curated and genomes that were not explicitly described in the literature as capable of

fixing carbon dioxide as their sole source of biosynthetic carbon were removed. Additionally, organisms possessing a sub-module of the Calvin cycle (M00166 or M00167) or that were from genera that were likely autotrophs were manually curated and those that were explicitly described in the literature as capable of fixing carbon dioxide as their sole source of biosynthetic carbon were added to the list of autotrophic genomes.

The PTS transporter analysis utilized the presence of KEGG modules which were described as PTS transporters. A total of 25 distinct PTS transporter types are described as separate modules in KEGG. The presence of each PTS transporter type was considered in calculations of PTS transporter abundance, so the maximum possible richness of PTS transporters encoded in any genome was 25.

I analyzed the *de novo* thiamine biosynthesis pathway to determine the number of genes present for all 1,167 genomes. A total of 12 biosynthetic steps are present in the canonical bacterial synthesis pathway (Jurgenson *et al.*, 2009), and 11 orthologous genes corresponding to steps in the pathway are annotated and present in the 1,167 genome dataset. Therefore, the maximum possible number of *de novo* thiamine synthesis genes encoded in a genome was considered to be 11. These 11 orthologs were split into two categories, those involved in thiamine recycling (3 orthologs) or uninvolved in thiamine recycling (8 orthologs) based on where in the pathway the gene product functioned relative to salvage of thiamine's metabolic precursors. Gene products catalyzing biosynthesis steps which take place after salvage of a metabolic precursor were considered involved in recycling, while gene products catalyzing biosynthesis steps which occur prior to salvage of a metabolic precursor were considered uninvolved in recycling (Jurgenson *et al.*, 2009).

Aligned 16s rRNA gene sequences for all bacteria in this study were downloaded from Silva (http://www.arb-silva.de). Three different sets of phylogenetic trees were built for this study. One tree set included the eight strains from growth efficiency experiment and 176 additional strains which were all included in the maximum growth rate analysis. A second tree set was built for the 1,167 strains in the comparative genomics analysis. A third tree set was built for the ten strains in a re-analysis of a study on translational power (Dethlefsen & Schmidt, 2007). For the 1,167 bacterial genomes in the comparative genomic analysis, if an aligned sequence from the genome was not available from Silva, an aligned sequence from a separate sequencing effort on the same strain or from the type strain of that species was utilized. Phylogenetic trees were built using the Arb software package using maximum likelihood estimation (RAxML 7.0.4) to generate the ten most likely trees using the GTRMIX substitution model with 25 rate categories and the new rapid hill climbing algorithm. I utilized The Living Tree Project's filter to ensure only those base positions which are conserved in 50% of all bacterial species were used in building the tree (Munoz *et al.*, 2011). All trees were built with 5 archael sequences, which were used to root the tree prior to pruning the archaeal tips. For each of the ten most likely trees built for the three datasets, a single tree with the most negative maximum likelihood was chosen for phylogenetic comparative method analyses.

Statistical analyses

The R statistical programming language was used for all statistical analyses (R Core Team, 2014). Base R packages were used for linear regression analyses. The R package lmodel2 was used for Model II major axis regression and standard major axis regression. Major axis regression accounts for residual error for both dependent and independent variables, and is also

not influenced by the arbitrary choice of dependent variables, so it was used in the ^{14}CUE vs. ^{3}H-CUE analysis. Major axis regression should not be used when variables are measured on different scales, so ranged major axis regression was also used to compare with the standard ordinary least squares (OLS) regression in the CUE versus growth rate & ^{14}CUE versus protein yield analyses. Corrected Akaike information criterion (AICc) was used to determine whether the explanatory variable *rrn* or $\log_2$-transformed *rrn* was a better model for the rate and efficiency data. The R package ggplot2 was used for plotting all figures (Wickham, 2009).

The R package MCMCglmm (Hadfield, 2010) was used for phylogenetic Poisson regression of thiamine biosynthesis orthologs and PTS transporter abundance. The R package phylolm was used for phylogenetic logistic and linear regression (Tung Ho & Ane, 2014) and evolutionary models were chosen empirically by choosing the model with the lowest AICc value. In MCMCglmm analysis, an inverse Wishart distribution was used for prior distributions of phylogenetic regressions (variance limit V=1 and belief parameter nu = 0.002). Guidelines for the package (MCMCglmm package course notes; http:// cran.r-project.org/web/packages/MCMCglmm) recommended these parameters for a flat prior, which makes no *a priori* assumptions of the variance, and it has been used in similar comparative bacterial analyses (Kümmerli *et al.*, 2014). AICc model averaging was used to determine the relative probability of regression models in the maximal growth rate and efficiency analyses which differed only in the transformation of the predictor variable *rrn* copy number.

The R package phytools was used for phylogenetic ranged major axis regression analysis of maximum growth rate vs. ^{14}CUE (Revell, 2012). Phytools was also used for phylogenetic principal coordinates analysis (pPCA). Covariance-based pPCA were performed utilizing Brownian motion as the underlying model of expected trait evolution. pPCA combines variables

into new axes which maximally summarize variation in low-dimensional space while accounting for non-independence of data points due to shared ancestry. The coordinates for each genome on the new pPCA axes were further analyzed using phylogenetic regression. The results of pPCA analyses were correlated with the explanatory variable *rrn* copy number, so orthologs that are part of the *rrn* operon were excluded from genome content pPCA. This included all orthologs for the genes encoding 16S rRNA, 23S rRNA, 5S rRNA, and all tRNAs. The R package phylolm was used for phylogenetic linear regression of the following dependent variables against $\log_2$-*rrn*: maximum growth rate, CUE, genome size, protein yield, translational power, and pPCA genome content scores. A random root Ornstein-Uhlenbeck evolutionary model used in the pPCA and genome size analyses, while Pagel's lambda was used in the growth rate, CUE, protein yield, and translational power analyses (Tung Ho & Ane, 2014). The R package ape was used to import NEXUS formatted tree files, store tree objects, and prune tips from the trees (Paradis *et al.*, 2004).

Results

Maximum growth rate is positively correlated with *rrn* copy number

Maximum reported bacterial growth rates are one metric of the rapid growth and are known to positively correlate with *rrn* copy number (Vieira-Silva & Rocha, 2010). I extend this observation by using a phylogenetic regression of growth rates which indicates that each doubling of *rrn* copy number leads to an approximate doubling of a bacterium's maximum recorded growth rate (Figure 4.1a and Table 4.2). This conclusion is based on the observation that $\log_2$-transformed *rrn* copy number (hereafter referred to as $\log_2$-*rrn*) better explains maximum growth rate variation than untransformed *rrn* counts using the model selection criteria

AICc (Table 4.3). The eight bacteria used in the efficiency experiments mirror the trend in the larger dataset, with a doubling of *rrn* copy number associated with an approximate doubling of maximum growth rate (Table 4.3). The correlation between maximum growth rate and $\log_2$-*rrn*

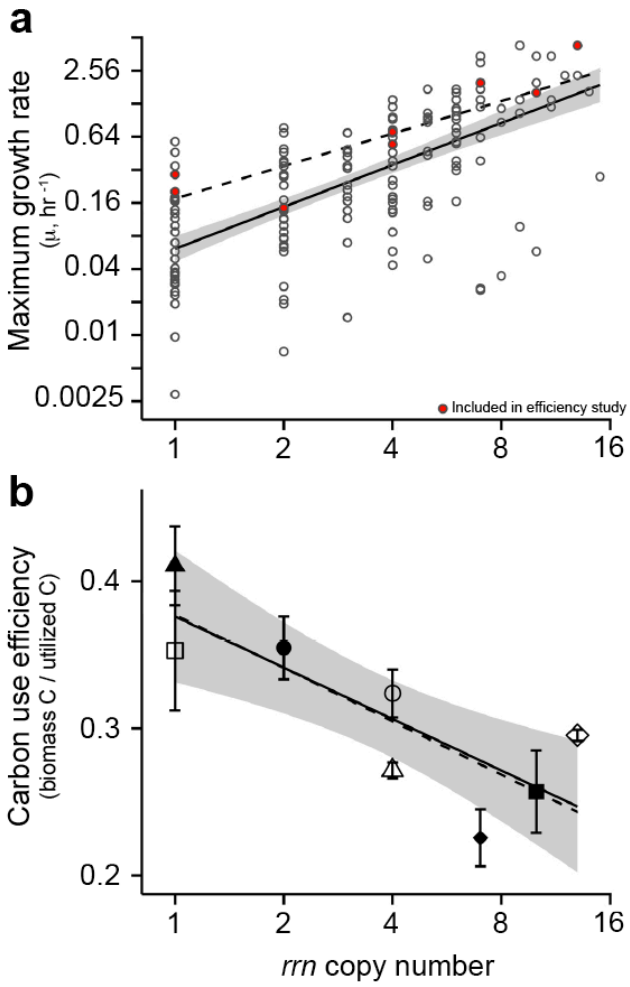


Figure 4.1: Metrics of rapid and efficient growth tactics. Maximum recorded growth rate (a, n=184) and carbon use efficiency (b, n=8) of bacteria. Non-phylogenetic ordinary least squares regression (solid lines) with 95% confidence bands (gray shading) and phylogenetic regression (dashed lines) reveal that these traits are inversely correlated with *rrn* copy number ($\log_2$-transformed). Error bars in panel b represent technical error from triplicate measurements. Strains represented in panel b are: *Sphingopyxis alaskensis* RB2256 (▲), *Acidobacteriaceae* sp. TAA166 (□), *Rhodospirillaceae* sp. PX3.14 (●), *Pseudomonadaceae* sp. HF3 (ʘ), *Micrococcaceae* sp. EC5 (△), *Escherichia coli* K12 MG1655 (◆), *Bacillus subtilis* Marburg ATCC 6051 (■), *Vibrio natriegens* ATCC 14048 (♢).

could not be explained by shared evolutionary history (Table 4.3) indicating it is not due to historical contingency during trait evolution. This relationship is consistent with the suggestion that transcription of the *rrn* operon limits growth rate (Aiyar *et al.*, 2002; Stevenson & Schmidt, 2004) and that a doubling in the capacity for ribosome synthesis is required for a doubling of growth rate.

Growth efficiency is negatively correlated with *rrn* copy number

Efficient growth is a postulated adaptation of oligotrophic bacteria (Giovannoni *et al.*, 2014), which often possess few *rrn* operon copies (Eichorst *et al.*, 2011; Lauro *et al.*, 2009), so I hypothesized that growth efficiency would be negatively correlated with *rrn* copy number. I measured the carbon use efficiency for eight aerobic, heterotrophic bacteria using two independent methods. CUE is a measure of the fraction of carbon consumed that is incorporated into biomass and it is equivalent to growth efficiency, progeny per resource, during unconstrained balanced growth (Roller & Schmidt, 2015). The two methods for measuring CUE were highly correlated ($R^2 = 0.73$) with a slope statistically indistinguishable from 1 (slope=1.44, 95% conf. int. = 0.80 – 3.07) and an intercept near, but statistically different from, 0 (intercept=-0.20, 95% conf. int. = -0.71 – -0.002). This indicates that to the best of our knowledge both estimates give equivalent measures of carbon use efficiency. The highest growth efficiencies in this study came from a model ocean oligotroph, *Sphingopyxis alaskensis* RB2256 (Lauro *et al.*, 2009), and soil strains PX3.14 and TAA166 which were isolated using strategies which favor oligotrophs, such as nutrient-poor media and long incubation times (Stevenson *et al.*, 2004; Klappenbach *et al.*, 2000).

The lowest growth efficiencies were observed in the notorious copiotrophic bacteria *Escherichia coli* and *Bacillus subtilis* (Klappenbach *et al.*, 2000; Dethlefsen & Schmidt, 2007; AL Koch, 2001). I found that each doubling of *rrn* copy number is associated with a 3.6% decrease in the amount carbon that is incorporated into biomass (Figure 4.1b, Table 4.2 and Table 4.3). Log_2-*rrn* best predicted CUE variation and this relationship could not be explained solely by shared ancestry among species (Table 4.2 and Table 4.3). To my knowledge, this is the first phylogenetically robust evidence that links efficient growth to *rrn* copy number.

What physiological mechanisms might underlie increased growth efficiency? Protein synthesis is an attractive target for explaining differences in growth efficiency because protein is an abundant macromolecule (Simon & Azam, 1989), it is expensive to make – 50-60% of ATP is used for polymerizing amino acids into protein during balanced growth (Stouthamer, 1973) – and a high ribosome content is necessary for rapid growth (Schaechter *et al.*, 1958; Stevenson & Schmidt, 2004; Fegatella *et al.*, 1998). One of my CUE methods measured ^{3}H-leucine incorporation into protein and O_2 consumption, and in interpreting this measurement in terms of protein yield—protein produced per oxygen consumed – I find that it is negatively correlated with log_2-*rrn* (Table 4.2, Table 4.3, Figure 4.2). The negative relationship between protein yield and log_2-*rrn* is not due to shared evolutionary history as the regression parameters did not differ when considering the effect of phylogeny in the data set (Table 4.3). I also examined a previously published study of translational performance (Dethlefsen & Schmidt, 2007) and found that mass-normalized translation rate is positively correlated with log_2-*rrn* (Table 4.3 and Figure 4.3).

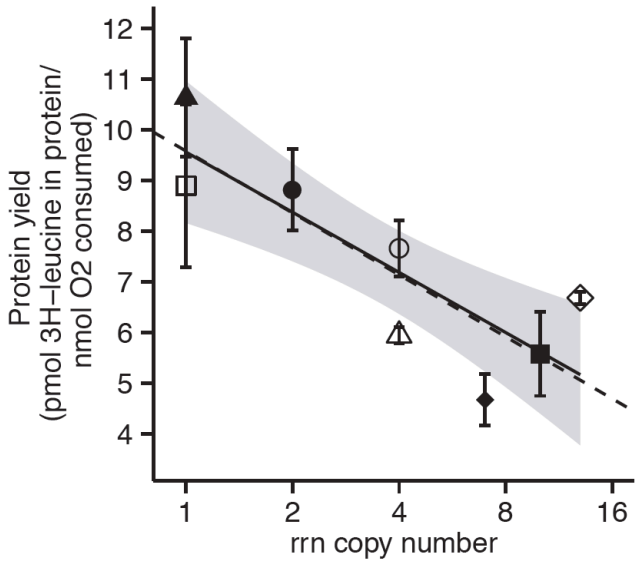


Figure 4.2: Protein yield correlates with $\log_2$-*rrn* copy number. OLS regression best fit (solid line) with 95% confidence band (gray shading) and phylogenetic regression (dashed line) demonstrate a negative relationship. Standard error bars represent technical error from triplicate measurements. Strains represented are: *Sphingopyxis alaskensis* RB2256 (▲), *Acidobacteriaceae sp.* TAA166 (□), *Rhodospirillaceae* sp. PX3.14 (●), *Pseudomonadaceae* sp. HF3 (⭘), *Micrococcaceae* sp. EC5 (△), *Escherichia coli* K12 MG1655 (◆), *Bacillus subtilis* Marburg ATCC 6051 (■), *Vibrio natriegens* ATCC 14048 (♢).

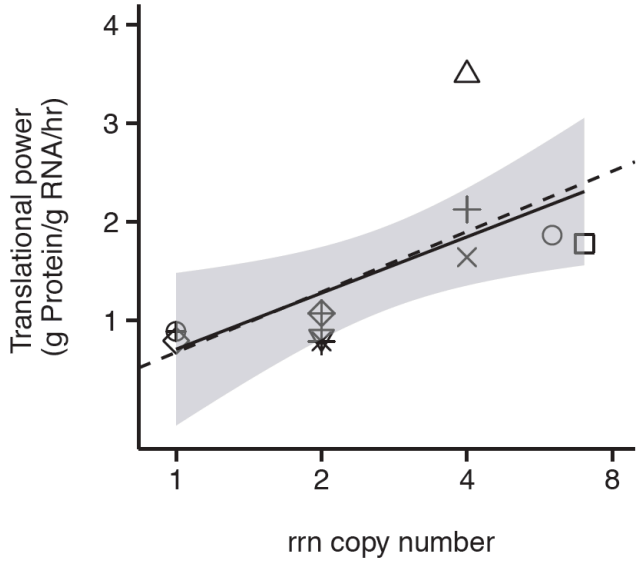


Figure 4.3: Translational power correlates with $\log_2$-*rrn* copy number. OLS regression best fit (solid line) with 95% confidence band (gray shading) and phylogenetic regression (dashed line) demonstrate a positive relationship. Distinct symbols in plot represent strains from ten different bacterial species measured in a separate study (Dethlefsen & Schmidt, 2007).

Trait influence	Trait	Effect size estimate	Significance
Nutrient metabolism	μ_{MAX}	μ_{MAX} doubles with *rrn* doubling	**p < 0.001**$^{\#}$
	Efficiency		
	^{3}H-CUE	-3.6% CUE with *rrn* doubling	**p < 0.004**$^{\#}$
	^{14}C-CUE	-3.2% CUE with *rrn* doubling	**p = 0.057**$^{\#}$
	Translational power	0.61 units with *rrn* doubling	**p < 0.014**$^{\#}$
	Translational yield	-1.14 units with *rrn* doubling	**p = 0.016**$^{\#}$
	Genome streamlining		
	Genome size	+ 0.66 Mbp from 1-15 *rrn*	**p < 0.001**$^{\#}$
	Thiamine biosynthesis	+ 3 biosynthetic steps from 1-15 *rrn*	**pMCMC<0.001^**
	Autotrophy	-5.1% probability from 1-15 *rrn*	p = 0.131★
Nutrient uptake	PTS transporters	< +1 PTS transporter from 1-15 *rrn*	**pMCMC<0.015^**
Nutrient sensing	Chemotactic motility	+11% probability from 1-15 *rrn*	**p < 0.035★**

Table 4.2: Summary of trait relationships with log$_2$-*rrn*. Phylogenetic regression summary statistics for postulated life history traits as a function of log$_2$-*rrn* copy number. Phylogenetic linear (#), logistic (★), and Poisson (^) regression methods depending on the nature of the response variable. Effect size calculated in terms of *rrn* doubling for linear regression models, and in terms of trait change over the extant *rrn* spectrum (1-15 copies) for non-linear regression models. Units for translational power are grams protein synthesized per gram RNA per hour (gProtein gRNA^{-1} hr^{-1}) and measurements are derived from the literature (Dethlefsen & Schmidt, 2007). Units for translation yield are pmol ^{3}H-leucine incorporated in protein per nmol O_2 consumed (pmolLeu nmolO_2^{-1})

Data set	Response variable	Predictor variable	Non-phylogenetic regression: Slope (1-tailed p-value)	Adj. R^2	Non-phylogenetic regression: ΔAICc* (probability log_2-*rrn* is better model)	Phylogenetic regression: Slope (1-tailed p-value)	Phylogenetic regression: ΔAICc* (probability log_2-*rrn* is better model)
176 bacteria (Vieira-Silva & Rocha, 2010; this study)	Log_2 Maximum growth rate	Log_2-*rrn*	1.27 (**<0.001**)	0.46	23.59 (**>0.999**)	0.98 (**<0.001**)	19.9 (**>0.999**)
8 bacteria (this study)	Log_2 Maximum growth rate	Log_2-*rrn*	1.11 (**<0.001**)	0.81	-2.24 (0.25)	1.03 (**<0.001**)	-3.08 (0.18)
8 bacteria (this study)	^{3}H-CUE	Log_2-*rrn*	-0.035 (**0.007**)	0.61	4.22 (0.89)	-0.036 (**0.004**)	4.57 (0.91)
8 bacteria (this study)	^{14}C-CUE	Log_2-*rrn*	-0.034 (0.062)	0.35	1.62 (0.69)	-0.032 (0.057)	1.62 (0.69)
8 bacteria (this study)	^{3}H-Protein Yield	Log_2-*rrn*	-1.19 (**0.005**)	0.71	4.76 (0.92)	-1.22 (**0.003**)	5.08 (0.93)
10 bacteria (Dethlefsen & Schmidt, 2007)	Translational power	Log_2-*rrn*	0.57 (**0.019**)	0.44	1.45 (0.67)	0.61 (**0.014**)	1.41 (0.67)

Table 4.3: Expanded rate and efficiency summary statistics. Regression statistics, with and without phylogenetic correction, for maximum growth rate, carbon use efficiency, translational yield, and translational power as a function of log_2-transformed *rrn* operon copy number. *AIC_C is expressed in terms of AIC_C *rrn* model - AIC_C log_2-*rrn* model and not the conventional AIC_C worst model - AIC_C best model. The probability of log_2-*rrn* being a better model than *rrn* comes from AIC_C model weighting using the AIC_C values of the two regression models. Phylogenetic regression models used Pagel's lambda to account for the effect of shared evolutionary history among traits.

Does *rrn* copy number correlate with postulated life history traits?

The results for the life history tactics of rapid and efficient growth suggest log_2-*rrn* may be a proxy for a bacterium's place along the life history spectrum. To investigate further, I assessed if postulated life history traits inferred from the genomes of 1,167 unique bacterial species are correlated with log_2-*rrn* . Chemotactic motility and PTS transporters are traits hypothesized to be adaptive in copiotrophs to find and exploit high nutrient conditions, but maladaptive in oligotrophs, due to their energetic costs (Stocker, 2012; Taylor & Stocker, 2012; Lauro *et al.*, 2009). I propose that the presence of chemotaxis and the richness of PTS transporter orthologs will be positively correlated with *rrn* copy number.

The probability that a bacterium encodes chemotactic motility increased as a function of *rrn* copy number (Table 4.2, Table 4.4, and Figure 4.4). On average, a bacterium with 15 *rrn* copies is 11% more likely (95% confidence interval = 0.5%-30%) to encode chemotactic motility than a 1 *rrn* bacterium when including phylogeny in the regression model. The relationship between $\log_2$-*rrn* and chemotaxis was stronger when phylogeny was not included in the regression model, but evolutionary history could not account for the relationship between these two variables.

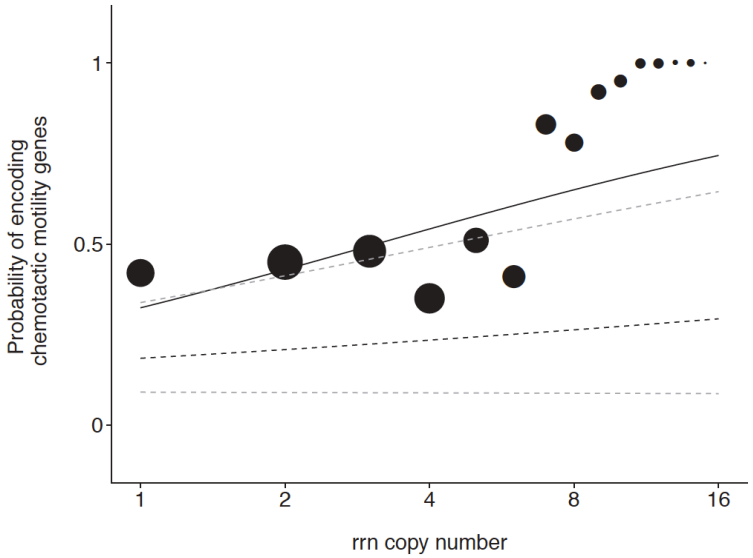


Figure 4.4: Chemotactic motility correlates with $\log_2$-*rrn* copy number. The proportion of genomes possessing chemotactic motility for each *rrn* copy number is represented by circles which are sized proportionally to the number of genomes (for *rrn* 1-15, respectively, N = 150, 262, 195, 165, 104, 96, 76, 46, 25, 22, 8, 8, 3, 5, 2). Logistic regression (black solid line) and phylogenetic logistic regression (best fit= black dashed line, confidence band=gray dashed lines) demonstrate a positive relationship.

The richness of PTS transporters is also positively related to $\log_2$-*rrn*, but this pattern is restricted to a subset of Firmicutes and Gammaproteobacteria. When phylogeny is included in the regression model the predicted PTS transporter richness for a 1 *rrn* bacterium is effectively no different from the prediction for a 15 *rrn* bacterium because both are much less than 1 PTS

transporter (Figure 4.5). While the presence of PTS transporters may be adaptive for certain clades of high-*rrn* bacteria, we cannot rule out the possibility that these traits are a product of historical contingency (Table 4.2 and Table 4.4).

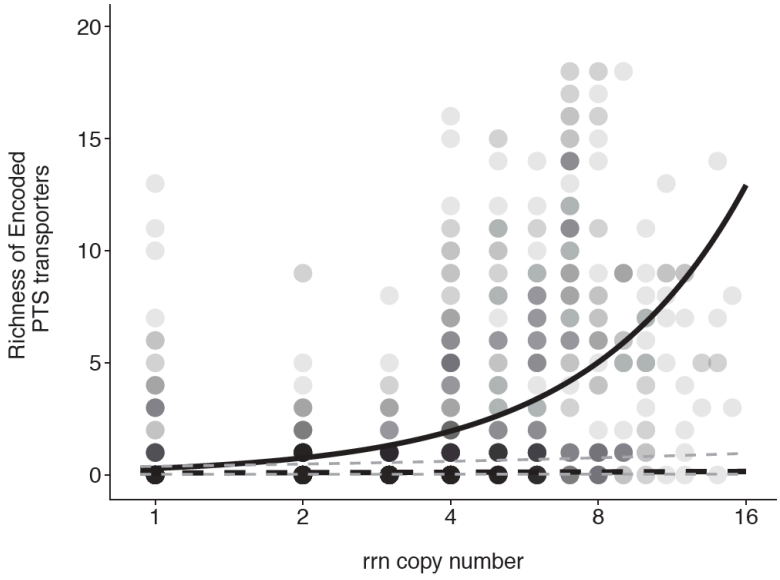


Figure 4.5: PTS transporter richness and $\log_2$-*rrn* copy number. The color density of each point represents the total number of genomes in each category (darker = more genomes). While Poisson regression model (black solid curve) indicates these traits are correlated, the best fit phylogenetic Poisson regression model (best fit=black dashed curve; confidence band=gray dashed curves) demonstrates no meaningful relationship between the variables was found after accounting for phylogenetic relationships among species.

On the other end of the life history spectrum, genome streamlining has been put forward as an oligotrophic adaptation which minimizes biosynthetic costs and increases nutrient use efficiency for oligotrophs (Giovannoni *et al.*, 2014). Streamlining has been implicated in thiamine biosynthesis gene loss for the oligotrophic SAR11 clade of Alphaproteobacteria (Carini *et al.*, 2014), which recycle a naturally abundant thiamine precursor rather than synthesize the molecule *de novo*. I propose that genome streamlining will cause genome size and the number of thiamine biosynthesis orthologs to be positively correlated with *rrn* copy number. Oxygenic photoautotrophy has also been hypothesized as an oligotrophic adaptation (Raven *et al.*, 2005),

but as organic resources are depleted from environments, we expect that all autotrophs – organisms that use inorganic carbon for biosynthesis – would have a large fitness advantage. I asked more generally if the presence of genes that permit autotrophic carbon fixation are correlated with *rrn* copy number.

Genome size is positively correlated with $\log_2$-*rrn* (Table 4.2 and Figure 4.6) and the phylogenetic regression model predicts an average 15 *rrn* copy bacterium possesses 0.66 Mbp more DNA than an average 1 *rrn* copy bacterium. Incorporating phylogeny slightly decreases the slope of the regression model, but shared ancestry cannot explain this relationship.

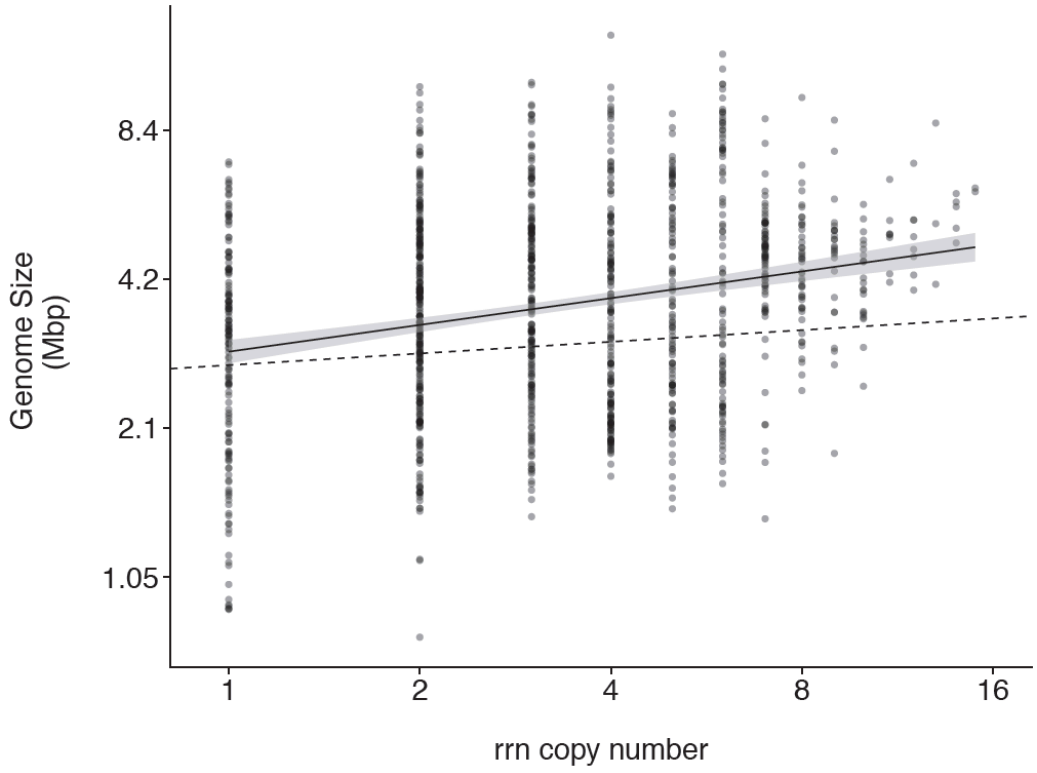


Figure 4.6: Log_2-transformed genome size correlates with $\log_2$-*rrn* copy number. The color density of each point represents the number of genomes at that location in the plot, with darker gray indicating more genomes. Linear regression (black line with gray confidence band) and phylogenetic linear regression (black dashed line) demonstrate a logarithmic, positive relationship between genome size and *rrn* copy number.

I interpret this result to reflect the process of genome streamlining, gene loss due to selection on efficient resource use, for two reasons. First, I excluded genomes of symbiotic, parasitic, and commensal bacteria which are undergoing genome degradation due to genetic drift. Additionally, the composition of the thiamine biosynthesis pathway provides support for selection leading to small genomes in bacteria with few *rrn* operons. Thiamine and its molecular precursors are secreted by microbes in many environments (Carini *et al.*, 2014; Strzelczyk & Leniarska, 1985), allowing auxotrophs to become dominant community members (Giovannoni *et al.*, 2005). High *rrn* copy number bacteria tend to possess more *de novo* thiamine biosynthesis genes (Figure 4.7, Table 4.2 and Table 4.4), while gene loss in low *rrn* bacteria is not distributed evenly in the pathway.

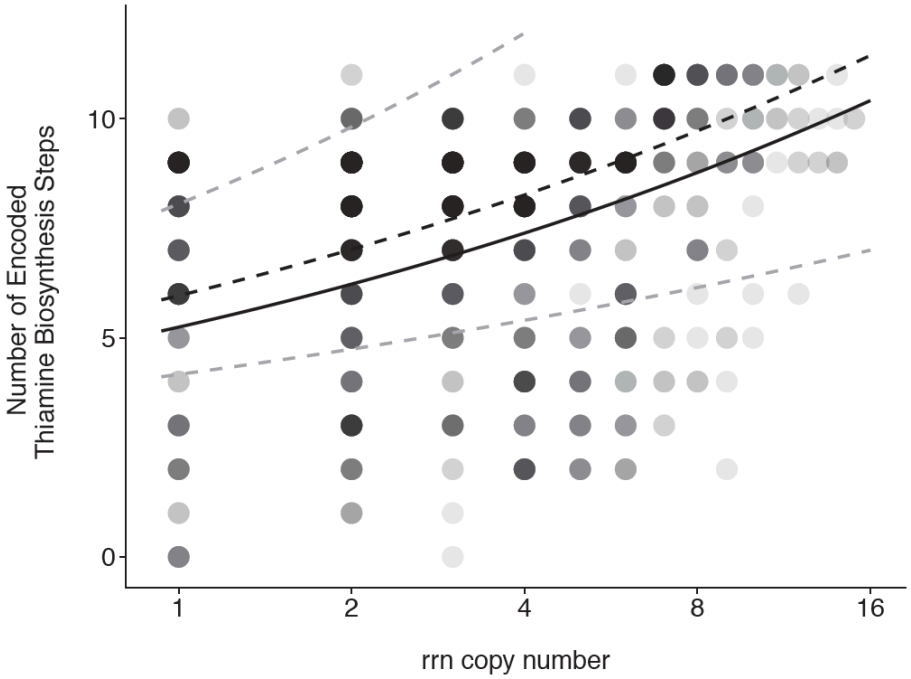


Figure 4.7: The number of encoded thiamine biosynthesis steps correlates with $\log_2$-*rrn* copy number. The color density of each point represents the total number of genomes in each category (darker = more genomes). Non-phylogenetic Poisson regression (black solid curve) and phylogenetic Poisson regression (best fit = black dashed curve; confidence bands = gray dashed curves) models demonstrate a positive relationship (Table 4.1 and Table 4.2).

Genes contributing to the recycling of thiamine or its precursors are equally present across the *rrn* spectrum, while genes uninvolved with recycling are less abundant in low *rrn* bacteria (Table 4.4). These findings are not consistent with gene loss due to random genetic drift, suggesting selection acts in low *rrn* bacteria to minimize unnecessary biosynthetic reactions when exogenous salvage of biosynthesis products is possible.

The probability of a bacterium being autotrophic, *i.e.*, encoding any of four autotrophic pathways – see Methods, increased with decreasing $\log_2$-*rrn* (Table 4.2, Table 4.4, and Figure 4.8). However, this trend could plausibly be explained by evolutionary history alone and this is likely due to the large number of oxygenic photoautotrophs in the dataset.

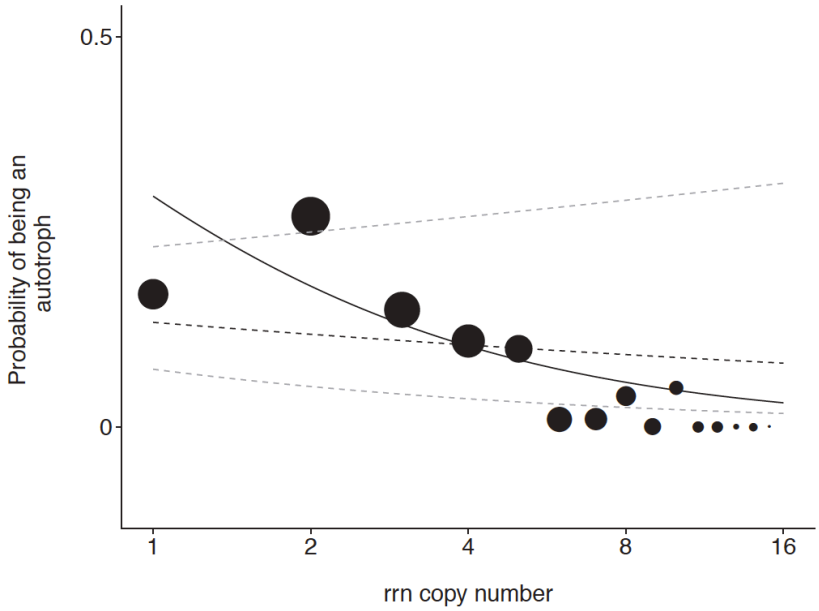


Figure 4.8: Autotrophy and *rrn* copy number. The proportion of genomes possessing autotrophy for each *rrn* copy number is represented by circles which are sized proportionally to the number of genomes (for *rrn* 1-15, respectively, N = 150, 262, 195, 165, 104, 96, 76, 46, 25, 22, 8, 8, 3, 5, 2). Logistic regression (black solid curve) indicates a negative relationship, but phylogenetic logistic regression (best fit = black, dashed curve; confidence band = gray, dashed curves) indicates that there is no relationship between these variables after accounting for shared ancestry among species.

Oxygenic photosynthesis is a complex trait involving many gene products. It has been demonstrated that complex traits are more likely to be conserved in deep-branching clades (Martiny *et al.*, 2012). The monophyletic Cyanobacteria all tend to have a low *rrn* copy number and make up a large proportion of the total autotrophs in the dataset. This combination of factors likely produces the strong phylogenetic signal seen for autotrophy, and reduces the predictive relationship of $\log_2$-*rrn* for this trait.

Trait	Regression Model	Non-phylogenetic regression Estimated effect size over full *rrn* spectrum (p-value)	Phylogenetic regression Estimated effect size over full *rrn* spectrum (significance measure)
Thiamine Biosynthesis (full pathway)	Poisson	+3 biosynthetic steps (**p<0.001**)	+3 biosynthetic steps (**pMCMC < 0.001**)
Thiamine biosynthesis (8 non-recycling orthologs)	Poisson	+2.5 biosynthetic steps (**p < 0.001**)	< +2.5 biosynthetic steps (**pMCMC < 0.001**)
Thiamine biosynthesis (all 8 non-recycling orthologs)	Logistic	+71.6% probability (**p < 0.001**)	+7.8% probability (**p = 0.02**)
Thiamine biosynthesis (3 recycling orthologs)	Poisson	< + 1 biosynthetic steps (p = 0.087)	< + 1 biosynthetic steps (**p = 0.006**)
Thiamine biosynthesis (all 3 recycling orthologs)	Logistic	+1.4% probability (p = 0.405)	-1% probability (p = 0.420)
Chemotactic motility	Logistic	+41.3% probability (**p<0.001**)	+11% probability (**p=0.035**)
Number of Encoded PTS system Transporters	Poisson	+12 PTS transporters (**p<0.001**)	< +1 PTS transporter (**pMCMC <0.003**)
Oxygenic photoautotrophy	Logistic	-7.7% probability (**p<0.001**)	<+0.1% probability (p=0.5)
Autotrophy	Logistic	-26.3% probability (**p<0.001**)	-5.1% probability (p=0.131)

Table 4.4: Expanded genomic trait summary statistics. Life history trait regression results from statistical models with and without phylogenetic correction. Effect size calculated in terms of *rrn* doubling for linear regression models, and in terms of trait change over the extant *rrn* spectrum (1-15 copies) for non-linear regression models.

Genome content covaries with *rrn* copy number

Building on the previous results, I explored $\log_2$-*rrn* copy number could explain variation in the ortholog content of entire bacterial genomes. I evaluated the presence of more than 7,000

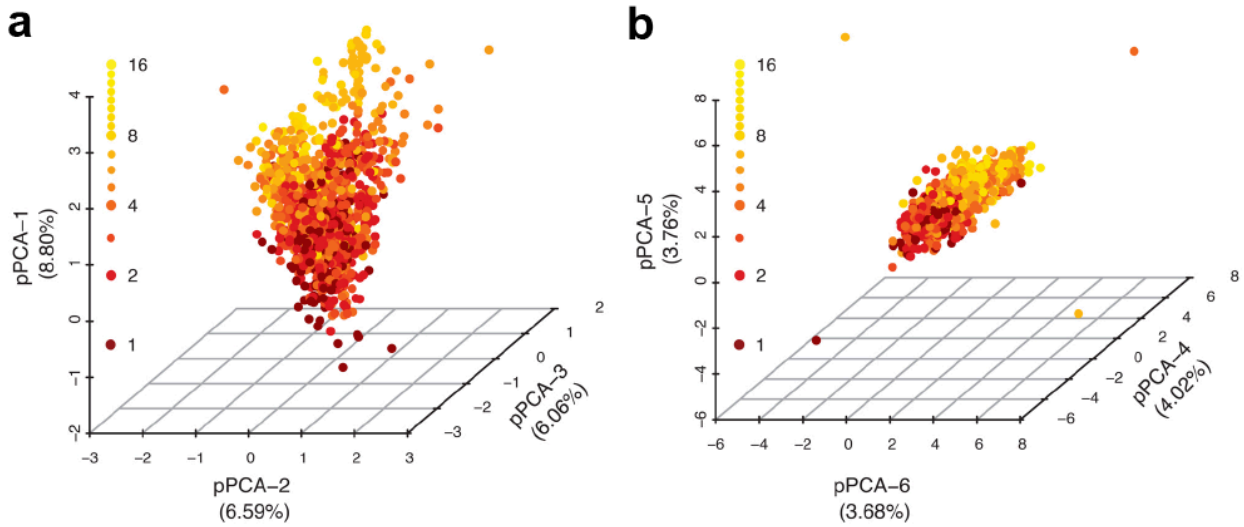
orthologs and 400 modules – combinations of orthologous genes which functionally interact as either a sub-pathway or enzyme complex (Kanehisa, 2013) – among 1,167 bacterial genomes using phylogenetic principal components analysis (pPCA) to control for shared evolutionary history. pPCA summarizes the variability present in the module and ortholog datasets by creating new axes which are combinations of the original variables while incorporating the evolutionary relationships among species. Although pPCA considers phylogeny, the position of each species on the new pPCA axes (often referred to as species scores) must be regressed against explanatory variables, in this case $\log_2$-*rrn*, using phylogenetic regression to sufficiently reduce the false positive rate to a commonly accepted level (Revell, 2009).

The pPCA analyses effectively reduced the dimensionality of both ortholog and module datasets; the first 10 axes of the ortholog analysis explained approximately 39% of the variation in the data, while the first 10 axes explained approximately 48% of the variation in the modules dataset. Performing phylogenetic linear regression on species scores for the modules pPCA analysis reveals eight of the first ten pPCA axes were correlated with $\log_2$-*rrn* (Figure 4.9a and c). A similar trend was seen in the ortholog pPCA analysis, where nine of the first ten pPCA axes were significantly correlated with $\log_2$-*rrn* (Figure 4.9b and c). This suggests life history evolution is a strong force driving genome content toward similar adaptations across the bacterial tree of life.

Discussion

Microbial ecologists have frequently classified bacteria based on their response to nutrient availability (Andrews & Rouse, 1982; Button, 1991; 1998), and the concepts of copiotrophy and oligotrophy have become the pervasive framework for describing this idea

(Kuznetsov *et al.*, 1979; AL Koch, 2001; Lauro *et al.*, 2009). While many acknowledge a spectrum of fitness variation between copiotrophy and oligotrophy, most studies exploring life history variation classify bacteria into only these two categories. I provide evidence in this study that *rrn* copy number is a quantitative proxy for life history which can delineate an entire spectrum of fitness variation in response to resource availability. Not only are rapid and efficient growth differentially beneficial for low and high *rrn* bacteria, but intermediate values of these



c

Axis	Variance explained	pPCA-*rrn* slope	pPCA-*rrn* p value
pPCA-1	8.80%	0.4781	**< 2.0x10^{-16}**
pPCA-2	6.59%	-0.1920	**< 2.2x10^{-16}**
pPCA-3	6.06%	0.1754	**< 2.2x10^{-16}**
pPCA-4	5.34%	-0.1046	**4.3x10^{-9}**
pPCA-5	4.48%	-0.1113	**8.5x10^{-11}**
pPCA-6	3.99%	0.0402	**0.0101**
pPCA-7	3.50%	-0.0497	**0.0086**
pPCA-8	3.21%	-0.0581	**0.0002**
pPCA-9	3.01%	-0.0253	0.0501
pPCA-10	2.72%	-0.0273	0.0555

Axis	Variance explained	pPCA-*rrn* slope	pPCA-*rrn* p value
pPCA-1	5.78%	0.1238	**0.0309**
pPCA-2	5.69%	-0.3431	**4.5x10^{-8}**
pPCA-3	4.88%	-0.2709	**2.3x10^{-7}**
pPCA-4	4.02%	0.5071	**< 2.0x10^{-16}**
pPCA-5	3.76%	0.4124	**< 2.2x10^{-16}**
pPCA-6	3.68%	0.5389	**< 2.0x10^{-16}**
pPCA-7	3.07%	-0.1195	**0.0035**
pPCA-8	2.90%	0.4040	**4.87x10^{-14}**
pPCA-9	2.59%	0.2642	**7.64x10^{-10}**
pPCA-10	2.55%	-0.0001	0.9969

Figure 4.9: Genome content covaries with log$_2$-*rrn* copy number. Phylogenetic principal component analysis (pPCA) of the genomic content of KEGG modules (a) or KEGG orthologs (b) in 1,167 unique bacterial species. Summary statistics for pPCA analysis and regressions of species scores against log$_2$-*rrn* are reported below their corresponding analysis (c). The three axes displayed in panels a & b correspond to pPCA axes which explain a high proportion of variance and have a strong regression slope.

growth phenotypes are observed for bacteria with intermediate *rrn* copy number. Additionally, postulated life history traits were correlated with *rrn* copy number in the expected direction. An important result from this study is that many of the life history trait-*rrn* correlations can not be explained by shared evolutionary history. This indicates that *rrn* copy number is a robust predictor of life history throughout the considerable phylogenetic breadth of the bacterial domain.

These results provide new insight into bacterial life histories and the link with *rrn* copy number provides a means to incorporate life histories into models of community dynamics and function. For instance, following perturbations that increase the availability of resources that favor rapid growth, there should be an increase in the abundance of bacteria with high *rrn* copy number bacteria. This has been observed in multiple ecosystems, including the response of planktonic bacteria to the Deepwater Horizon oil spill in the Gulf of Mexico. Following the spill, there was a bloom of hydrocarbon-degrading bacteria classified as members of the genus *Colwellia* (Valentine *et al.*, 2010; Redmond & Valentine, 2012). Well-characterized relatives of these bacteria are known hydrocarbon degraders (Baelum *et al.*, 2012) and have 9 *rrn* copies in their genome (Methé *et al.*, 2005). In terrestrial environments, early successional bacteria encode for more *rrn* operons than late successional bacteria (Shrestha *et al.*, 2007) and bacteria that responded most quickly to the addition of 2,4-D (an herbicide that is metabolized by bacteria) had more *rrn* copies than those that responded slowly (Klappenbach *et al.*, 2000). In host-associated microbial communities, a bloom of high *rrn* copy number *Enterobacteriaceae* during antibiotic-associated diarrhea (Young & Schmidt, 2004) is coincident with the temporary increase in carbohydrates entering the colon. As suggested from these studies, the impact of altered nutrient flux following an environmental perturbation can be evaluated by inferring *rrn*

copy number from molecular surveys of bacterial 16S rRNA genes (Kembel *et al.*, 2012; Angly *et al.*, 2014). Inferring the *rrn* copy number distribution in soil community surveys should also allow for improved predictions of CUE in natural systems. Recent soil carbon modeling efforts suggest improvements can come from allowing variable CUE parameters and the inclusion of microbial community composition data (Allison *et al.*, 2010; Wieder *et al.*, 2013). Bacterial genome sequencing is rapidly outpacing physiological characterization, and it is becoming increasingly common that the only thing known about a bacterium is its genome sequence. Findings from this study can be applied to help generate hypotheses about the natural history and physiology of these bacteria.

Theoretical biologists have proposed a tradeoff in the rate and efficiency of heterotrophic growth is inevitable based on thermodynamic constraints of ATP production (Pfeiffer *et al.*, 2001). My results are consistent with a rate-efficiency tradeoff: there is a negative relationship between $\log_2$-transformed maximum growth rate and growth efficiency (Figure 4.10). This evidence does not demonstrate an evolutionary tradeoff between the rate and efficiency of growth, which would require an experimental evolution approach. Understanding the mechanisms underlying rapid and efficient bacterial growth is essential if managing microbiomes becomes a priority for human and environmental health. I have demonstrated that *rrn* copy number is a quantitative marker of the life history tactics of rapid and efficient growth and provide a phylogenetically informed approach which can allow for new insights into the genomic features underlying a key dimension of bacterial fitness.

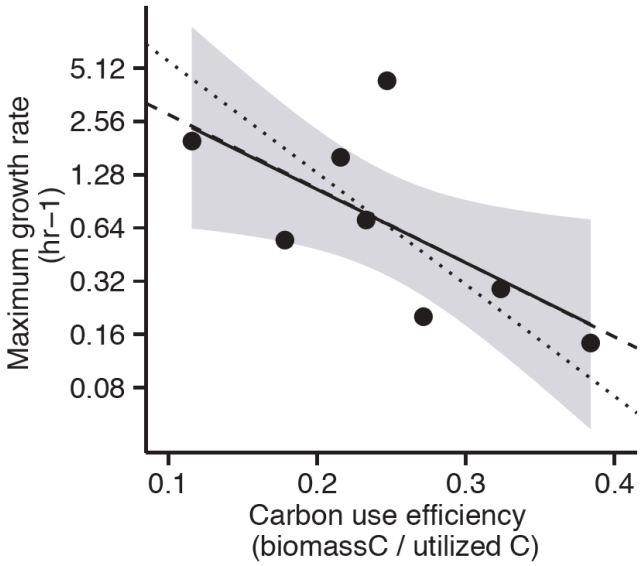


Figure 4.10: Growth rate and efficiency are inversely correlated. Maximum recorded growth rate and ^{14}CUE of 8 bacteria from the efficiency study. OLS regression (solid black line) with 95% confidence band (gray shading), phylogenetic regression (dashed line) and phylogenetic RMA regression (dotted line) demonstrate a negative relationship.^{14}CUE OLS regression: R^2 =0.44, slope = -13.75 , 1-tailed p = 0.037; ^{14}CUE RMA phylogenetic regression: slope = -25.75, 1-tailed p = 0.033. ^{14}CUE phylogenetic regression: slope = -13.90, p = 0.037; ^{14}CUE phylogenetic RMA slope = -20.94, $p < 0.001$.

REFERENCES

REFERENCES

Ackermann M. (2015). The usefulness of evolutionary principles: predicting the unexpected. *Environmental Microbiology Reports* **7**:4–5.

Aiyar SE, Gaal T, Gourse RL. (2002). rRNA Promoter Activity in the Fast-Growing Bacterium Vibrio natriegens. *Journal of Bacteriology* **184**:1349–1358.

Allison SD, Wallenstein MD, Bradford MA. (2010). Soil-carbon response to warming dependent on microbial physiology. *Nature Geoscience* **3**:336–340.

Andrews JH, Rouse DI. (1982). Plant pathogens and the theory of r-and K-selection. *American Naturalist*.

Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. (2014). CopyRighter: a rapid tool for improving theaccuracy of microbial community profiles throughlineage-specific gene copy number correction. **2**:1–13.

Baelum J, Borglin S, Chakraborty R, Fortney JL, Lamendella R, Mason OU, *et al.* (2012). Deep-sea bacteria enriched by oil and dispersant from the Deepwater Horizon spill. *Environmental Microbiology* **14**:2405–2416.

Blomberg S, Garland J, Ives A. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**:717–745.

Blomberg SP, Garland T. (2002). Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology* **15**:899–910.

Button DK. (1991). Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the michaelis constant. *Appl Environ Microbiol* **57**:2033–2038.

Button DK. (1998). Nutrient uptake by microorganisms according to kinetic parameters from theory as related to cytoarchitecture. *Microbiology and Molecular Biology Reviews* **62**:636–645.

Carini P, Campbell EO, Morré J, Sañudo-Wilhelmy SA, Thrash JC, Bennett SE, *et al.* (2014). Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. *ISME J* **8**:1727–1738.

Conn HJ. (1930). The identity of Bacillus subtilis. *The Journal of Infectious Diseases*.

Datta S, Costantino N, Court DL. (2006). A set of recombineering plasmids for gram-negative bacteria. *Gene* **379**:109–115.

Dethlefsen L, Schmidt TM. (2007). Performance of the translational apparatus varies with the ecological strategies of bacteria. *Journal of Bacteriology* **189**:3237–3245.

Eagon R. (1962). Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *Journal of Bacteriology* **83**:736.

Eichorst SA, Breznak JA, Schmidt TM. (2007). Isolation and Characterization of Soil Bacteria That Define Terriglobus gen. nov., in the Phylum Acidobacteria. *Appl Environ Microbiol* **73**:2708–2717.

Eichorst SA, Kuske CR, Schmidt TM. (2011). Influence of Plant Polymers on the Distribution and Cultivation of Bacteria in the Phylum Acidobacteria. *Appl Environ Microbiol* **77**:586–596.

Fegatella F, Lim J, Kjelleberg S, Cavicchioli R. (1998). Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. *Appl Environ Microbiol* **64**:4433–4438.

Felsenstein J. (1985). Phylogenies and the Comparative Method. *The American Naturalist* **125**:1–15.

Fierer N, Bradford M, Jackson R. (2007). Toward an ecological classification of soil bacteria. *Ecology* **88**:1354–1364.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* 1–13.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**:1242–1245.

Gorlach K, Shingaki R, Morisaki H, Hattori T. (1994). Construction of eco-collection of paddy field soil bacteria for population analysis. *Journal of General and Applied Microbiology* **40**:509–517.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, *et al.* (2012). Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* **3**:e00252–12–e00252–12.

Hadfield J. (2010). MCMC Methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* **33**.

Jurgenson CT, Begley TP, Ealick SE. (2009). The Structural and Biochemical Foundations of Thiamin Biosynthesis. *Annu Rev Biochem* **78**:569–603.

Kanehisa M. (2013). Chemical and genomic evolution of enzyme-catalyzed reaction networks. *FEBS Letters* **587**:2731–2737.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**:D199–D205.

Kembel SW, Wu M, Eisen JA, Green JL. (2012). Incorporating 16S Gene Copy Number

Information Improves Estimates of Microbial Diversity and Abundance Mering, Von, C (ed). *PLoS Comp Biol* **8**:e1002743.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**:1328–1333.

Koch A. (2001). Oligotrophs versus copiotrophs. *Bioessays* **23**:657–661.

Koch H, Galushko A, Albertsen M, Schintlmeister A, Gruber-Dorninger C, Lücker S, *et al.* (2014). Growth of nitrite-oxidizing bacteria by aerobic hydrogen oxidation. *Science* **345**:1052–1054.

Kuznetsov S, Dubinina G, Lapteva N. (1979). Biology of oligotrophic bacteria. *Annual Reviews in Microbiology* **33**:377–387.

Kümmerli R, Schiessl KT, Waldvogel T, McNeill K, Ackermann M. (2014). Habitat structure and the evolution of diffusible siderophores in bacteria Van Baalen, M (ed). *Ecol Lett* **17**:1536–1544.

Lauro F, McDougald D, Thomas T, Williams T, Egan S, Rice S, *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences* **106**:15527–15533.

Martiny AC, Treseder K, Pusch G. (2012). Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7**:830–838.

Methé BA, Nelson KE, Deming JW, Momen B, Melamud E, Zhang X, *et al.* (2005). The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. *Proc Natl Acad Sci USA* **102**:10913–10918.

Miller LD, Russell MH, Alexandre G. (2009). Diversity in Bacterial Chemotactic Responses and Niche Adaptation. 1st ed. Elesvier Inc.

Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer K-H, *et al.* (2011). Release LTPs104 of the All-Species Living Tree. *Syst Appl Microbiol* **34**:169–170.

Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**:289–290.

Pfeiffer T, Schuster S, Bonhoeffer S. (2001). Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**:504–507.

Porter SL, Wadhams GH, Armitage JP. (2011). Signal processing in complex chemotaxis pathways. *Nat Rev Micro* **9**:153–165.

Prosser JI. (2015). Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nat Rev Micro* **13**:439–446.

R Core Team. (2014). R: A Language and Environment for Statistical Computing. *wwwR-projectorg*. http://www.R-project.org/ (Accessed July 23, 2015).

Raven JR, Andrews M, Quigg A. (2005). The evolution of oligotrophy: implications for the breeding of crop plants for low input agricultural systems. *Annals of Applied Biology* **146**:261–280.

Redmond MC, Valentine DL. (2012). Natural gas and temperature structured a microbial community response to the Deepwater Horizon oil spill.

Revell LJ. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**:217–223.

Revell LJ. (2009). Size-correction and principal components for interspecific comparative studies. *Evolution* **63**:3258–3268.

Roller BRK, Schmidt TM. (2015). The physiology and ecological implications of efficient growth. *ISME J* **9**:1481–1487.

Schaechter M, Maaloe O, Kjeldgaard N. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced growth of Salmonella typhimurium. *Microbiology* **19**:592.

Schut F, de Vries EJ, Gottschal JC, Robertson BR, Harder W, Prins RA, *et al.* (1993). Isolation of Typical Marine Bacteria by Dilution Culture: Growth, Maintenance, and Characteristics of Isolates under Laboratory Conditions. *Appl Environ Microbiol* **59**:2150–2160.

Schut F, Gottschal JC, Prins RA. (1997). Isolation and characterisation of the marine ultramicrobacteriumSphingomonassp. strain RB2256. *FEMS Microbiol Rev* **20**:363–369.

Shrestha PM, Noll M, Liesack W. (2007). Phylogenetic identity, growth-response time and rRNA operon copy number of soil bacteria indicate different stages of community succession. *Environmental Microbiology* **9**:2464–2474.

Simon M, Azam F. (1989). Protein content and protein synthesis rates of planktonic marine bacteria. *Marine ecology progress series Oldendorf* **51**:201–213.

Singh BK, Bardgett RD, Smith P, Reay DS. (2010). Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nat Rev Micro* **8**:779–790.

Sorokin DY, cker SLU, Vejmelkova D, Kostrikina NA, Kleerebezem R, Rijpstra WIC, *et al.* (2012). Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi. *ISME J* **6**:2245–2256.

Stearns SC. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften* **87**:476–486.

Stevenson BS, Eichorst SA, Wertz JT, Schmidt TM, Breznak JA. (2004). New Strategies for

Cultivation and Detection of Previously Uncultured Microbes. *Appl Environ Microbiol* **70**:4748–4755.

Stevenson BS, Schmidt TM. (2004). Life history implications of rRNA gene copy number in Escherichia coli. *Appl Environ Microbiol* **70**:6670.

Stocker R. (2012). Marine Microbes See a Sea of Gradients. *Science* **338**:628–633.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. (2014). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*.

Stouthamer AH. (1973). A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* **39**:545–565.

Strzelczyk E, Leniarska U. (1985). Production of B-group vitamins by mycorrhizal fungi and actinomycetes isolated from the root zone of pine (Pinus sylvestris L.). *Plant and soil*.

Taylor JR, Stocker R. (2012). Trade-Offs of Chemotactic Foraging in Turbulent Water. *Science* **338**:675–679.

Tung Ho LS, Ane C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic biology* **63**:397–408.

Valentine DL, Kessler JD, Redmond MC, Mendes SD, Heintz MB, Farwell C, *et al.* (2010). Propane respiration jump-starts microbial response to a deep oil spill. *Science* **330**:208–211.

Vieira-Silva S, Rocha EPC. (2010). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet* **6**:e1000808.

Wickham H. (2009). ggplot2: elegant graphics for data analysis. Springer: New York.

Wieder WR, Bonan GB, Allison SD. (2013). Global soil carbon projections are improved by modelling microbial processes. *Nature Climate Change* **3**:1–4.

Young VB, Schmidt TM. (2004). Antibiotic-Associated Diarrhea Accompanied by Large-Scale Alterations in the Composition of the Fecal Microbiota. *Journal of Clinical Microbiology* **42**:1203–1206.

CHAPTER 5

Axes of life history variation explain the genome content of bacteria

Abstract

Characterizing the niche and life history of a bacterium using traditional methods is a slow and challenging process. The diversity of bacteria in nature is immense and the cultivation of environmental bacteria is often technically demanding. Given that the scale of unexplored bacterial diversity is much greater than our ability to characterize it using traditional laboratory methods, new solutions must be sought to improve our understanding of bacteria. I propose that the torrent of genome sequences being produced provides an opportunity to understand the ecological and evolutionary factors influencing bacteria. In this study I explore how genome content correlates with important axes of the environment thought to influence a microbe's niche and life history. I implement phylogenetic comparative methods to control for the influence of shared ancestry on genome similarity. This approach highlights that niche and life history factors explain patterns in genome content among diverse bacterial species. The number of rRNA operons (*rrn*) present in a bacterium's genome provides a proxy for its place on the life history spectrum from copiotrophy to oligotrophy. *rrn* copy number is a strong predictor of genome content after controlling for evolutionary history, suggesting that resource competition is important for bacterial genome evolution. Genome content similarity was only marginally explained by oxygen requirements, and not at all by the temperature range of bacteria, after controlling for evolutionary history. Examination of the genome features driving the relationship

between genome content and life history provides new insight into how bacteria adapt to life at different resource concentrations.

Introduction

Genome sequencing has revolutionized microbiology. It has shed light on genome differences within species (Tettelin *et al.*, 2005; Medini *et al.*, 2008) and provided a glimpse at previously unobserved candidate phyla (Brown *et al.*, 2015). Extensive genome sequencing is also beginning to illuminate the ecological and evolutionary forces acting within and between bacterial populations (Cordero & Polz, 2014). Small genomes are found in bacteria with either extremely small or very large effective population size, but this shared feature is the outcome of disparate processes. Small effective population size amplifies the power of genetic drift and can lead to genome degradation (Mccutcheon & Moran, 2011), while large effective population size increases the influence of selection, which can purge costly genome features via streamlining (Giovannoni *et al.*, 2014). The genome content of bacteria can also be influenced by environmental factors. Distantly related bacteria residing in mammalian GI tracts are more similar to each other in gene content than are genomes of the same phylogenetic relatedness but which come from other environments. This similarity suggests that shared selective pressures are acting on gut bacteria (Zaneveld *et al.*, 2010). These examples show the great potential for genomics to improve our understanding of the evolutionary ecology of bacteria.

Genomes have been used to inform the challenging endeavor of characterizing the niche and the life history of a bacterium. The range of environmental conditions where a bacterium can persist is its ecological niche, while the pattern of fitness within the niche is the bacterium's life history. The distribution of protein domains among bacterial genomes is related to both their phylogeny and their environmental preferences (Suen *et al.*, 2007), supporting the idea that

ecology and evolution interact to influence bacterial genome content. Resource concentration is a key dimension of bacterial life history variation. Oligotrophic bacteria have a higher relative fitness under low resource concentrations, while copiotrophic bacteria are favored when resources are abundant. Recent studies comparing the genome content of copiotrophs and oligotrophs have concluded that there is a genomic basis to this axis of life history variation (Luo *et al.*, 2013; Lauro *et al.*, 2009). Exciting as these findings are, these studies were restricted primarily to Proteobacteria and the comparisons between life histories were confounded by phylogeny. It is possible that shared ancestry alone explains the genomic differences found in these comparisons. Life history evolution in other phylogenetic groups might have taken a completely separate path. Therefore, understanding if diverse bacterial genomes share adaptations to the environmental pressures exerted by extreme resource concentrations remains an open question.

The goal of this study is to explore more broadly how bacterial genome content relates to three axes of a bacterium's niche: oxygen concentration, temperature, and resource concentration. I ask if these three niche variables could explain the ortholog content of approximately one thousand bacterial genomes after controlling for the influence of shared evolutionary history. All of these variables are likely to have a major influence on competitive fitness in natural environments and they have been shown to covary with genome features (Vieira-Silva & Rocha, 2010; Sabath *et al.*, 2013; Morris & Schmidt, 2013; Wu & Moore, 2010; Lauro *et al.*, 2009). I utilized the number of rRNA operon (*rrn*) copies encoded in the genome as a proxy for resource concentration preference (Chapter 4), while information on bacterial oxygen requirements and temperature ranges was collected from other studies. The relative influence of these axes of niche and life history variation on genome variation is not well understood, nor is

their interaction with shared evolutionary history. With this study I examine the extent to which different aspects of niche and life history variation can explain genome content similarity among bacterial species. I also explore universal genomic signatures associated with the copiotroph-oligotroph spectrum across the bacterial tree of life.

Materials and methods

Genomic data and metadata

The genomes of 1,167 bacterial species were analyzed using annotations provided by the Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa *et al.*, 2013) and as described in Chapter 4 of this thesis. Briefly, the current version of the database was downloaded in May 2014 and was used to assert the presence of every ortholog and module – combinations of orthologs inferred to function in concert – present in all genomes and their *rrn* operon copy number (Stoddard *et al.*, 2014). Genomes were excluded from further analysis if they displayed symptoms of genome degradation due to genetic drift, while attempting to preserve the presence of streamlined genomes in the dataset (Mccutcheon & Moran, 2011; Giovannoni *et al.*, 2014). A single representative genome for each unique bacterial species in the dataset was chosen to reflect the central tendency of *rrn* copy number for the species, among other criteria, and led to the final number of 1,167 genomes included in this study. NCBI taxonomy was also extracted from the KEGG database for each genome.

The oxygen requirements and temperature ranges associated with these genomes was then extracted from the Integrated Microbial Genomes (IMG) database (Markowitz *et al.*, 2013). Oxygen requirement was classified into six categories by IMG: obligate anaerobe, anaerobe, facultative, microaerophilic, aerobe and obligate aerobe. I merged the obligate categories into

their respective aerobe and anaerobe categories, resulting in four categories of oxygen requirement. Temperature range was classified into seven categories by IMG: psychrophile, psychrotrophic, psychrotolerant, mesophile, thermotolerant, thermophile and hyperthermophile. I merged psychrotrophic and psychrophile into one category named psychrophile. I also merged psychrotolerant, mesophile and thermotolerant into a single category named mesophile, resulting in four categories of temperature range. Ultimately, a subset of 932 genomes was generated that contained an estimate of *rrn* copy number, temperature range and oxygen requirement.

Phylogenetic tree construction

A phylogenetic tree for the 1,167 genomes in this study was built using the same methods used as in Chapter 4 of this thesis. Briefly, I downloaded aligned 16S rRNA gene sequences from Silva (http://www.arb-silva.de) for as many of the 1,167 genomes as possible. If an aligned sequence from the genome was not available through Silva, an aligned sequence from a separate sequencing effort on the same strain or from the type strain of that species was utilized. Phylogenetic trees were built using maximum likelihood estimation (RAxML 7.0.4) to generate the ten most likely trees based on the GTRMIX substitution model with 25 rate categories and the new rapid hill climbing algorithm in the software program Arb. The only base positions used to build the tree were those which are conserved in 50% of all bacterial species (Munoz *et al.*, 2011). All trees included 5 archael sequences for rooting, which were pruned prior to use of the tree in statistical analysis. A single tree with the most negative maximum likelihood was chosen for phylogenetically informed comparative analyses. This tree was then pruned to generate smaller trees when subsets of the dataset were analyzed.

Statistical analyses

The R statistical programming language was used for all analyses (R Core Team, 2014). The R package vegan was used to perform non-metric multi-dimensional scaling (nMDS) of genome content using a the Jaccard coefficient, which excludes joint absence in the dataset from further analysis (Oksanen *et al.*, 2014). The R package ape was used to import NEXUS formatted tree files, store tree objects, and prune tips from the trees (Paradis *et al.*, 2004).

The R package phytools was used for phylogenetic principal coordinates analysis (pPCA). Correlation-based pPCA and covariance-based pPCA were performed utilizing Brownian motion as the underlying model of expected trait evolution (Revell, 2012). pPCA combines variables into new axes which maximally summarize variation in low-dimensional space while accounting for non-independence of data points due to shared ancestry. The coordinates for each genome on the new pPCA axes were further analyzed using phylogenetic comparative methods, and this combination of methods decreases the false positive rate of hypothesis tests to their expected level (Revell, 2009). Correlation-based PCA standardizes the variance among all variables and is typically done when variables with different measurement units are present in the dataset. Covariance-based PCA does not standardize variance among variables. The results of pPCA analyses were correlated with the explanatory variable *rrn* copy number, so orthologs that are part of the *rrn* operon were excluded from genome content pPCA. This included all orthologs for the genes encoding 16S rRNA, 23S rRNA, 5S rRNA, and all tRNAs. The R package phylolm was used for phylogenetic linear regression of pPCA genome scores, with a random root Ornstein-Uhlenbeck evolutionary model used in the analyses (Tung Ho & Ane, 2014). Phylogenetic logistic regression of orthologs/modules as a function of $\log_2$-transformed *rrn* copy number was also performed using the phylolm package for

modules/orthologs with the strong pPCA loadings. The R package geiger was used to implement phylogenetic multivariate analysis of variance (MANOVA) on the pPCA coordinates of bacterial genomes (Harmon *et al.*, 2007). This method first performs a standard MANOVA and the generated Pillai-Bartlett F-statistic is then compared to a distribution of F-values simulating no relationship between variables on the provided phylogenetic tree. I ran phylogenetic MANOVA analyses using 1000 trait simulations. The R packages scatterplot3d (Ligges & Mächler, 2002) and ggplot2 (Wickham, 2009) were used to generate figures.

Results

Genome content is correlated with shared evolutionary history

I performed a preliminary examination of genome content variation among 1,167 unique bacterial species. Non-metric multidimensional scaling (nMDS) was implemented to summarize variation in two datasets which measured the presence or absence of orthologs or modules among genomes. These datasets contained information on the presence of over 7,000 orthologs and over 400 modules in all 1,167 genomes. The distance between points in nMDS ordination qualitatively illustrate similarity, so genomes close together are more likely to be similar than genomes which are further apart. The nMDS analysis indicates that a bacterium's phylum is a strong factor influencing its genome content because the three major phyla in the analysis (Firmicutes, Proteobacteria, and Actinobacteria) cluster into distinct groups (Figure 5.1). Phylogeny has previously been shown to explain a significant fraction of genome content variation (Zaneveld *et al.*, 2010; Snel *et al.*, 1999), and these results support that claim. It is interesting to note that genome content appears to correlate with changes in *rrn* copy number in parallel among the three major phyla. For example, genomes with low *rrn* copy number in the

modules ordination tend to co-occur in the bottom and left sections of their phylum. Higher *rrn* copy number is generally observed as a shift in position along the nMDS axis towards the top and right sections of each phylum. These parallel patterns suggest that genome content may covary with nutritional preference among the three major phyla in this study. However, these patterns could still result from phylogenetic signal below the phylum level and phylogenetic ordination methods are required to test this idea.

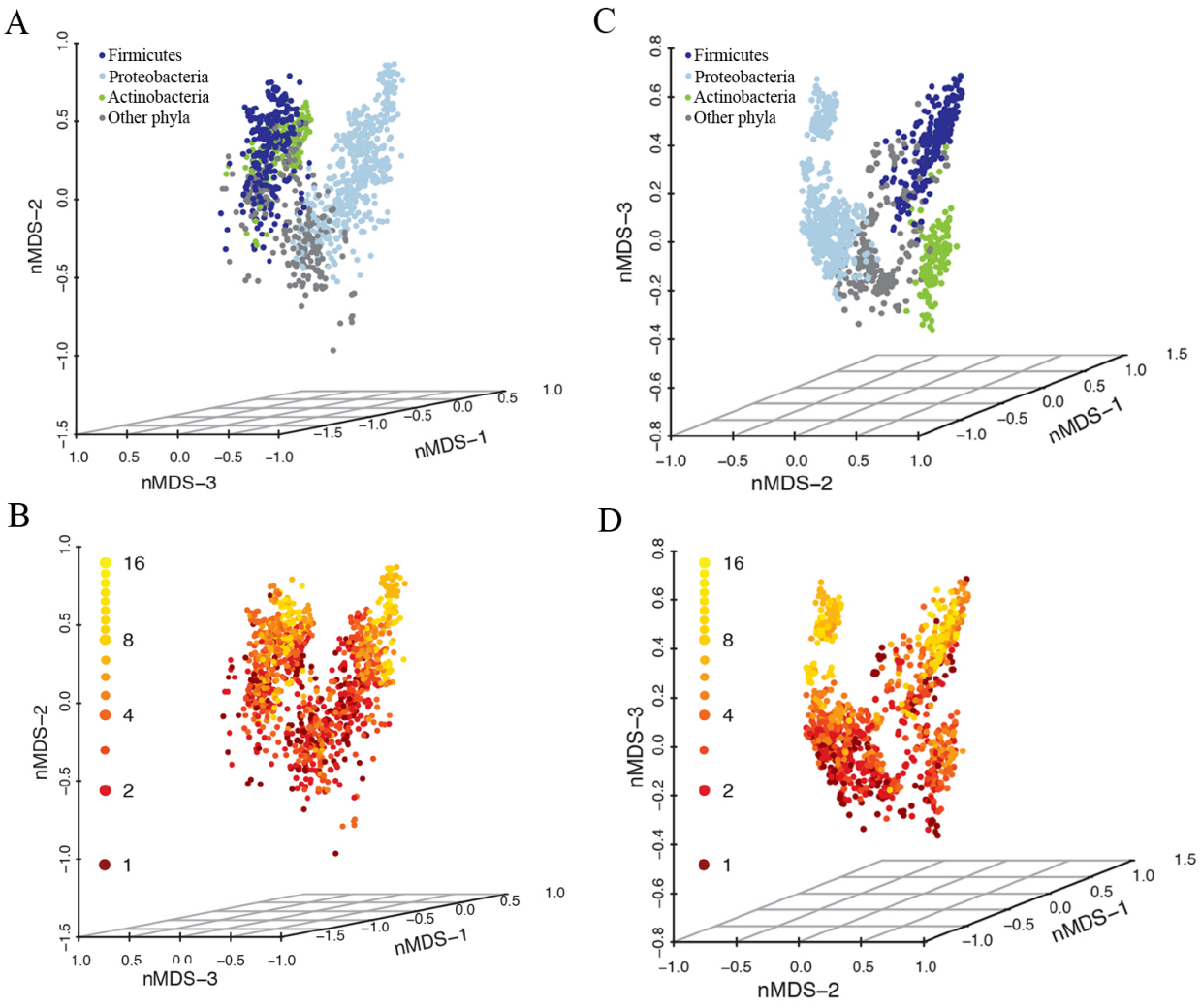


Figure 5.1: Genome content is related to evolutionary history. nMDS analysis of module (A and B) and ortholog (C and D) genome content, colored by phylum (A and C) or *rrn* operon copy number (B and D). A total of 10 nMDS dimensions were generated in the analysis (ortholog stress=0.038; module stress=0.058) and the first 3 are visualized.

Genome content is correlated with life history variation

I explored if genome content could be explained by three axes of niche and life history variation in bacteria: oxygen requirement, temperature range, and resource concentration. While oxygen requirement and temperature range were directly assessed, a proxy was used for a bacterium's resource concentration preference, $\log_2$-transformed *rrn* copy number (hereafter referred to as $\log_2$-*rrn*). Metadata describing the oxygen requirement and temperature range was available for 932 of the 1,167 genomes with known *rrn* copy number, so I restricted initial analysis of variables influencing genome content to these 932 genomes.

Phylogenetic principal components analysis (pPCA) was implemented to summarize the variance in ortholog and module content among bacterial genomes. Both correlation-based pPCA and covariance-based pPCA gave similar results, so unless otherwise noted correlation-based pPCA is reported for simplicity. pPCA was effective at collapsing the variation present in the genome content datasets. The first 50 pPCA axes explained a large fraction of the variation in each dataset, approximately 53% for modules and approximately 42% for orthologs. I tested if three niche variables could predict genome content variation on the first 50 pPCA axes using phylogenetic MANOVA.

A genome's position on the first 50 pPCA axes was significantly associated with all three niche variables, but only *rrn* copy number remained as a significant predictor of genome content after controlling for phylogeny (Table 5.1). This was true for both orthologs and modules datasets and when considering different numbers of pPCA axes in the analysis. Oxygen requirement was significantly associated with the first 10 pPCA axes of the ortholog dataset, but not the modules dataset. The location of genomes on first three pPCA axes for the ortholog dataset (Figure 5.2 and Figure 5.3) and the modules dataset (Figure 5.4) were visualized and

support the phylogenetic MANOVA results. There is clear clustering of genomes with similar *rrn* copy values on the first three pPCA axes, while it is much more difficult to observe genomes clustering by oxygen requirement and temperature range. *rrn* copy number has more explanatory power than the other two niche variables, suggesting resource concentration has a stronger influence on genome content variation than oxygen availability or temperature.

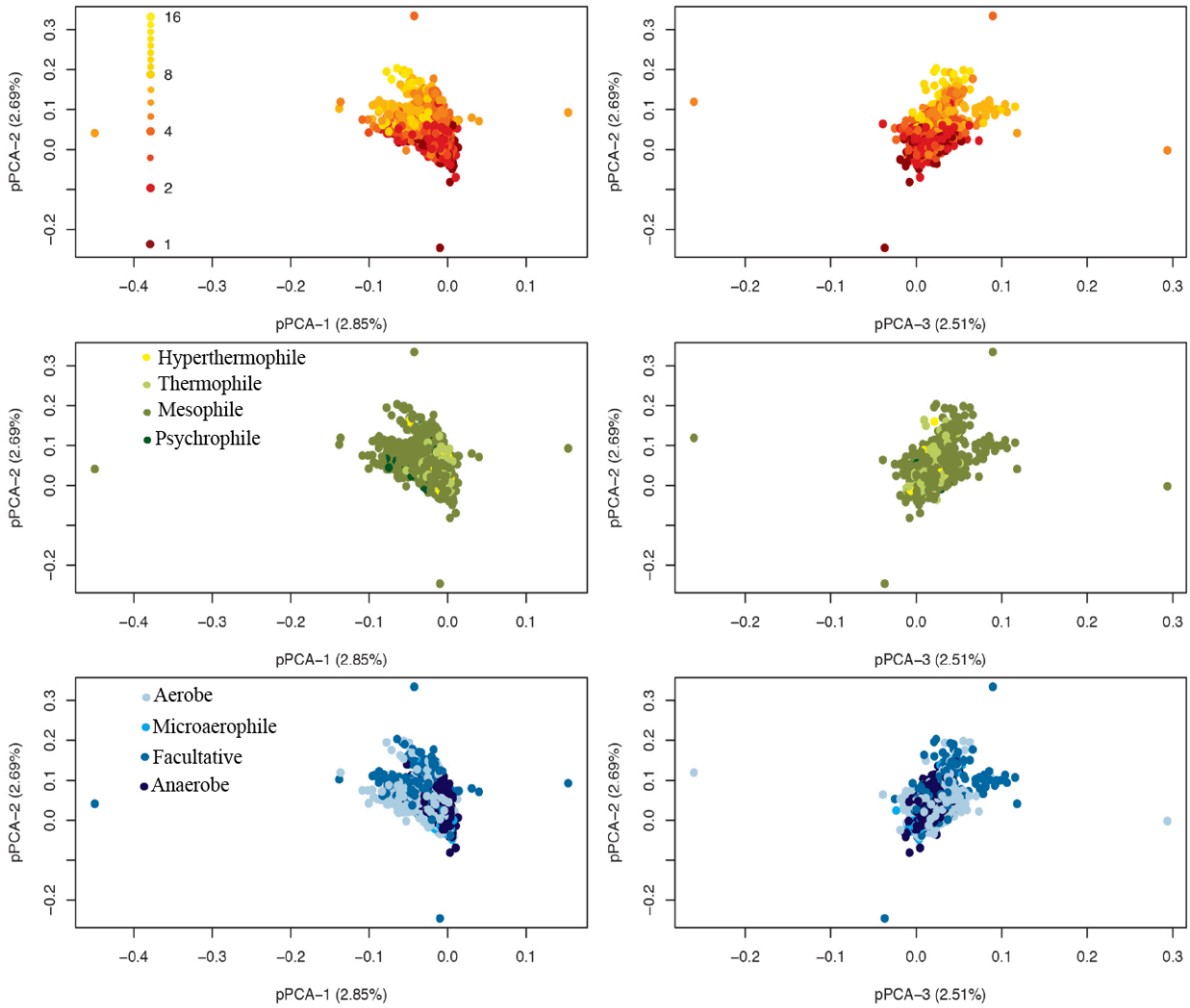


Figure 5.2: Correlation-based pPCA axes 1-3 of ortholog genome content. Clustering of genomes with similar *rrn* copy number is evident (A and B) while clustering is difficult to discern for temperature (C and D) and oxygen requirement (E and F).

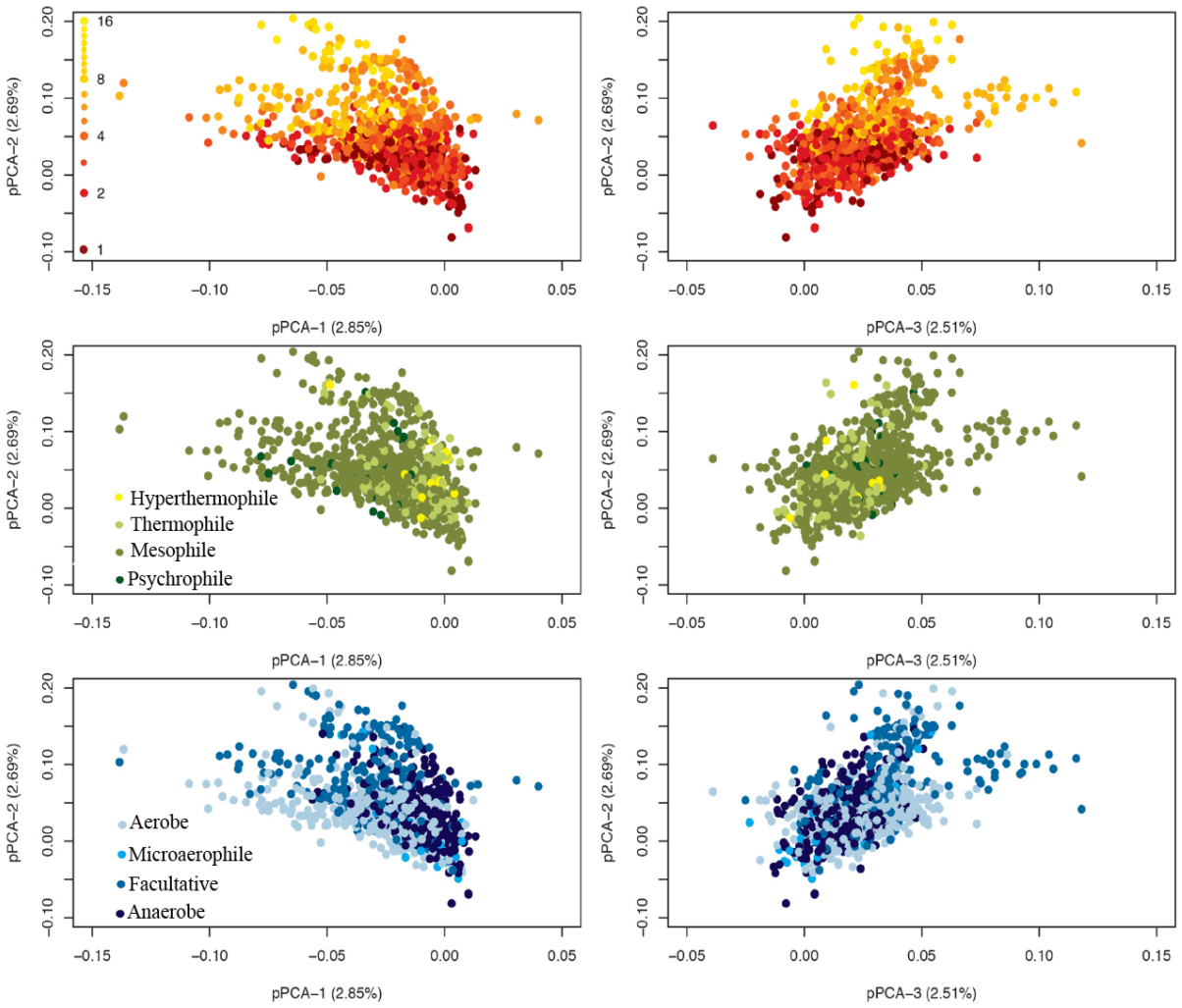


Figure 5.3: Correlation-based pPCA axes 1-3 of ortholog genome content, alternate view. Re-plotting of Figure 5.2, with correlation-based pPCA axes 1-3 of ortholog genome content trimmed to better observe patterns in the central cluster of genomes. Four data points with extreme pPCA axis coordinates are not visible in these plots. Clustering of genomes with similar *rrn* copy number is evident (A and B) while clustering is difficult to discern for temperature (C and D) and oxygen requirement (E and F).

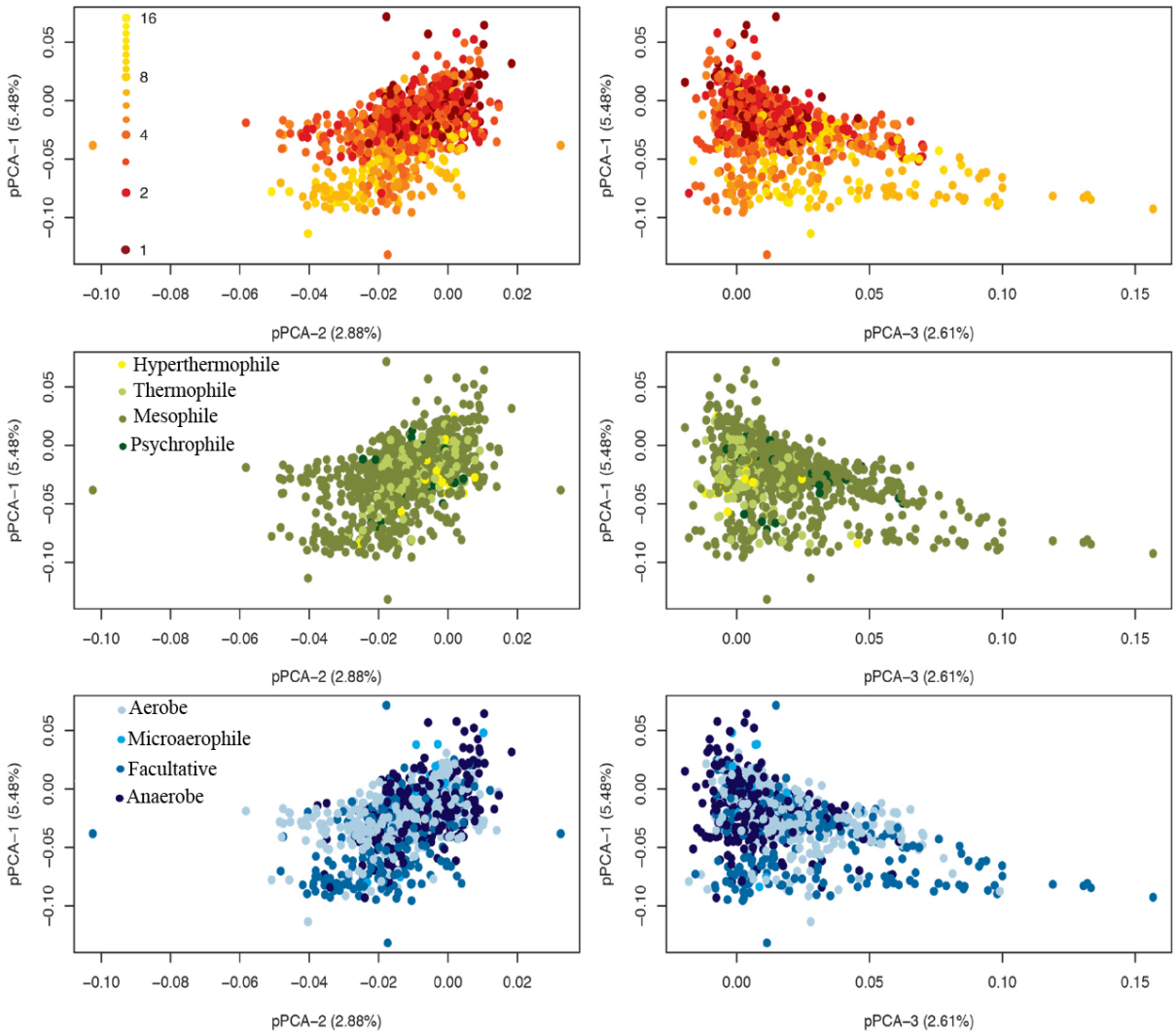


Figure 5.4: Correlation-based pPCA axes 1-3 of module genome content. Clustering of genomes with similar *rrn* copy number is evident (A and B) while clustering is difficult to discern for temperature (C and D) and oxygen requirement (E and F).

Explanatory variable	pPCA axes considered	Orthologs p (std.)	Orthologs p (phy)	Modules p (std.)	Modules p (phy)
Temperature	50	3.21E-25	1.000	5.04E-40	1.000
Temperature	30	8.33E-27	1.000	4.83E-26	1.000
Temperature	10	3.20E-13	0.965	1.66E-26	0.606
Temperature	5	1.17E-06	0.910	5.61E-18	0.431
O_2 requirement	50	2.45E-153	0.672	6.31E-133	0.977
O_2 requirement	30	7.65E-126	0.570	2.68E-125	0.563
O_2 requirement	10	1.10E-90	**0.041**	1.40E-87	0.068
O_2 requirement	5	2.92E-64	**0.036**	6.17E-55	0.103
rrn copies	50	4.00E-158	**<0.001**	1.63E-146	**<0.001**
rrn copies	30	2.42E-159	**<0.001**	3.02E-140	**<0.001**
rrn copies	10	4.79E-113	**<0.001**	5.21E-118	**<0.001**
rrn copies	5	8.54E-87	**<0.001**	1.90E-84	**<0.001**

Table 5.1: Genome content is associated with a bacterium's niche. Ortholog and module genome scores on correlation-based pPCA axes were consistently related to the *rrn* copy number of the genome based on phylogenetic MANOVA. The first 10 ortholog pPCA axes can also be explained by oxygen.

To get a better understanding of the influence of *rrn* copy number on genome content, I returned to the full 1,167 genome dataset. Including these two hundred additional genomes did not alter the relationship previously observed between ortholog or module content and *rrn* copy number. The first 50 pPCA axes explained approximately 40% of variation in the ortholog genome content, while the first 50 module pPCA axes explained approximately 53% of variation in that dataset. Phylogenetic MANOVA on all 50 ortholog and module pPCA axes demonstrated associations between genome content and *rrn* copy number ($p < 0.001$ for both orthologs and modules). Although phylogenetic MANOVA was useful for comparing the influence of three niche/life history dimensions on genome content, this method converts *rrn* copy number into a categorical variable. Previous results (Chapter 4) indicate $\log_2$-*rrn* is a quantitative proxy of life history variation, and so a regression analysis offers more statistical power than ANOVA to discern relationships between pPCA axes and *rrn* copy number (Cottingham *et al.*, 2005). I

performed phylogenetic linear regression using $\log_2$-*rrn* as a predictor for each of the first fifteen pPCA axes of the 1,167 genome dataset using both covariance- and correlation-based pPCA (Table 5.2 and Table 5.3). Regardless of the type of pPCA used or the dataset analyzed, at least 10 of the first 15 pPCA axes had significant regression slopes and these correlations could not be explained by shared ancestry.

	Orthologs			Modules		
pPCA Axis	Percent variance explained by axis	Slope	p	Percent variance explained by axis	Slope	p
1	2.54	-0.0115	**4.75E-09**	4.49	-0.0179	**<2.2E-16**
2	2.34	0.0363	**<2.2E-16**	3.53	-0.0028	**2.36E-06**
3	2.31	-0.0063	**5.93E-05**	2.50	-0.0067	**<2.2E-16**
4	1.98	0.0034	**0.01712**	2.25	0.0043	**2.30E-10**
5	1.64	-0.0019	0.1559	2.13	0.0018	**2.41E-04**
6	1.62	0.0019	0.2402	1.87	0.0014	**0.0030**
7	1.39	0.0003	0.8326	1.79	-0.0060	**7.27E-15**
8	1.33	0.0141	**5.05E-15**	1.71	0.0077	**<2.2E-16**
9	1.19	-0.0112	**9.48E-09**	1.59	0.0055	**<2.2E-16**
10	1.06	0.0251	**<2E-16**	1.47	0.0022	**2.89E-07**
11	1.03	0.0238	**<2.2E-16**	1.31	-0.0020	**4.00E-06**
12	0.95	0.0042	**0.0002**	1.23	0.0026	**0.0450**
13	0.93	-0.0375	**<2.2E-16**	1.16	-0.0056	**4.89E-16**
14	0.91	-0.0053	**7.62E-06**	1.06	0.0040	**0.0002**
15	0.85	0.0425	**<2.2E-16**	1.03	0.0010	0.4122

Table 5.2: Genome content is related to a bacterium's life history, correlation pPCA. Phylogenetic linear regression of the first 15 correlation-based pPCA axes of genome content as a function of $\log_2$-transformed *rrn* copy number for the 1,167 bacterial dataset.

	Orthologs			Modules		
pPCA Axis	Percent variance explained by axis	Slope	p	Percent variance explained by axis	Slope	p
1	5.78	0.1238	**0.0309**	8.80	0.4781	**<2E-16**
2	5.70	-0.3431	**4.48E-08**	6.59	-0.1920	**<2.2E-16**
3	4.88	-0.2709	**2.23E-07**	6.06	0.1754	**<2.2E-16**
4	4.02	0.5071	**<2E-16**	5.34	-0.1046	**4.32E-09**
5	3.76	0.4124	**<2.2E-16**	4.48	-0.1113	**8.47E-11**
6	3.68	0.5389	**<2E-16**	3.99	0.0402	**0.0101**
7	3.07	-0.1195	**0.0035**	3.50	-0.0497	**0.0086**
8	2.90	0.4040	**4.87E-14**	3.21	-0.0581	**0.0002**
9	2.59	0.2642	**7.64E-10**	3.01	-0.0253	0.0501
10	2.55	-0.0001	0.9969	2.72	-0.0274	0.0555
11	2.42	0.0396	0.3220	2.54	-0.2019	**<2.2E-16**
12	2.38	0.1336	**0.0006**	2.24	0.0142	0.3287
13	2.32	-0.1231	**0.0012**	2.09	0.0497	**0.0001**
14	2.12	0.0763	0.0509	2.01	-0.0118	0.3293
15	1.92	0.0219	0.4994	1.68	0.0022	0.8811

Table 5.3: Genome content is related to a bacterium's life history, covariance pPCA. Phylogenetic linear regression of the first 15 covariation-based pPCA axes of genome content as a function of log_2-transformed *rrn* copy number.

Genome features underlying life history variation are shared among bacterial phyla

Taken together, these results confirm the idea that a bacterium's place on the life history spectrum from copiotrophy is related to its genome content. These results also suggest that life history evolution's influence on genome content can lead distantly related species to possess similar genome content. If life history evolution truly drives changes in genome content than we expect two things: 1) when bacterial phyla are analyzed separately they should each display log_2-*rrn* correlations with genome content and 2) the orthologs and modules driving patterns within major phyla should be similar among phyla and the all bacteria analysis.

Over 75% of the species in this analysis can be attributed to three bacterial phyla, the Actinobacteria, Proteobacteria, and Firmicutes. These phyla each contain more than 100 species which span a wide range of *rrn* copy numbers, making them ideal candidates to test if genome

content universally varies with *rrn* copy number. Correlation-based pPCA was performed for each of these phylum datasets, and the first ten pPCA axes were regressed against $\log_2$-*rrn* copy

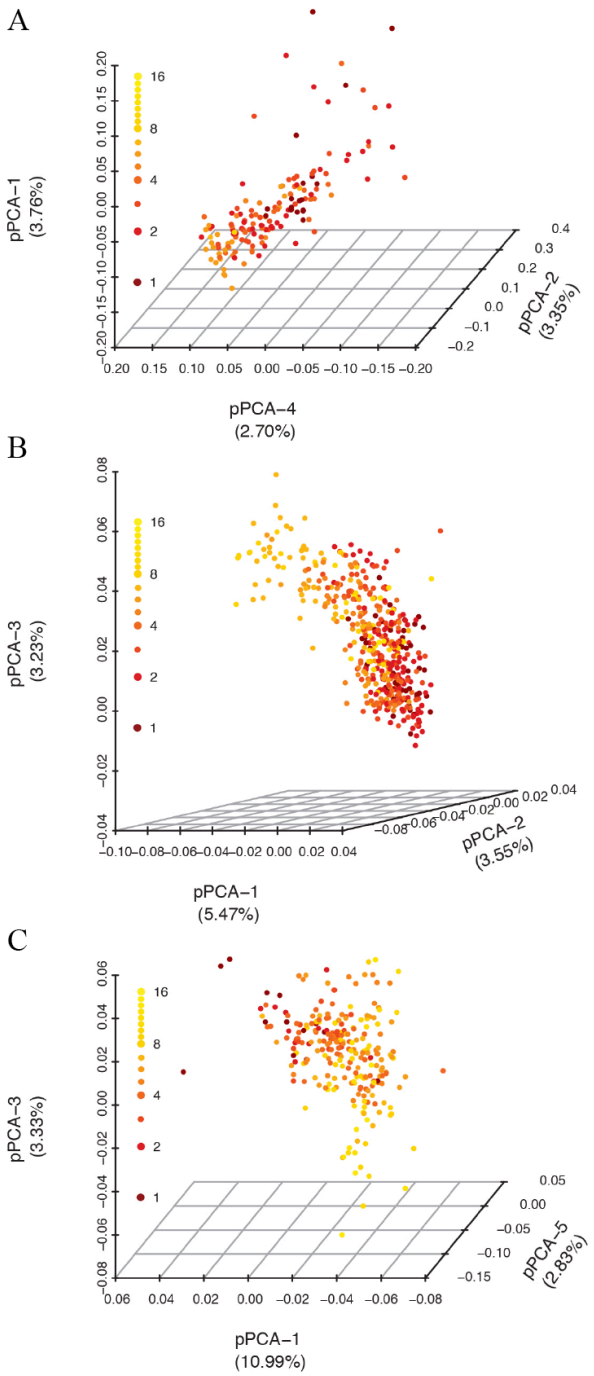


Figure 5.5: Module genome content is correlated with life history. Correlation-based pPCA of modules within three bacterial phyla. Axes which significantly correlate with $\log_2$-*rrn* are depicted for the Actinobacteria (A, N=156), Proteobacteria (B, N=484) and Firmicutes (C, N =253).

number. For both ortholog and module datasets, all three phyla had at least three of their first 10 pPCA axes significantly correlate with $\log_2$-*rrn* (Table 5.4, Table 5.5, Table 5.6, and Figure 5.5).

	Orthologs			Modules		
pPCA Axis	Percent variance explained by axis	Slope	p	Percent variance explained by axis	Slope	p
1	3.74	0.0057	0.2906	3.76	-0.0229	**4.13E-05**
2	3.22	-0.0779	**0.0379**	3.35	-0.0211	**0.0157**
3	3.06	-0.0349	**0.0014**	2.90	0.0090	**2.41E-06**
4	2.74	0.0395	**1.24E-08**	2.70	0.0178	**0.0073**
5	2.48	-0.0080	0.3582	2.55	-0.0039	0.5726
6	2.15	-0.0397	0.1212	2.44	-0.0013	0.7379
7	2.06	-0.0416	0.1428	2.30	-0.0088	0.1218
8	1.86	0.0289	0.1873	2.23	0.0077	**0.0097**
9	1.73	-0.0370	**0.0085**	2.22	-0.0054	0.4137
10	1.70	0.0245	0.3132	2.07	0.0082	0.0809

Table 5.4: Actinobacteria genome content is related to life history. Phylogenetic linear regression of the first 10 correlation based pPCA axes of genome content as a function of $\log_2$-transformed *rrn* copy number for the Actinobacteria dataset (N = 156).

	Orthologs			Modules		
pPCA Axis	Percent variance explained by axis	Slope	p	Percent variance explained by axis	Slope	p
1	3.53	-0.0070	**0.0244**	5.47	-0.0067	**1.27E-10**
2	3.23	0.0060	**0.0218**	3.55	-0.0065	**2.86E-15**
3	2.81	-0.0008	0.7431	3.23	0.0075	**8.23E-14**
4	2.39	-0.0049	0.0700	3.11	-0.0038	**1.76E-07**
5	2.34	0.0055	**0.0162**	2.67	-0.0140	**<2.2E-16**
6	1.92	-0.0121	**6.89E-06**	2.49	-0.0039	**1.45E-05**
7	1.61	0.0176	**1.09E-11**	2.35	-0.0027	**3.80E-05**
8	1.51	0.0421	**<2.2E-16**	1.97	-0.0185	**3.17E-09**
9	1.47	0.0257	**<2E-16**	1.86	0.0018	**0.0103**
10	1.33	-0.0353	**<2E-16**	1.72	0.0053	**2.57E-10**

Table 5.5: Proteobacteria genome content is related to life history. Phylogenetic linear regression of the first 10 correlation based pPCA axes of genome content as a function of $\log_2$-transformed *rrn* copy number for the Proteobacteria dataset (N=484).

	Orthologs			Modules		
pPCA Axis	Percent variance explained by axis	Slope	p	Percent variance explained by axis	Slope	p
1	5.62	-0.0428	**<2.2E-16**	10.99	-0.0179	**<2.2E-16**
2	3.52	0.0101	**0.0027**	4.21	0.0018	0.0869
3	2.17	-0.0008	0.7431	3.33	-0.0349	**0.0014**
4	1.97	0.0395	**1.24E-08**	2.96	-0.0026	**0.0190**
5	1.82	-0.0367	**7.66E-12**	2.83	-0.0104	**3.32E-08**
6	1.79	0.0084	0.2018	2.43	0.0045	**0.0146**
7	1.70	0.0087	0.1321	2.26	-0.0001	0.9491
8	1.51	0.0018	0.7484	2.05	0.0010	0.4825
9	1.49	-0.0086	0.1170	1.92	-0.0064	**5.56E-05**
10	1.42	<-0.001	0.9987	1.83	0.0019	0.2169

Table 5.6: Firmicutes genome content is related to life history. Phylogenetic linear regression of the first 10 correlation based pPCA axes of genome content as a function of log_2-transformed *rrn* copy number for the Firmicutes dataset (N=253).

To explore this further I examined which orthologs and modules loaded most strongly on pPCA axes correlated with log_2-*rrn* in the all bacteria and three major phyla analyses. To systematically evaluate all pPCA loadings from all analyses of orthologs and modules I performed the following procedure. First, I evaluated which of the first 10 pPCA axes in each analysis correlated significantly with log_2-*rrn*. I then extracted all of the loadings on any of the correlated axes and extracted the 100 loadings on any of these axes with the largest magnitude. Finally, I performed phylogenetic logistic regression on the original presence/absence data for each of these 100 modules or orthologs against log_2-*rrn* with all 1,167 genomes. The complete list of all orthologs and modules which have significant correlations with log_2-rrn among all 1,167 bacteria is provided (Table 5.7), and I will highlight some of the major findings across all of the analyses.

Both the modules and the orthologs loadings indicated secretion systems were significantly correlated with increased *rrn* copy number across all bacteria. Most intriguing was the type III secretion system, which had eleven orthologs with strong loadings in the all bacteria

covariance-based pPCA analysis. Each of these eleven orthologs had positive regression slope versus $\log_2$-*rrn* (11 orthologs, all $p<0.02$). Additionally, the type III secretion system module had strong loadings in the all bacteria and Proteobacteria correlation-based pPCA analyses and it was positively correlated with $\log_2$-*rrn* in the follow-up logistic regression analysis ($p<0.001$). The type III secretion system mediates host-association for both pathogenic and mutualistic bacteria through injecting effector proteins into the cytosol of the host (Silver *et al.*, 2007; Soto *et al.*, 2009; Sachs *et al.*, 2011). Other secretion systems which had strong pPCA loadings and positive regression slopes included the Sec (module, $p=0.005$), type I (module, $p < 0.001$), type VI secretion systems (10 orthologs, all $p<0.002$).

The biosynthesis and import of the compatible solute glycine betaine was also related to *rrn* copy number among all bacteria. Glycine betaine is the preferred compatible solute for most bacteria when coping with osmotic stress and it can be present in millimolar concentrations within bacterial cells (Csonka & Hanson, 1991). The module for the biosynthesis of glycine betaine from choline had a strong loading within Actinobacteria pPCA analysis, while the module for importing glycine betaine had strong loadings in the Proteobacteria and Firmicutes pPCA. Follow-up regression indicated that both modules were significantly correlated with $\log_2$-*rrn* among all 1,167 bacteria (transport, $p=0.003$; biosynthesis, $p=0.006$). It appears that copiotrophic bacteria are either better equipped to survive when solute concentrations rapidly change, or they may simply be more likely to experience osmotic stress than oligotrophic bacteria. The difference in loading patterns for the glycine betaine modules also suggests that preferences for producing or taking up glycine betaine may exist at the phylum level. However, further research is needed to test this hypothesis.

Discussion

Bacteria colonize a diverse range of habitats on Earth, where they are exposed to a wide range of environmental conditions. Over long evolutionary timescales, bacteria have adapted to a huge range of environmental pressures and I explored if three axes of environmental variation could explain the genome content of extant bacteria. Taken together, all approaches from this study demonstrate that genome content is intimately linked with a bacterium's niche and life history. *rrn* copy number is by far the strongest predictor of genome content that was analyzed in this study, which suggests that resource competition may be a fundamental player in the genome evolution of bacteria. Additionally, oxygen concentration may also play a role in determining the genome content of diverse bacterial species. My inability to detect any genome variation explained by temperature may reflect a poor sampling of the temperature ranges favored by bacteria. However, a great deal of imbalance in *rrn* copy number and oxygen requirement were also present in the dataset so I find this an unlikely, but still possible alternative hypothesis.

The orthologs and modules driving genome content to covary with *rrn* copy number broadly fit into our understanding of the biology of copiotrophic and oligotrophic bacteria. The finding that glycine betaine import and synthesis is more probable in copiotrophic bacteria fits well with the idea that oligotrophs are thought to be relatively passive in terms of their response to environmental change (Fegatella & Cavicchioli, 2000; Ostrowski *et al.*, 2001). Additionally, this finding may explain why many oligotrophic bacteria can not be cultured on rich medium in the laboratory (Koch, 2001). The strong relationship between secretion and copiotrophy has been previously observed in marine bacteria, and this may be linked to the preference for particle-association by copiotrophic bacteria or that (Lauro *et al.*, 2009). In chapter 4, I hypothesized that encoding a diverse set of phosphoenolpyruvate:carbohydrate phosphotransferase system (PTS)

transporters was a copiotrophic adaptation. While a correlation was present in the data it did not hold up statistically after incorporating phylogeny into the regression model. Many individual PTS transporters were more probable in high *rrn* genomes (Table 5.7), even when accounting for phylogeny, so there does appear to be some link between these transporters and copiotrophy.

Our understanding of the ecological and evolutionary forces acting on microbes in nature is still in its infancy. Genome sequencing is a promising tool which can help microbiologists integrate physiological insight with evolutionary perspective to better understand bacteria in their natural environments.

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
1167	correlation	M00018	2.017E-01	1.079E-02	17.1
1167	correlation	M00022	1.529E-01	3.303E-02	13.7
1167	correlation	M00036	1.247E+00	4.290E-02	5.5
1167	correlation	M00049	5.959E-01	1.356E-02	8.8
1167	correlation	M00061	4.644E-01	3.411E-03	15.5
1167	correlation	M00211	2.953E-01	2.690E-02	12.7
1167	correlation	M00213	7.195E-01	1.209E-02	7.8
1167	correlation	M00215	3.431E-01	2.023E-02	12.3
1167	correlation	M00217	9.869E-01	1.198E-02	12.7
1167	correlation	M00226	8.008E-01	1.904E-02	6.9
1167	correlation	M00268	1.263E+00	6.512E-07	23.6
1167	correlation	M00271	5.443E-01	4.019E-04	19.8
1167	correlation	M00279	4.724E-01	9.526E-03	10.7
1167	correlation	M00282	5.068E-01	2.046E-02	9.8
1167	correlation	M00287	9.724E-01	1.712E-02	5.2
1167	correlation	M00303	1.005E+00	1.111E-03	21.3
1167	correlation	M00305	1.413E+00	2.340E-03	27.4
1167	correlation	M00306	2.279E+00	1.653E-06	28.1
1167	correlation	M00332	9.116E-01	8.424E-04	10.0
1167	correlation	M00339	7.401E-01	9.635E-05	16.3
1167	correlation	M00362	3.044E-01	3.917E-02	9.7
1167	correlation	M00435	3.673E-01	2.895E-02	9.2
1167	correlation	M00439	2.945E-01	3.016E-02	11.9
1167	correlation	M00446	8.158E-01	2.687E-02	7.4
1167	correlation	M00450	6.097E-01	2.548E-02	8.2

Table 5.7: Genome features which load strongly on pPCA axes and correlate with *rrn*. Results of follow-up logistic regression performed on top 100 loadings on any pPCA axis correlated with $\log_2$-*rrn* across all analyses. Only modules or orthologs with phylogenetic logistic regression slopes significantly different from 0 ($p < 0.05$) when considered across all bacteria. Effect size is percent change in predicted probability from 1-15 *rrn*.

Table 5.7 (cont'd)

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
1167	correlation	M00454	2.150E-01	4.919E-02	12.8
1167	correlation	M00473	1.680E+00	2.122E-04	16.9
1167	correlation	M00477	1.059E+00	4.895E-03	8.2
1167	correlation	M00486	1.431E+00	3.280E-03	10.2
1167	correlation	M00495	1.201E+00	9.495E-03	8.4
1167	correlation	M00506	5.500E-01	2.493E-02	8.1
1167	correlation	M00549	2.072E-01	1.583E-02	17.9
1167	correlation	M00551	1.077E+00	1.187E-04	15.4
1167	correlation	M00568	1.825E+00	8.664E-04	17.4
1167	correlation	M00569	8.510E-01	2.310E-02	6.0
1167	correlation	M00582	1.887E-01	2.257E-03	17.3
1167	correlation	M00583	2.104E-01	2.497E-02	6.6
1167	correlation	M00632	2.749E-01	1.702E-02	15.4
1167	covariation	M00018	2.017E-01	1.079E-02	17.1
1167	covariation	M00022	1.529E-01	3.303E-02	13.7
1167	covariation	M00036	1.247E+00	4.290E-02	5.5
1167	covariation	M00049	5.959E-01	1.356E-02	8.8
1167	covariation	M00126	3.892E-01	8.169E-04	22.5
1167	covariation	M00178	7.652E-01	2.499E-17	60.8
1167	covariation	M00207	2.158E-01	9.633E-03	20.0
1167	covariation	M00213	7.195E-01	1.209E-02	7.8
1167	covariation	M00221	2.392E-01	1.741E-02	17.0
1167	covariation	M00223	2.520E-01	1.496E-02	16.6
1167	covariation	M00239	5.110E-01	5.598E-09	46.1
1167	covariation	M00268	1.263E+00	6.512E-07	23.6
1167	covariation	M00282	5.068E-01	2.046E-02	9.8
1167	covariation	M00299	2.280E-01	1.391E-02	18.5
1167	covariation	M00303	1.005E+00	1.111E-03	21.3
1167	covariation	M00305	1.413E+00	2.340E-03	27.4
1167	covariation	M00306	2.279E+00	1.653E-06	28.1
1167	covariation	M00332	9.116E-01	8.424E-04	10.0
1167	covariation	M00339	7.401E-01	9.635E-05	16.3
1167	covariation	M00362	3.044E-01	3.917E-02	9.7
1167	covariation	M00435	3.673E-01	2.895E-02	9.2
1167	covariation	M00436	5.229E-01	1.716E-04	22.2
1167	covariation	M00439	2.945E-01	3.016E-02	11.9
1167	covariation	M00446	8.158E-01	2.687E-02	7.4
1167	covariation	M00450	6.097E-01	2.548E-02	8.2
1167	covariation	M00473	1.680E+00	2.122E-04	16.9
1167	covariation	M00486	1.431E+00	3.280E-03	10.2
1167	covariation	M00495	1.201E+00	9.495E-03	8.4
1167	covariation	M00529	-6.667E-01	2.285E-02	-4.3
1167	covariation	M00530	3.256E-01	1.406E-02	13.8
1167	covariation	M00549	2.072E-01	1.583E-02	17.9
1167	covariation	M00568	1.825E+00	8.664E-04	17.4
1167	covariation	M00582	1.887E-01	2.257E-03	17.3
1167	covariation	M00632	2.749E-01	1.702E-02	15.4
1167	correlation	K00791	3.088E-01	1.040E-02	8.1
1167	correlation	K00928	7.722E-01	3.810E-03	8.9

Table 5.7 (cont'd)

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
1167	correlation	K01872	1.386E+00	4.536E-02	2.3
1167	correlation	K01939	7.068E-01	3.186E-02	6.7
1167	correlation	K01951	1.921E+00	2.939E-02	7.6
1167	correlation	K02867	1.127E+00	4.258E-02	1.7
1167	correlation	K02879	1.705E+00	4.270E-02	3.4
1167	correlation	K03046	9.420E-01	3.674E-02	3.8
1167	correlation	K03076	1.651E+00	1.674E-03	5.1
1167	correlation	K03217	9.316E-01	1.591E-02	3.3
1167	correlation	K03501	8.360E-01	1.131E-02	7.9
1167	correlation	K03687	1.699E+00	1.234E-03	7.0
1167	correlation	K03979	1.352E+00	1.027E-02	2.9
1167	correlation	K08227	7.210E-01	3.386E-02	8.7
1167	covariation	K01039	5.530E-01	2.586E-03	14.4
1167	covariation	K01040	5.492E-01	2.508E-03	14.5
1167	covariation	K01581	1.498E-01	1.196E-02	14.5
1167	covariation	K01643	4.084E-01	1.210E-02	12.0
1167	covariation	K01785	3.604E-01	5.457E-05	31.8
1167	covariation	K03168	1.123E+00	2.917E-03	9.5
1167	covariation	K03219	8.893E-01	1.837E-04	14.6
1167	covariation	K03220	7.203E-01	5.667E-04	12.1
1167	covariation	K03222	6.977E-01	1.233E-03	12.8
1167	covariation	K03223	5.571E-01	1.434E-02	8.0
1167	covariation	K03224	6.706E-01	1.480E-03	12.6
1167	covariation	K03225	7.012E-01	1.811E-03	12.0
1167	covariation	K03226	6.977E-01	1.233E-03	12.8
1167	covariation	K03227	6.867E-01	1.385E-03	11.9
1167	covariation	K03228	8.566E-01	1.772E-04	15.6
1167	covariation	K03229	7.715E-01	5.509E-04	14.0
1167	covariation	K03230	6.977E-01	1.233E-03	12.8
1167	covariation	K03838	1.520E+00	1.663E-02	6.2
1167	covariation	K07248	5.703E-01	2.951E-04	14.7
1167	covariation	K08154	1.780E+00	1.167E-03	15.4
1167	covariation	K08227	7.210E-01	3.386E-02	8.7
1167	covariation	K08682	8.934E-01	2.671E-02	7.6
1167	covariation	K09758	5.692E-01	1.011E-03	13.7
1167	covariation	K10117	2.444E-01	3.374E-02	13.3
1167	covariation	K11890	1.125E+00	2.873E-05	21.2
1167	covariation	K11895	9.130E-01	3.586E-03	13.2
1167	covariation	K11896	8.878E-01	4.026E-03	12.8
1167	covariation	K11897	9.555E-01	4.738E-04	16.1
1167	covariation	K11900	8.817E-01	1.953E-03	15.1
1167	covariation	K11901	8.906E-01	2.084E-03	14.9
1167	covariation	K11902	7.356E-01	2.915E-03	12.3
1167	covariation	K11903	9.109E-01	6.394E-04	17.6
1167	covariation	K11904	1.194E+00	1.370E-05	20.6
1167	covariation	K11907	9.388E-01	3.168E-04	18.8
1167	covariation	K12055	9.941E-01	1.832E-04	15.7
1167	covariation	K13069	1.324E+00	9.150E-04	12.9
1167	covariation	K13929	5.862E-01	1.943E-03	13.7

Table 5.7 (cont'd)

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
1167	covariation	K13930	6.371E-01	8.151E-04	14.8
1167	covariation	K13932	6.073E-01	1.603E-03	13.9
1167	covariation	K13933	5.658E-01	9.653E-03	10.0
1167	covariation	K13934	6.104E-01	1.436E-03	14.1
1167	covariation	K13935	1.119E+00	9.919E-06	18.3
1167	covariation	K15551	4.508E-01	8.922E-03	11.5
1167	covariation	K15552	4.217E-01	1.481E-02	10.6
1167	covariation	K15737	6.334E-01	5.789E-03	12.8
1167	covariation	K15790	4.865E-01	2.490E-02	8.9
Firmicutes	correlation	M00001	6.716E-01	4.736E-03	39.4
Firmicutes	correlation	M00002	1.009E+00	3.828E-03	20.7
Firmicutes	correlation	M00004	8.352E-01	3.769E-02	22.3
Firmicutes	correlation	M00018	5.160E-01	1.020E-02	38.6
Firmicutes	correlation	M00049	1.056E+00	6.044E-03	38.8
Firmicutes	correlation	M00050	9.386E-01	7.061E-04	33.8
Firmicutes	correlation	M00087	1.617E+00	1.595E-02	20.8
Firmicutes	correlation	M00096	6.340E-01	4.462E-02	25.1
Firmicutes	correlation	M00119	6.969E-01	1.874E-03	50.3
Firmicutes	correlation	M00157	8.212E-01	5.925E-04	41.9
Firmicutes	correlation	M00183	1.347E+00	8.912E-07	54.3
Firmicutes	correlation	M00188	8.113E-01	2.672E-04	62.4
Firmicutes	correlation	M00193	7.002E-01	2.738E-02	24.2
Firmicutes	correlation	M00208	5.457E-01	1.357E-02	35.8
Firmicutes	correlation	M00211	4.863E-01	9.081E-03	42.5
Firmicutes	correlation	M00219	1.034E+00	4.128E-02	20.6
Firmicutes	correlation	M00221	5.992E-01	2.108E-02	33.4
Firmicutes	correlation	M00222	4.928E-01	3.537E-02	21.9
Firmicutes	correlation	M00239	7.885E-01	7.556E-05	60.3
Firmicutes	correlation	M00298	-8.012E-01	1.797E-03	-35.6
Firmicutes	correlation	M00299	4.129E-01	2.773E-02	37.3
Firmicutes	correlation	M00307	5.047E-01	7.554E-03	44.4
Firmicutes	correlation	M00335	5.054E-01	4.482E-03	45.6
Firmicutes	correlation	M00360	2.255E+00	3.056E-06	70.7
Firmicutes	correlation	M00434	5.418E-01	9.319E-03	41.1
Firmicutes	correlation	M00439	4.340E-01	2.276E-02	38.7
Firmicutes	correlation	M00476	2.699E+00	1.134E-03	63.4
Firmicutes	correlation	M00479	2.039E+00	2.073E-02	25.6
Firmicutes	correlation	M00484	1.646E+00	7.292E-03	42.3
Firmicutes	correlation	M00495	6.254E-01	4.360E-02	25.2
Firmicutes	correlation	M00506	1.985E-01	4.540E-02	15.2
Firmicutes	correlation	M00549	3.586E-01	3.881E-02	32.8
Firmicutes	correlation	K00981	1.079E+00	8.291E-03	18.1
Firmicutes	correlation	K01839	6.327E-01	1.323E-02	36.8
Firmicutes	correlation	K01921	1.315E+00	1.297E-02	15.5
Firmicutes	correlation	K01939	1.291E+00	4.833E-02	34.4
Firmicutes	correlation	K01951	1.204E+00	8.211E-04	23.1
Firmicutes	correlation	K02528	1.097E+00	2.660E-02	12.7
Firmicutes	correlation	K02824	8.162E-01	3.657E-04	52.2
Firmicutes	correlation	K02860	9.464E-01	2.473E-02	29.1

Table 5.7 (cont'd)

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
Firmicutes	correlation	K02886	1.094E+00	4.702E-02	8.8
Firmicutes	correlation	K02899	1.035E+00	1.673E-02	12.6
Firmicutes	correlation	K03043	1.049E+00	1.286E-02	13.6
Firmicutes	correlation	K03046	1.953E+00	1.994E-04	30.6
Firmicutes	correlation	K03217	1.948E+00	4.330E-04	32.0
Firmicutes	correlation	K03431	1.120E+00	2.636E-02	10.4
Firmicutes	correlation	K03501	1.368E+00	4.425E-03	25.0
Firmicutes	correlation	K03589	1.516E+00	7.397E-03	56.8
Firmicutes	correlation	K03687	1.989E+00	1.450E-02	50.1
Firmicutes	correlation	K04078	8.653E-01	1.840E-02	15.1
Firmicutes	correlation	K04096	1.456E+00	3.506E-04	25.3
Firmicutes	correlation	K04567	2.181E+00	5.296E-05	43.4
Firmicutes	correlation	K06207	6.666E-01	3.337E-02	30.9
Firmicutes	correlation	K06309	5.662E-01	2.992E-02	24.3
Firmicutes	correlation	K06867	1.484E+00	8.472E-03	24.5
Firmicutes	correlation	K06949	1.100E+00	2.140E-03	45.2
Firmicutes	correlation	K07030	6.584E-01	7.430E-03	43.6
Firmicutes	correlation	K07462	1.165E+00	8.789E-04	27.9
Firmicutes	correlation	K09748	1.480E+00	5.632E-05	31.9
Firmicutes	correlation	K09787	1.328E+00	9.737E-04	30.6
Proteobacteria	correlation	M00027	9.551E-01	1.335E-03	29.3
Proteobacteria	correlation	M00053	4.299E-01	2.651E-02	24.8
Proteobacteria	correlation	M00136	1.297E+00	1.313E-02	15.1
Proteobacteria	correlation	M00150	7.919E-01	8.488E-04	38.5
Proteobacteria	correlation	M00176	5.272E-01	8.278E-04	43.2
Proteobacteria	correlation	M00208	4.486E-01	2.511E-03	39.5
Proteobacteria	correlation	M00213	8.852E-01	7.753E-04	31.6
Proteobacteria	correlation	M00217	1.337E+00	8.755E-03	28.4
Proteobacteria	correlation	M00226	9.367E-01	8.477E-03	21.6
Proteobacteria	correlation	M00238	4.374E-01	3.144E-03	36.5
Proteobacteria	correlation	M00266	2.208E+00	1.790E-03	37.1
Proteobacteria	correlation	M00268	2.302E+00	1.420E-04	45.8
Proteobacteria	correlation	M00270	2.360E+00	5.521E-03	33.3
Proteobacteria	correlation	M00271	1.290E+00	3.008E-03	23.5
Proteobacteria	correlation	M00275	2.809E+00	2.425E-03	72.6
Proteobacteria	correlation	M00277	2.596E+00	5.580E-05	56.4
Proteobacteria	correlation	M00282	6.074E-01	4.028E-02	14.4
Proteobacteria	correlation	M00303	2.654E+00	6.681E-07	68.7
Proteobacteria	correlation	M00305	7.321E-01	2.982E-05	34.5
Proteobacteria	correlation	M00324	1.385E+00	5.532E-09	78.3
Proteobacteria	correlation	M00332	1.269E+00	1.015E-05	37.7
Proteobacteria	correlation	M00339	8.487E-01	2.851E-06	43.7
Proteobacteria	correlation	M00446	9.573E-01	5.325E-04	44.3
Proteobacteria	correlation	M00450	7.533E-01	1.066E-03	38.1
Proteobacteria	correlation	M00471	5.529E-01	1.986E-02	23.8
Proteobacteria	correlation	M00473	2.377E+00	2.361E-05	59.2
Proteobacteria	correlation	M00477	8.610E-01	1.305E-02	16.7
Proteobacteria	correlation	M00486	1.747E+00	1.533E-03	32.0
Proteobacteria	correlation	M00491	8.513E-01	1.697E-02	17.2

Table 5.7 (cont'd)

Dataset	pPCA analysis	Genome feature	Slope	p	Effect size
Proteobacteria	correlation	M00504	5.710E-01	2.055E-02	23.4
Proteobacteria	correlation	M00538	-1.459E+00	1.740E-02	-5.4
Proteobacteria	correlation	M00545	5.893E-01	1.802E-02	13.8
Proteobacteria	correlation	M00551	9.224E-01	4.144E-04	30.3
Proteobacteria	correlation	M00568	9.479E-01	4.178E-04	32.0
Proteobacteria	correlation	M00569	-8.518E-01	1.688E-02	-9.0
Proteobacteria	correlation	M00577	-1.057E+00	3.977E-02	-4.5
Proteobacteria	correlation	M00582	7.866E-01	1.029E-02	17.1
Proteobacteria	correlation	M00605	5.128E-01	3.106E-02	20.6
Proteobacteria	correlation	M00631	1.730E+00	4.243E-04	35.5
Proteobacteria	correlation	M00632	1.025E+00	2.905E-04	44.7
Proteobacteria	correlation	K02763	6.074E-01	4.028E-02	14.4
Proteobacteria	correlation	K02764	6.237E-01	1.824E-02	17.4
Proteobacteria	correlation	K02765	6.237E-01	1.824E-02	17.4
Proteobacteria	correlation	K02840	1.276E+00	2.498E-02	27.2
Proteobacteria	correlation	K03838	1.760E+00	1.340E-02	18.6
Proteobacteria	correlation	K07862	7.369E-01	1.266E-04	47.9
Proteobacteria	correlation	K08682	2.408E+00	2.502E-03	47.6
Proteobacteria	correlation	K11744	1.368E+00	2.856E-02	13.5
Proteobacteria	correlation	K12151	1.568E+00	2.654E-02	15.0
Proteobacteria	correlation	K12290	1.481E+00	4.240E-02	12.7
Proteobacteria	correlation	K14392	1.846E+00	2.697E-03	55.4
Proteobacteria	correlation	K15983	1.168E+00	1.911E-02	12.7
Proteobacteria	correlation	K16044	9.648E-01	3.620E-02	14.6
Proteobacteria	correlation	K16050	1.168E+00	1.911E-02	12.7
Actinobacteria	correlation	M00087	-1.076E+00	2.816E-03	-44.8
Actinobacteria	correlation	M00135	1.040E+00	1.514E-02	54.7
Actinobacteria	correlation	M00157	1.094E+00	9.765E-04	56.7
Actinobacteria	correlation	M00196	6.329E-01	1.574E-02	54.1
Actinobacteria	correlation	M00233	1.470E+00	9.515E-06	76.3
Actinobacteria	correlation	M00238	1.167E+00	9.307E-04	52.7
Actinobacteria	correlation	M00239	8.007E-01	4.244E-03	52.1
Actinobacteria	correlation	M00258	1.275E+00	4.521E-02	13.3
Actinobacteria	correlation	M00302	1.148E+00	1.303E-02	60.7
Actinobacteria	correlation	M00436	8.276E-01	2.455E-03	57.4
Actinobacteria	correlation	M00546	1.858E+00	1.853E-02	73.1
Actinobacteria	correlation	M00555	8.523E-01	5.548E-03	65.0
Actinobacteria	correlation	M00603	9.893E-01	5.826E-03	66.0

REFERENCES

REFERENCES

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, *et al.* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**:208–211.

Cordero OX, Polz MF. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Micro* **12**:263–273.

Cottingham KL, Lennon JT, Brown BL. (2005). Knowing when to draw the line: designing more informative ecological experiments. *Frontiers in Ecology and the Environment* **3**:145–152.

Csonka LN, Hanson AD. (1991). Prokaryotic osmoregulation: genetics and physiology. *Annu Rev Microbiol* **45**:569–606.

Fegatella F, Cavicchioli R. (2000). Physiological responses to starvation in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. *Appl Environ Microbiol* **66**:2037–2044.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* 1–13.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. (2007). GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**:129–131.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**:D199–D205.

Koch A. (2001). Oligotrophs versus copiotrophs. *Bioessays* **23**:657–661.

Lauro F, McDougald D, Thomas T, Williams T, Egan S, Rice S, *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences* **106**:15527–15533.

Ligges U, Mächler M. (2002). Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software* **8**:1–20.

Luo H, Csuros M, Hughes AL, Moran MA. (2013). Evolution of divergent life history strategies in marine alphaproteobacteria. *mBio* **4**:–.

Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Pillay M, *et al.* (2013). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42**:D560–D567.

Mccutcheon JP, Moran NA. (2011). Extreme genome reduction in symbiotic bacteria. *Nature Publishing Group* **10**:13–26.

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, *et al.* (2008). Microbiology in the post-genomic era. *Nat Rev Micro*.

Morris RL, Schmidt TM. (2013). Shallow breathing: bacterial life at low O. 1–8.

Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer K-H, *et al.* (2011). Release LTPs104 of the All-Species Living Tree. *Syst Appl Microbiol* **34**:169–170.

Oksanen J, Blanchet FG, Kindt R, Pegendre L, Minchin PR, O'Hara RB, *et al.* (2014). vegan: Community Ecology Package. *CRANR-projectorg*. http://CRAN.R-project.org/package=vegan (Accessed July 23, 2015).

Ostrowski M, Cavicchioli R, Blaauw M, Gottschal JC. (2001). Specific growth rate plays a critical role in hydrogen peroxide resistance of the marine oligotrophic ultramicrobacterium sphingomonas alaskensis strain RB2256. *Appl Environ Microbiol* **67**:1292–1299.

Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**:289–290.

R Core Team. (2014). R: A Language and Environment for Statistical Computing. *wwwR-projectorg*. http://www.R-project.org/ (Accessed July 23, 2015).

Revell LJ. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**:217–223.

Revell LJ. (2009). Size-correction and principal components for interspecific comparative studies. *Evolution* **63**:3258–3268.

Sabath N, Ferrada E, Barve A, Wagner A. (2013). Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biology and Evolution* **5**:966–977.

Sachs JL, Essenberg CJ, Turcotte MM. (2011). New paradigms for the evolution ofbeneficial infections. *Trends in Ecology & Evolution* **26**:202–209.

Silver AC, Kikuchi Y, Fadl AA, Sha J. (2007). Interaction between innate immune cells and a bacterial type III secretion system in mutualistic and pathogenic associations.

Snel B, Bork P, Huynen MA. (1999). Genome phylogeny based on gene content. *Nat Genet* **21**:108–110.

Soto MJ, Domínguez-Ferreras A, Pérez-Mendoza D, Sanjuán J, Olivares J. (2009). Mutualism versus pathogenesis: the give-and-take in plant-bacteria interactions. *Cellular Microbiology* **11**:381–388.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. (2014). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*.

Suen G, Goldman BS, Welch RD. (2007). Predicting Prokaryotic Ecological Niches Using Genome Sequence Analysis Butler, G (ed). *PLoS ONE* **2**:e743.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, *et al.* (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA* **102**:13950–13955.

Tung Ho LS, Ane C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic biology* **63**:397–408.

Vieira-Silva S, Rocha EPC. (2010). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet* **6**:e1000808.

Wickham H. (2009). ggplot2: elegant graphics for data analysis. Springer: New York.

Wu H, Moore E. (2010). Association analysis of the general environmental conditions and prokaryotes' gene distributions in various functional groups. *Genomics* **96**:27–38.

Zaneveld JR, Lozupone C, Gordon JI, Knight R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research* **38**:3869–3879.