UNIVERSITY OF CALIFORNIA,
IRVINE

Ecological and evolutionary processes contributing to the formation of bacterial populations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Alexander Bennett Chase

Dissertation Committee
Professor Jennifer B.H. Martiny, Chair
Professor Brandon S. Gaut
Professor Adam C. Martiny

2018

DEDICATION

To

my mother and father,

Steve and Lisa

my number one fans and THE best support system in the world

And to

my friends and family,

for reminding me there exists another world outside the lab

And, finally, to

my incredible wife and main editor,

Liz

without you, I wouldn't be where I am today

you continue to inspire and motivate me everyday

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation, and the work included within, would not have been possible without the love and support from my friends and family, the guidance from my colleagues, and the always necessary “meetings” in Microbial Group. To my wife, Liz: This has been an insane journey filled with excitement, frustration, panic, but most importantly love. We went through so much to get here (including getting married!) and could not have done this without your constant support, motivation, and especially your patience. My parents, Steve and Lisa, and my sister, Carissa: you have always given me your support and instilled the confidence in myself to chase my passions. You have provided me with a loving home and family that is second to none. My advisor and mentor, Jennifer Martiny: you are an inspiration and, without you, none of this literally would have been possible! You are an incredible mentor and friend as you always allowed me to pursue my scientific curiosities and fostered my growth both as a person and as a scientist. Brandon and Adam: thank you for constantly suffering through my committee meetings and providing me with thoughtful feedback to improve everything in here. And finally, thank you to the lab members of the Martiny lab, past and present, for creating a fun and exciting work place, obviously including bowling and karaoke nights! There were a lot of people that contributed to this work and it was truly a privilege to work with you all. This was a long and hard journey and I could not have finished without each of you – thank you!

I would like to acknowledge and thank the funding sources who made my dissertation research possible. Fellowship support was provided by the department of Ecology and Evolutionary Biology at the University of California, Irvine, the US Department of Education Graduate Assistance in Areas of National Need (GAANN) fellowship, and the US Department of Energy Office of Science Graduate Student Research (SCGSR) fellowship. Research support was provided by grants from the NSF Division of Environmental Biology, and the DOE Office of Science Biological and Environmental Research program.

Chapter 1 was published with the permission of CSIRO Publishing. The text of this chapter is a reprint of the material as it appears in the journal Microbiology Australia. Chapter 2 was published with the permission of Frontiers Media. The text of this chapter is a reprint of the material as it appears in the journal Frontiers in Microbiology. Chapter 3 was published with the permission of the American Society for Microbiology. The text of this chapter is a reprint of the material as it appears in the journal mBio. Chapter 4 was published with the permission of Elsevier Publishing Group. The text of this chapter is a reprint of the material as it appears in the journal Environmental Microbiology.

CURRICULUM VITAE

Alexander B. Chase

Department of Ecology and Evolutionary Biology
3300 Biological Sciences III
University of California
Irvine, CA 92697

W: 949.824.0532
E: abchase at uci dot edu

Research Interests

Biogeographical patterns and the role microbes play in regulating and/or promoting ecosystem processes
Evolutionary forces that shape environmental microbial population dynamics using (meta)genomics

Education

2014 – 2018	Doctor of Philosophy (Ph.D.) Biological Sciences	*University of California – Irvine (UCI)*
2014 – 2017	Master of Science (M.S.) Biological Sciences	*University of California – Irvine*
2014 – 2015	Data Science Certificate	*University of California – Irvine*
2008 – 2012	Bachelor of Science (B.S.)	*University of California – Los Angeles (UCLA)*

Professional Appointments

2019 – *Postdoctoral Researcher*
Center for Biotechnology and Medicine, Scripps Institution of Oceanography
Principle Investigator: Paul Jensen

2018 *Research Affiliate*
Earth and Environmental Sciences, Lawrence Berkeley National Laboratory
Principle Investigators: Eoin Brodie and Ulas Karaoz

2014 – 2018 *Research Assistant*
Department of Ecology and Evolutionary Biology, UCI
Principle Investigator: Jennifer B.H. Martiny

NSF DEB-1457160. Collaborative Research: Manipulating microbial community composition along a climate gradient to analyze microbial function and response

2012 – 2013 *Research Associate III*
Department of Ecology and Evolutionary Biology, UCLA
Principle Investigators: Stephen P. Hubbell and Brant C. Faircloth

NSF DEB-1136626. Dimensions: Testing the potential of pathogenic fungi to control the diversity, distribution, and abundance of tree species in a Neotropical forest community
NSF DEB-1146440. Collaborative Research: Genealogical reconstruction, foliar chemistry, and community dynamics in a Neotropical rain forest landscape

2009 – 2010 *Research Assistant*
National Center for Research on Evaluation, Standards, and Student Testing, UCLA

Publications

7 T Gallagher, J Phan, A Oliver, **AB Chase**, W England, S Wandro, C Hendrickson, S Riedel, and K Whiteson. 2018. Cystic fibrosis-associated *Stenotrophomonas maltophilia* strain-specific adaptations and responses to pH. Journal of Bacteriology (In Revision).

6 MBN Albright, **AB Chase**, and JBH Martiny. 2018. Experimental evidence that stochasticity contributes to bacterial composition and functioning in a decomposer community. mBio (In Review).

5 **AB Chase**, Z Gomez-Lunar, AE Lopez^, J Li, SD Allison, AC Martiny, JBH Martiny. 2018. Emergence of soil bacterial ecotypes along a climate gradient. Environmental Microbiology 20:4112-4126.

4 **AB Chase** and JBH Martiny. 2018. The importance of resolving biogeographic patterns of microbial microdiversity. Microbiology Australia 39:5-8.

3 **AB Chase**, U Karaoz, EL Brodie, Z Gomez-Lunar, AC Martiny, and JBH Martiny. 2017. Microdiversity of an abundant terrestrial bacterium encompasses extensive variation in ecologically relevant traits. mBio 8:e01809-17.

2 **AB Chase**, P Arevalo, MF Polz, R Berlemont, and JBH Martiny. 2016. Evidence for ecological flexibility in the cosmopolitan genus *Curtobacterium*. Frontiers in Microbiology 7:1874.

1 MB Nelson, **AB Chase**, JBH Martiny, *et al*. 2016. The Microbial Olympics 2016. Nature Microbiology 1:16122.

^ Undergraduate Mentee

Book Chapters

AB Chase, KL Dolan, DJ Mohamed, and JBH Martiny. 2017. Microbial Biodiversity. Encyclopedia of Biodiversity. Elsevier Press. Amsterdam.

Honors & Awards

2019 – 2020	Scripps Institution of Oceanography Postdoctoral Scholar Fellowship ($129,000)
2018	U.S. Department of Energy Student Sponsorship ($400)
2018	William D. Redfield Graduate Fellowship Award, UCI ($1,000)
2018	Office of Science Graduate Student Research (SCGSR) Award, U.S. Department of Energy ($11,500)
2016 – 2018*	Graduate Assistance in Areas of National Need (GAANN) Fellowship, U.S. Department of Education ($29,202)
2018	GAANN Research Award ($2,000)
2017	GAANN Travel Award ($1,865)
2015 – 2016*	Ecology & Evolutionary Biology Department Travel Grant, UCI ($500)
2014 – 2019*	Ecology & Evolutionary Biology Department Fellowship, UCI ($2,000)
2012	*Gamma Sigma Alpha* inductee
2012	*Order of Omega* inductee
2008	Assistance League Scholarship for Excellence in Community Service

* Grant/Fellowship renewed

Presentations

AB Chase*, Z Gomez-Lunar, AE Lopez^, J Li, SD Allison, AC Martiny, and JBH Martiny. 2018. Differential distributions of bacterial ecotypes along a climate gradient. Lake Arrowhead Microbial Genomics Conference. Lake Arrowhead, CA. Poster Presentation. **Award Winning Poster**.

AB Chase*, Z Gomez-Lunar, AE Lopez^, J Li, SD Allison, AC Martiny, and JBH Martiny. 2018. Emergence of soil bacterial ecotypes along a climate gradient. ISME. Leipzig, Germany. Oral and Poster Presentation.

Z Gomez-Lunar*, **AB Chase**, JBH Martiny, and AC Martiny. 2018. Drought conditions decrease respiration rates of most abundant bacterial taxa in the southern California grassland litter. ASM Microbe. Atlanta, GA.

AB Chase* and JBH Martiny. 2018. Microbial Biogeography: From Macro- to Micro-. Ecology and Evolutionary Biology Graduate Student Symposium. Irvine, CA. Oral Presentation.

AB Chase* and JBH Martiny. 2018. Biogeographic patterns of bacterial microdiversity. Ecology and Evolutionary Biology Graduate Recruitment. Irvine, CA. Oral Presentation.

Z Gomez-Lunar*, **AB Chase**, JBH Martiny, and AC Martiny. 2017. Leaf litter decomposition and drought tolerance in *Curtobacterium*. Gordon Research Conferences: Applied and Environmental Microbiology. South Hadley, MA.

AE Lopez^*, **AB Chase**, Z Gomez-Lunar, AC Martiny, and JBH Martiny. 2017. Physiological characterization of microdiversity within *Curtobacterium*. UCI Undergraduate Research Symposium. Irvine, CA.

AB Chase*, AC Martiny, U Karaoz, EL Brodie, and JBH Martiny. 2017. Clade-specific responses of *Curtobacterium* in a leaf litter community. Department of Energy Genomics Science Program. Washington D.C. Poster Presentation.

AB Chase*, U Karaoz, EL Brodie, and JBH Martiny. 2016. Clade-specific responses of an abundant bacterium in a leaf litter community. Symposium on Recent Advancements in Data Science. Irvine, CA. Poster Presentation.

AB Chase*, U Karaoz, EL Brodie, and JBH Martiny. 2016. Analysis of an abundant bacterial genus in a leaf litter community. Lake Arrowhead Microbial Genomics Conference. Lake Arrowhead, CA. Poster Presentation.

AB Chase* and JBH Martiny. 2016. A case for a model organism in terrestrial decomposition. Ecology and Evolutionary Biology Graduate Student Symposium. Irvine, CA. Oral Presentation.

AB Chase*. 2015. *Curtobacterium*: a model organism for terrestrial decomposition. Organization for Tropical Studies. San José, Costa Rica. Oral Presentation.

* Presenting author ^ Undergraduate Mentee

Advanced Coursework and Workshops

UC San Diego

2018 Microbiome Research and Human Health

San Diego Supercomputer Center Summer Institute at UC San Diego

2017 High Performance Computing and Data Science

UCI Department of Ecology and Evolutionary Biology

2018 EEB Education Part IV – Educational Design
2017 EEB Education Part III – CURE Course Design
2017 EEB Education Part II – Effective Mentoring
2016 EEB Education Part I – Scientific Learning and Course Design
2016 Special Topics in Evolution
2016 Writing Proposals
2015 Quantitative Statistical Methods
2015 Special Topics in Ecology
2014 Physiological Plant Ecology

UCI Data Science Certificate Program

2016 Introduction to Next Generation Sequencing Analysis
2015 Advanced R Topics
2015 Predictive Modeling with Python
2015 Unix, Git, and Python Software Carpentry Workshop

Miscellaneous Coursework

2017	MiGA: Microbial Genomes Atlas	Georgia Institute of Technology	Invited NSF Workshop
2016	Bayesian Inference Phylogenetics	University of California, Irvine	Graduate Short Course
2015	Tropical Biology: Ecology	Organization for Tropical Studies	Graduate Course

Teaching

2018 Advanced Molecular Biology (UCI – Department of Molecular Biology and Biochemistry)
Implemented a Course-based Undergraduate Research Experience (CURE) to understand human microbiomes with an emphasis on how diet can affect human gut microbiomes

2016 Intro to Linux (UCI – Data Science Initiative)
Co-led a data science training course to best exploit the bash shell for both interactive work and batch jobs, moving and simple manipulation of data, as well as short introductions to programming in bash, Perl, and R

2016 Phylogenetics for Ecologists (UCI – Department of Ecology and Evolutionary Biology)
Co-led a departmental seminar for faculty and graduate students about the utilization of the bash shell for multi-sequence alignments and phylogenetic tree construction, as well as statistical inference in R

Teaching Assistant

2018	Biological Science 94	Organisms to Ecosystem	Robin Bush
2016	Biological Science 93 (Admin)	DNA to Organisms	Kim Green
2016	Biological Science E142W	Philosophy of Biology	Francisco J. Ayala
2015	Biological Science 93	DNA to Organisms	Marcelo Wood
2015	Biological Science E189	Environmental Ethics	Peter Bowler
2014	Biological Science E150	Conservation Biology	Steve Weller

Student Mentoring and Outreach

Undergraduate Mentees

2016 – 2017 Alberto Lopez (Independent Research Student)
Advised through the UCI Minority Science Program – Maximizing Access to Research Careers (MARC) and the Undergraduate Research Training Program to conduct microbiology work and computational analyses
Currently a Ph.D. student at Northwestern University Feinberg School of Medicine

Outreach

2017 AAUW Tech Trek (UCI)
Led a hands-on microbiology workshop for the American Association of University Women (AAUW) Tech Trek camp, which focuses on promoting science, technology, engineering, and math (STEM) for middle school girls to broaden interest and accessibility

2013 – 2014 Robison Elementary After-School Coordinator (Tucson Unified School District, Arizona)
Coordinated after school programs for elementary school students, including academic activities geared towards promoting STEM and inter-school athletic competitions

2011 – 2012 Project Literacy (UCLA)
Participated in weekly meetings with underrepresented elementary school students from Compton, CA to promote reading and writing in students who are falling below their recommended grade reading levels

2011 – 2012 Watts Tutorial Program (UCLA)
Participated in an on-campus tutoring and mentoring program for children of the Watts and William Mead government housing projects to stimulate interest in higher education for underrepresented students

ABSTRACT OF THE DISSERTATION

Ecological and evolutionary processes contributing to the emergence of bacterial populations

by

Alexander Bennett Chase

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2018

Professor Jennifer B.H. Martiny, Chair

Despite our ability to characterize diverse microbial communities, we currently lack the capability to identify the mechanisms affecting a given taxon's response to environmental conditions and its functional consequence. This problem partly stems from the common practice of characterizing microbial communities using conserved marker genes, such as the 16S rRNA region, which mask ecologically-relevant genetic and phenotypic variation. For example, most microbial studies that target the 16S rRNA region, cluster similar sequences into operational taxonomic units (OTUs) that represents millions of years of evolutionary history. My work has explored how much genetic variation may exist within these OTU designations and identified how genetic variation corresponds to phenotypic variation (overview in Chapter 1). Overall, I have focused on the ecological and evolutionary mechanisms driving environmental niche partitioning and speciation within a bacterial taxon to link genotypic variation to functional roles.

Utilizing an abundant soil bacterium, *Curtobacterium*, I have demonstrated that isolates within this *Actinobacteria* genus harbored extensive genomic diversity within a single OTU (Chapter 2) that reflected large phenotypic variation even within a single field site. Using extensive isolation efforts with the integration of genomic and metagenomic data, I have identified distinct genomic clusters that would otherwise be masked by traditional microbial analyses (Chapter 3). Further, this vast genomic diversity corresponded to distinct phenotypes denoting fine-scale niche partitioning and the emergence of bacterial ecological populations along a regional climate gradient (Chapter 4). And while ecological variation may drive large-scale geographic distributions, the evolutionary mechanisms contributing to microbial speciation and diversity are less understood. As such, I analyzed the genetic diversity within a single ecotype to identify could identify distinct populations (groups of individuals recombining more with one another than among groups) in a heterogeneous soil system (Chapter 5). In conclusion, my research has provided evidence that ecological and evolutionary processes both contribute to the response of bacteria to environmental conditions.

CHAPTER 1

The importance of resolving biogeographic patterns of microbial microdiversity

For centuries, ecologists have used biogeographic patterns to test the processes governing the assembly and maintenance of plant and animal communities (Lomolino *et al.*, 2006). Similarly, evolutionary biologists have used historical biogeography (e.g., phylogeography) to understand the importance of geological events as barriers to dispersal that shape species distributions (Avise, 2000). As the field of microbial biogeography initially developed, the utilization of highly conserved marker genes, such as the 16S ribosomal RNA gene, stimulated investigations into the biogeographic patterns of the microbial community as a whole. Here, we propose that we should now consider the biogeographic patterns of microdiversity, the fine-scale genetic diversity observed within a traditional ribosomal-based taxonomic unit.

Biogeography investigates how ecological and evolutionary processes influence the distribution of biodiversity and the structure of contemporary communities (Wiens and Donoghue, 2004). Historically, biogeographic patterns of plants and animals are studied at the species level and describe large-scale patterns of species' distributions. In contrast, the vast majority of microbial biogeographic studies investigate patterns by sampling the entire community at broad taxonomic designations. Typically, these studies define operational taxonomic units (OTUs) using a highly conserved ribosomal marker gene, usually the 16S rRNA gene for bacteria and archaea. However, the decision of which genetic region to target, and in particular the genetic resolution of that region, can influence the biogeographic patterns

observed (Cho and Tiedje, 2000). While these conserved regions can capture a large breadth of the microbial community, these regions, by their very nature, limit the detection of finer-scale genetic variation. By resolving diversity within the OTU designation, we can detect ecological and evolutionary processes occurring at this fine taxonomic scale that might otherwise be overlooked.

What OTU-based biogeography can and can't tell us

It is now well established that microbial communities assayed by traditional OTU designations display distinct biogeographic patterns over space and time. These patterns have been identified in environments ranging from marine (Giovannoni *et al.*, 1996; Garcia-Martinez and Rodriguez-Valera, 2000), to terrestrial (Fierer and Jackson, 2006), and to human-associated systems (Consortium, 2012). Combined with abiotic and biotic data from the sampled environment, such patterns can provide initial hypotheses about the ecological processes shaping microbial community assemblages (Hanson *et al.*, 2012). Thousands of microbial studies now demonstrate that OTU-based patterns primarily reflect the importance of selection of environmental conditions based on correlations between microbial composition and the environment (Figure 1A). These patterns indicate that OTUs comprising each microbial community vary in their ability to tolerate various abiotic and biotic conditions, suggesting partitioning of environmental resources and niche spaces among taxa in the community.

While environmental variables explain much of the variation in microbial composition, many studies also find that some variation is correlated to the geographic distances between communities (Dumbrell *et al.*, 2009; Hanson *et al.*, 2012). This observation can be illustrated with a distance-decay curve, or a negative correlation between the similarity in microbial

composition with geographic distance between pairwise samples (J. B. H. Martiny *et al.*, 2006) (Figure 1B). If this negative relationship holds after accounting for environmental variation, then the pattern suggests that ecological drift, caused by stochastic fluctuations in demographic patterns, contributes to variation in community composition (Hubbell, 2001; McGill, Maurer, *et al.*, 2006). Further, since ecological drift depends on restricted dispersal, the pattern gives insight into the degree of dispersal limitation between the sampled communities. A caveat to such studies is that it is impossible to completely account for environmental variation, and the environment is spatially autocorrelated. However, such OTU-based studies suggest that the ecological processes of both environmental selection and ecological drift contribute to biogeographic patterns at this broad genetic resolution (J. B. H. Martiny *et al.*, 2006).

In contrast to ecological processes, biogeographic patterns of OTU-based analyses are unlikely to detect patterns shaped by evolutionary processes. This limitation is due to the broad resolution of conserved marker genes. Variation in these genetic regions capture relatively distant evolutionary divergences, especially when clustered at 97% sequence similarity. Indeed, a 3% sequence divergence in the 16S rRNA gene, the most common level of OTU clustering, represents roughly 150 million years of evolutionary history (Ochman *et al.*, 1999), or before the origin of modern birds (Pereira and Baker, 2006). In other words, biogeographic patterns for birds at this taxonomic level would mask all diversification within the group. Similarly, the use of such conserved marker genes for microbes will generally preclude detecting biogeographic patterns emerging from evolutionary processes, such as endemism and niche conservatism, as observed for macroorganisms assessed at the species or population level.

What is microbial microdiversity

Studies based on 16S sequences have been instrumental in identifying ecological patterns and their underlying processes at relatively broad genetic resolutions. However, it is increasingly clear that there is extensive genetic diversity within 16S-based OTUs, so-called microdiversity, in environmental habitats (Moore *et al.*, 1998; Acinas *et al.*, 2004; Jaspers and Overmann, 2004). For example, a natural population of the bacterioplankton *Vibrio splendidus* contained >1000 distinct genotypes, even when clustered at >99% 16S rRNA sequence similarity (Thompson *et al.*, 2005). Based on their very nature, conserved marker genes lack the variability to resolve fine-scale diversity within an OTU. Even with the implementation of exact sequence variants (ESVs), the 16S rRNA gene simply cannot resolve the fine-scale variation among closely related microbial lineages (Lan *et al.*, 2016). Thus, different approaches are needed to investigate the biogeographic patterns of this vast genetic diversity.

Beyond identifying genetic microdiversity, a key question is whether this genetic variation is phenotypically relevant (Larkin and Martiny, 2017). Investigations into microdiverse marine bacterial taxa suggest that they vary in physiological traits including preferences for particular abiotic conditions (Jaspers and Overmann, 2004; Johnson, Zinser, Coe, Mcnulty, *et al.*, 2006). Further, some of this trait variation within OTU-based taxa appears to be phylogenetically conserved within microdiverse clades (Martiny *et al.*, 2009, 2015), although resolving the phylogeny of such closely related strains is often difficult with 16S sequences (Figure 2A). Instead, taxon-specific marker genes or, ideally full genome sequences, can often resolve microdiverse clades and reveal which traits are shared among particular phylogenetic clades (Figure 2B). For example, an analysis of strain diversity of an abundant leaf litter

bacterium, *Curtobacterium*, exhibited extensive variation in the degree of polymeric carbohydrate degradation and temperature preference among microdiverse clades (Chase *et al.*, 2017). Thus, more resolved genetic and physiological studies can help to establish the phylogenetic distribution of traits.

What biogeographic patterns of microdiversity can tell us

The presence of trait variation among microdiverse clades suggests that microdiversity will exhibit distinctive biogeographic patterns. If this trait variation corresponds to different ecological preferences, then the environment should select for specific clades under variable conditions. Indeed, different bacterial ecotypes, or ecological populations (Cohan, 2001), have repeatedly been shown to vary in their spatial distribution. Thus, closely-related clades appear to partition niche space in the environment that would normally be masked at the OTU level (Figure 2C). For example, at the OTU level, the globally distributed cyanobacterium, *Prochlorococcus*, shows a broad preference for low-nutrient and warmer waters (Flombaum *et al.*, 2013). However, microdiverse clades of *Prochlorococcus* exhibit distinct spatial distribution patterns shaped by additional environmental factors, including light availability and temperature (Moore *et al.*, 1998; A. C. Martiny *et al.*, 2006; Johnson, Zinser, Coe, Mcnulty, *et al.*, 2006). Thus, biogeographic patterns of microdiversity can elucidate the importance of key environmental parameters governing niche differentiation that may not be identifiable at the OTU designation.

Perhaps even more importantly, a focus on microdiversity can reveal evolutionary processes that would otherwise be masked at a broader genetic resolution. Thus far, few environmental studies have targeted microbial diversity at a fine enough scale to investigate

how evolutionary mechanisms, such as mutation and genetic drift, can lead to differential biogeographic patterns (Martiny *et al.*, 2009; Andam *et al.*, 2016). Those examples that do exist find evidence for evolutionary processes contributing to spatial patterns. In one such example, reduced dispersal between hot spring populations of the archaeon thermophile *Sulfolobus*, restricted gene flow to allow diversification to occur among geographic regions (Whitaker *et al.*, 2003; Cadillo-Quiroz *et al.*, 2012). Similarly in terrestrial soils, dispersal limitation at regional spatial scales structures bacterial populations of *Streptomyces* along a latitudinal gradient (Choudoir *et al.*, 2016). With the increased availability of computational tools to study population genomics (Shapiro *et al.*, 2012) and the incorporation of gene flow networks (Hehemann *et al.*, 2016), we expect that more studies will consider the spatial distribution of microdiversity. Such studies are likely to illuminate the effects of evolutionary processes on microbial diversity in the environment, including the presence of biogeographic barriers and the degree of microbial endemism (Polz *et al.*, 2013) (Figure 2D).

Conclusions

Future progress in microbial biogeography necessitates moving beyond the OTU designation. While OTU-based studies will continue to play an important role in microbial biogeography, an intensified focus on finer-genetic diversity will uncover thus-far unidentified ecological and evolutionary patterns. However, these studies will require targeted sampling of particular microbial taxa rather than the entire community. Generally, this effort will require moving beyond targeting the 16S rRNA gene; even ESVs of this region will not be able to distinguish microbial populations at a fine enough genetic scale. And while extensive shotgun metagenomic and targeted amplicon sampling can reveal co-occurrence of novel microdiversity

associated with distinct environmental conditions (Woebken *et al.*, 2008; Brown *et al.*, 2012; Malmstrom *et al.*, 2013), these studies are dependent on the interpretation of genomic potential for ecological diversity. Therefore, there is still a need to link the genomic variation to functional traits that will define ecotypes. The return to isolation-based studies to gather relevant genetic and physiological information will better inform environmental metagenomic studies investigating microbial microdiversity. By expanding the focus to microbial microdiversity and implementing targeted environmental studies, we can better understand the ecological and evolutionary processes generating microbial biogeographic patterns as macroecologists have done for decades.

Figures

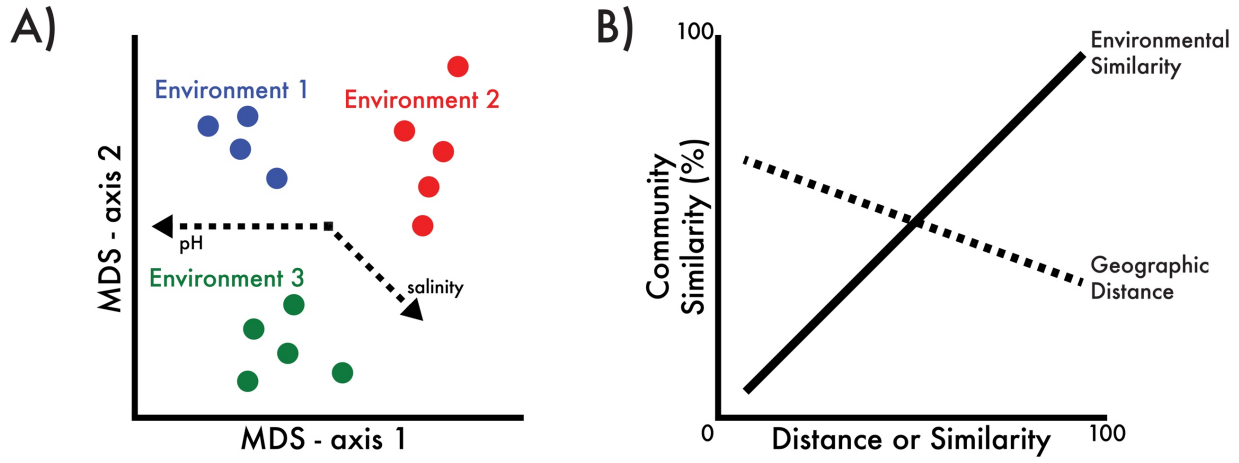


Figure 1.1 Hypothetical community analyses from OTU-based studies. **A)** An ordination plot depicting community composition across three environments with the main environmental factors driving compositional differences indicated with dashed arrows. Each point represents a sampled microbial community, with points closer to one another indicating higher similarity in community composition. **B)** Community similarity among a collection of samples is often positively correlated to environmental similarity (grey line) and negatively correlated with geographic distance (black dashed line, also called a distance-decay curve). The influence of strong environmental selection on the community is reflected in the positive correlation with increasing environmental similarity, while the influence of ecological drift is reflected in the negative correlation with increasing geographic distance between samples.

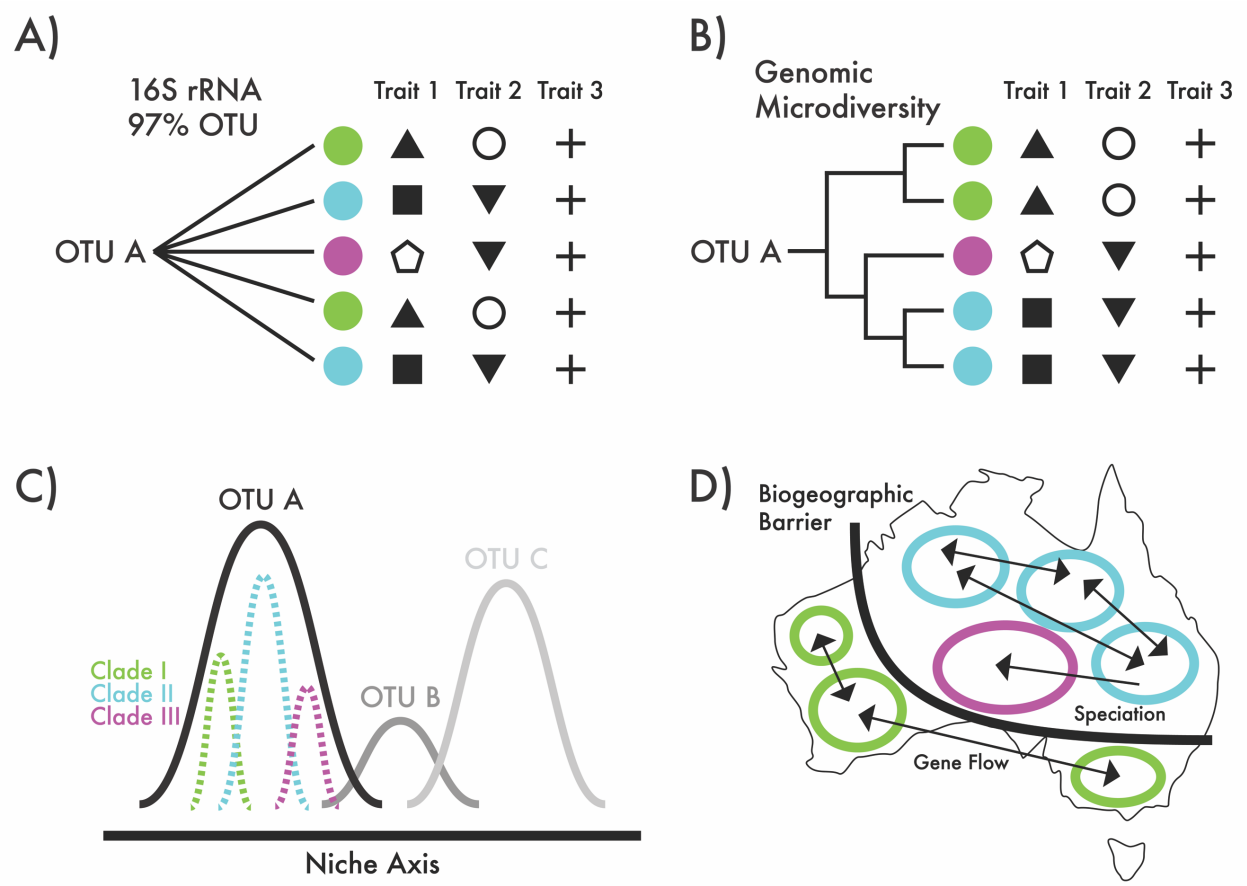


Figure 1.2 Detection of ecological and evolutionary processes within OTU A with microdiverse Clades I (green), II (blue), and III (pink). **A)** The 16S rRNA gene often cannot resolve phylogenetic relationships within a 16S-based OTU and, subsequently, the distribution of traits among clades. **B)** Genomic sequences or multi-locus sequence analyses (MLSA) of marker genes can resolve phylogenetic relationships at a finer-scale revealing, in this hypothetical example, that strains within clades share more similar traits. **C)** Trait variation within microdiverse taxa can promote resource partitioning in the environment leading to fine-scale niche differentiation among clades (represented in colored dashed lines) that would otherwise be masked at the OTU level (black line represents the total niche for OTU A). **D)** Investigating genetic differentiation within OTUs is more likely to reveal dispersal limitation (measured by gene flow between clade populations) and the presence of biogeographic barriers that contribute to microbial diversification. In this hypothetical example, black arrows represent gene flow between populations of microdiverse clades, where limited gene flow (no arrows connecting green with the blue and pink populations) suggests the presence of biogeographic barriers.

CHAPTER 2

Evidence for ecological flexibility in the cosmopolitan genus *Curtobacterium*

ABSTRACT

Assigning ecological roles to bacterial taxa remains imperative to understanding how microbial communities will respond to changing environmental conditions. Here we analyze the genus *Curtobacterium* as it was found to be the most abundant taxon in a leaf litter community in southern California. Traditional characterization of this taxon predominantly associates it as the causal pathogen in the agricultural crops of dry beans. Therefore, we seek to conduct a broad investigation into this genus to ask whether its high abundance in our soil system is in accordance with its role as a plant pathogen or if alternative ecological roles are needed. By collating >24,000 16S rRNA sequences with 120 genomes across the Microbacteriaceae family, we show that *Curtobacterium* has a global distribution with a predominant presence in soil ecosystems globally. Moreover, this genus harbors a high diversity of genomic potential for the degradation of carbohydrates, specifically with regards to structural polysaccharides. We conclude that *Curtobacterium* may be responsible for the degradation of organic matter within litter communities.

INTRODUCTION

Traditional ecological characterization of microorganisms often narrowly defines their roles in terms of interspecies interactions. Such limited classification of interactions ignores the dynamic alterations of life cycles indicative of microorganisms in changing environmental conditions (Redman *et al.*, 2001; Kogel *et al.*, 2006; Newton *et al.*, 2010). Depending on the environment, microbes can transition from symbiont to pathogen (Johnson *et al.*, 1997) or drastically alter their life history strategy altogether. For instance, endophytic fungi transition to decomposers after the leaves fall off its host plant (Osono, 2006; Korkama-Rajala *et al.*, 2008). Such flexibility in ecological roles may also explain why *Curtobacterium*, a bacterial genus traditionally viewed as a plant pathogen (Hsieh *et al.*, 2005), was recently found to be the dominant bacterium in the leaf litter of a Mediterranean-like grassland community (Matulich *et al.*, 2015).

Members of the *Curtobacterium* genus are Gram-positive, obligately aerobic chemoorganotrophs in the family Microbacteriaceae, phylum Actinobacteria (Evtushenko and Takeuchi, 2006). The habitat of *Curtobacterium* is described mainly in association with plants and especially, the phyllosphere (Komagata *et al.*, 1965; Behrendt, Ulrich, Schumann, Naumann, and K.-I. Suzuki, 2002). Indeed, most studies investigating *Curtobacterium* focus on its role as an economically important plant pathogen (Huang *et al.*, 2009; Osdaghi, Taghavi, *et al.*, 2015). The best-studied pathovar, *C. flaccumfaciens* pv. *flaccumfaciens*, is the causal agent of bacterial wilt in dry beans worldwide with reports on five continents (Wood and Easdown, 1990; Harveson *et al.*, 2006; Soares *et al.*, 2013; Osdaghi, Pakdaman Sardrood, *et al.*, 2015). The

disease harbors a high degree of genetic and phenotypic diversity (Hedges, 1926; Conner *et al.*, 2008) even within a single host (Agarkova *et al.*, 2012).

Although economically important, *C. flaccumfaciens* is the only species of *Curtobacterium* associated with plant pathogenesis (Young *et al.*, 1996), and there is evidence the other *Curtobacterium* species perform other ecological roles. For instance, isolates have been identified as endophytic symbionts (Sturz *et al.*, 1997, 1999; Elbeltagy *et al.*, 2000; Araújo *et al.*, 2001; Bulgari *et al.*, 2009). Similar to other beneficial endophytes (Benhamou *et al.*, 2000; Taghavi *et al.*, 2009), *Curtobacterium* can elicit plant defense responses (Bulgari *et al.*, 2011) and reduce disease symptoms (Lacava *et al.*, 2007). The genus has also been found to associate with roots and promote plant growth (Sturz *et al.*, 1997). Even the presence of *C. flaccumfaciens* in the rhizosphere induced a systematic resistance in cucumber plants to pathogens (Raupach and Kloepper, 1998) and promoted plant growth (Raupach and Kloepper, 2000). *Curtobacterium* can also be found in soil (Ohya *et al.*, 1986; Aizawa *et al.*, 2007; Kim *et al.*, 2008) with an ability to persist on plant debris (Júnior Silva *et al.*, 2012), although as a non-spore forming bacterium, the genus might be assumed to be a poor survivor in soil (Vidaver, 1982).

Our previous work in a Mediterranean-like grassland community revealed that a *Curtobacterium* taxon (defined by $\geq$ 97% similarity in 16S rRNA sequence) was the most abundant bacterium in leaf litter, the top layer of soil. The leaf litter community at this site is dominated by bacteria with a bacteria to fungi biomass ratio up to 30:1 (Alster *et al.*, 2013). The community is highly diverse, but uneven; three phyla (Actinobacteria, Bacteroidetes, and Proteobacteria) made up 95% of total bacterial abundance (Matulich *et al.*, 2015). Further

analysis revealed that *Curtobacterium* constituted ~18% of 16S rRNA sequences amplified directly from 177 litter samples over a two-year period (Matulich *et al.*, 2015). This high abundance was further supported by sequenced metagenomes from the same grassland. These samples suggested that >8% of the reads fall within Microbacteriaceae (Berlemont *et al.*, 2014), most likely an underestimate due to lack of representation of *Curtobacterium* in genomic databases.

Given its dominance in grassland litter, this current study investigates the potential for *Curtobacterium* to play ecological roles other than a plant pathogen, and in particular, as a decomposer. We asked: 1) What is the geographic and habitat distribution of the genus? 2) Is the phylogenetic diversity of *Curtobacterium* related to its habitat distribution? and 3) What is the genus' genomic potential to degrade recalcitrant carbohydrates? To address these questions, we isolated and sequenced 14 *Curtobacterium* strains from grassland litter. Then, we combined our genome sequences with publicly-available sequences from a variety of habitats and locations, collating >24,000 Microbacteriaceae 16S rRNA sequences. Finally, we investigated the genomic diversity of *Curtobacterium* with regards to its ability to degrade carbohydrates, an important attribute for litter decomposition. We searched for glycoside hydrolases (GHs), enzymes that target specific glycosidic bonds of carbohydrates (including cellulose and xylan in plant cell walls). We conclude that the genus *Curtobacterium* is cosmopolitan in terrestrial ecosystems and may be, at an intrageneric level, involved in a variety of ecological roles including decomposition of organic matter.

MATERIALS AND METHODS

Geographic Distribution

To investigate the geographic extent of *Curtobacterium*, we searched for *Curtobacterium* sequences within the open reference dataset of the Earth Microbiome Project (EMP) (Gilbert *et al.*, 2014). We obtained 41 unique *Curtobacterium* OTUs with metadata from 14,096 uploaded samples.

To gather additional *Curtobacterium* sequences, we used BLAST to search for sequences similar to eight *Curtobacterium* 16S rRNA gene sequences from the GreenGenes "Core Set" database (DeSantis *et al.*, 2006) against the GenBank nr database (Benson *et al.*, 2008). Additional sequences were identified using the keyword search: "Microbacteriaceae Curtobacterium 16S ribosomal RNA gene". After removing redundant entries and 16S rRNA sequences that could not be identified to the genus level, 11,484 unique sequences remained.

We extracted metadata from either corresponding GenBank files, the EMP 10k merged mapping file, or manually reviewed the published literature to identify the isolation source and location of all retrieved *Curtobacterium* sequences. Each sequence was assigned to one of seven ecosystems: animal microbiome, aquatic, artificial, atmosphere, human microbiome, ice, or terrestrial. Terrestrial samples were further divided into six categories: plant, plant roots, plant seeds, rock, sediment, and soil.

The geographic distribution of the EMP and GenBank sequences were plotted using the R library 'rworldmap' (South, 2011). For samples with minimal location data (mainly from the GenBank dataset), we used a publicly available dataset from Google Developers[1] to assign

[1] https://developers.google.com/public-data/docs/canonical/countries_csv

approximate longitude and latitude coordinates based on the state, providence, and/or country of origin.

Phylogenetic Diversity

To establish a robust phylogenetic distribution, we downloaded 16S rRNA gene sequences from the SILVA SSU r123 database (Pruesse *et al.*, 2007) on August 6, 2015. Sequences were obtained using SILVA's assigned taxonomy, yielding 1,519 *Curtobacterium* sequences and 24,835 Microbacteriaceae sequences. Due to variability in taxonomic nomenclature by various databases, we confirmed all taxonomic assignments of all downloaded sequences. First, we assigned taxonomy with QIIME v1.6 (Caporaso *et al.*, 2010) using the UCLUST consensus taxonomy assigner (Edgar, 2010) against the GreenGenes reference database (May 2013 revision; (DeSantis *et al.*, 2006)). Next, we compared these taxonomic assignments to those assigned using the RDP Classifier (Wang *et al.*, 2007). After removing sequences incorrectly assigned to *Curtobacterium* and/or Microbacteriaceae and other low quality sequences (<80% identity, <700 bp), 12,469 sequences remained.

To select a subset of this sequence diversity for phylogenetic analysis, we clustered the filtered sequences and the sequences of our litter isolates (see below) using QIIME v1.9. We defined OTUs at 97% identity with UCLUST using the optimal flag for OTU picking, and selected representative sequences for each OTU. The representative sequences were assigned a taxonomic designation at the genus level using a combination of UCLUST, BLAST, and the RDP Classifier. Specifically, genera designations for the representative sequences were only assigned when at least two of the aforementioned taxonomic designations matched at the genus level. We aligned the sequences using the Infernal Alignment Tool (Nawrocki *et al.*, 2009). Gaps

common in >90% of aligned sequences were manually removed, resulting in a 1900 bp alignment. OTU representative sequences that contained >25% gap regions were also removed. As a result, the sequences obtained from the EMP database were too short (~100-250 bp; mean size=134 bp) to integrate in the phylogeny with the full 16S rRNA gene obtained from other datasets. A maximum likelihood tree with 100 bootstrap replications was constructed with RAxML v8.0, using the GTR + Gamma distribution model (Stamatakis, 2014). The tree was visualized using the Interactive Tree of Life (iTOL; (Letunic and Bork, 2006)).

The pipeline above was modified slightly to investigate the phylogenetic diversity within the *Curtobacterium* genus. This analysis incorporated all available 16S rRNA genes (n=1532) from GenBank, SILVA, and litter isolates assigned to *Curtobacterium*. OTUs were clustered at 99% similarity to provide finer taxonomic resolution and included a sister genus, *Frigoribacterium*, as an outgroup.

Genomic Analysis of Litter Isolates

Isolation and Identification of Litter Isolates

Bacteria from litter were isolated from two grassland global change experiments. Isolates from the Loma Ridge Global Change Experiment (LRGCE) (in Irvine, California, USA [33° 44' N, 117° 42' W]; (Potts *et al.*, 2012)) were previously identified and presented in (Mouginot *et al.*, 2014). Briefly, leaf litter particles were suspended in saline and inoculated onto nutrient-limited media plates made from Loma Ridge litter leachate and incubated at room temperature. For this study, additional strains were isolated from the Boston-Area Climate Experiment (BACE) [42° 23' N, 71° 12' W] (Tharayil *et al.*, 2011) using the same protocol on

Boston litter leachate media. Individual colonies were streaked onto LB plates three times to ensure clonal isolation.

To identify *Curtobacterium* isolates, the 16S rRNA gene was PCR amplified and sequenced. Individual colonies were boiled for 1 min in 50 μL of sterile dH_2O prior to PCR amplification. Next, 3.0 μL of the boiled bacterial colony was added to the PCR cocktail containing 0.3 μL of Taq polymerase (5 units/μL), 15.0 μL of Premix F (Epicentre, Madison, WI), and 50 μM of each primer in a final volume of 30 μL. We amplified 1500 bp of the 16S rRNA gene using the pA (5'-AGAGTTTGATCCTGGCTCAG-3') and pH (5'-AAGGAGGTGATCCAGCCGCA-3') primers (Edwards *et al.*, 1989). Forward and reverse strands were trimmed and merged using Geneious v6.1 (Drummond *et al.*, 2011) under the default parameters. Isolate identity was tentatively assigned using the best-identified match with the blastn alignment (Altschul *et al.*, 1997) within GenBank. In total, 34 Microbacteriaceae isolates were identified, including 17 *Curtobacterium* isolates.

Whole Genome Analysis

This Whole Genome Shotgun project including the genome sequences of 14 *Curtobacterium*, 1 *Frigoribacterium*, and 1 *Plantibacter* isolates deposited at GenBank under BioProject PRJNA342146 with accessions MJGI00000000-MJGX00000000. Paired-end 100 bp x 100 bp whole genome sequencing libraries with a mean gap size of 400 bp were prepared from genomic DNA using the Nextera DNA Library Preparation Kit (Illumina Inc., San Diego, CA, USA). Genomes were sequenced on an Illumina HiSeq 2500 apparatus (Illumina Inc., San Diego, CA, USA) at the Whitehead Institute Genome Technology Core (Cambridge, MA). After quality

trimming and removal of short (<30 bp) reads, an initial de novo assembly was performed in CLC Genomics Workbench (CLC Bio, Cambridge, MA, USA) using the default parameters.

Genomes (fully assembled and whole genome shotgun assembly) belonging to the Microbacteriaceae were retrieved from the Pathosystems Resource Integration Center (PATRIC) database (Wattam *et al.*, 2014). To annotate these downloaded genomes and our isolate genomes, we first assigned open reading frames (ORFs) sequences as called by Prodigal v2.6 (Hyatt *et al.*, 2010). Genomic ORFs were searched against the Pfam database (Finn *et al.*, 2016) for the presence of protein families using HMMer (Johnson *et al.*, 2010). We identified the GH families as in (Berlemont and Martiny, 2013) and compiled the number of occurrences of each GH family in each genome. To create a phylogeny of the whole genome sequences, the 16S rRNA region of each genome was predicted using Barrnap[2]. The resulting sequences were used for phylogenetic reconstruction as described above.

RESULTS

Geographic Distribution of Curtobacterium

We isolated 17 *Curtobacterium* strains from two invasive grassland sites. Although similar in their vegetation, LRGCE and BACE sites are 4130 km apart across the North American continent. Yet, from these sites, *Curtobacterium* strains comprised 10% and 15% of culturable isolates in LRGCE and BACE, respectively. Beyond these two terrestrial sites, data collected from a wide array of studies and isolation sources reveal that *Curtobacterium* is an abundant and globally distributed taxon. In total, we obtained 3360 16S rRNA sequences with corresponding metadata from GenBank and the EMP databases. The genus was found on all continents,

[2] http://www.vicbioinformatics.com/software.barrnap.shtml

ranging from the Arctic to the Antarctic (Figure 1). The majority of sequences were isolated from North America (61.6%), while there was a lack of representation in the Southern hemisphere, most likely due to sampling effort. Australia, South America, Africa, and Antarctica accounted for only 15.3% of all sequences.

Curtobacterium has been identified in all designated ecosystems, including animal microbiome, aquatic, artificial, atmosphere, human microbiome, ice, and terrestrial (Supplementary Table 1). The human and animal microbiome comprised 26.9% and 12.9% of all obtained *Curtobacterium* sequences, respectively. *Curtobacterium* sequences from humans were comprised almost exclusively of samples originating from skin, while those from animals were primarily collected from the gut. Most *Curtobacterium* sequences (32.6%) from the EMP dataset were from human microbiome samples, reflecting the emphasis on humans in this dataset. In contrast, only 10.8% of *Curtobacterium* sequences retrieved from GenBank were associated with the human microbiome. After excluding human microbiome samples, over 63% of all sequences originated from terrestrial ecosystems. Specifically, 14% of all sequences were extracted from a plant source and 21% from soil. Sequences from the GenBank database revealed a stronger association with 70.1% of sequences being classified into a terrestrial ecosystem (Supplementary Table 1). Terrestrial samples from the GenBank database included 58.9% from plants and 28.4% from soil.

Phylogenetic Diversity

The Microbacteriaceae sequences clustered into 971 OTUs at a 97% similarity level. Considering only OTUs with more than 10 sequence representatives, the remaining 183 OTUs represented 19 genera (Figure 2). The 10 *Curtobacterium* OTUs form a well-supported

(bootstrap support of 89%) monophyletic clade. Their closest relatives belong to the *Rathayibacter* and *Pseudoclavibacter* genera. The 17 *Curtobacterium* litter isolates from the Loma Ridge and Boston sites clustered together into five OTUs. Two *Curtobacterium* OTUs contained only one litter isolate despite being in the top 25 of the most abundant OTUs in the SILVA database.

To examine *Curtobacterium* diversity at a finer genetic resolution, we clustered the 1074 total sequences retrieved from GenBank and SILVA with our 22 isolates at a 99% similarity level. This yielded 100 *Curtobacterium* and 7 *Frigoribacterium* OTUs, a sister genus. Excluding singletons, the remaining 52 OTUs represented 1014 *Curtobacterium* sequences with 764 of those sequences containing metadata originating from GenBank entries. Of these sequences, 582 (74%) sequences were isolated from a terrestrial ecosystem. Due to some OTUs containing sequence with low metadata, distribution of ecosystem preference across phylogeny was not possible. However, there were OTUs detected solely in one ecosystem (e.g., OTU 25 was only found in terrestrial ecosystems), while others OTUs were detected in a variety of ecosystems (e.g., OTU 38 contained all seven assigned ecosystems). At the level of 99% sequence similarity, most (10 out of 18) of the litter isolates clustered into one abundant OTU (86). This abundantly universal OTU contains over 202 sequences isolated from all seven assigned ecosystems.

Genomic Potential for Carbohydrate Degradation

Full genomes were used to compare the genomic diversity of glycoside hydrolases within Microbacteriaceae. We included 14 *Curtobacterium*, 1 *Frigoribacterium*, and 1 *Plantibacter* isolates from our leaf litter sites at LRGCE and BACE. The *Curtobacterium*

assemblies produced an average genome size of 3.76 Mbp in an average of 78 contigs (mean maximum contig length of 582,567 bp), with an average GC content of 70.47% (Table 1).

Combining these genomes with the 104 publicly available genomes retrieved from the PATRIC database reveal that strains within Microbacteriaceae contain many diverse GH families. Across the 120 genomes, we identified 7,355 potential glycoside hydrolases (GHs) and carbohydrate binding modules (CBMs) representing 63 GH/CBM families (Supplementary Table 2). The most common and ubiquitous families belonged to those targeting starch (GH13, CBM48) and oligosaccharides (GH1, 2, and 3). These GH families were present in most genomes with 92.5% of the genomes containing at least one copy of GH13. GHs that targeted more recalcitrant carbohydrates such as fructan, dextran, mixed polysaccharides, animal polysaccharides, plant polysaccharides, cellulose, chitin, and xylan were also detected in a variety of genomes across Microbacteriaceae, albeit at a lower frequency (Table 2).

GH content is highly variable across genera. Some genera were not able to process any structural polysaccharides (cellulose, chitin, or xylan) and were constrained to the targeting of oligosaccharides and starch (see Table 2). Others, like *Pseudoclavibacter*, lack any of the identified GH families that process simpler substrates such as starch, and, presumably, are only capable of processing more complex carbohydrates. A few genera have the genomic potential to digest all identified substrates. Specifically, *Curtobacterium* can digest all identified substrates at a frequency almost double the family average, particularly with regard to structural polysaccharides. Individual strains with the potential to breakdown and digest all three structural polysaccharides appear to be restricted within the genera *Curtobacterium*

(N=11 genomes; including 8 litter isolates), *Clavibacter* (N=6 genomes), and *Microbacterium* (N=6 genomes).

The average richness of GH families present in a Microbacteriaceae genome was 19.3 GH/CBM families (Table 2). However, GH/CBM richness varied widely across genomes; a *Leucobacter* genome contained only 1 GH family while one *Microbacterium* species, *Microbacterium sp. SUBG005* (accession number JNNT00000000), had 35 GH families. The litter isolates belonging to *Curtobacterium* had an above average richness of 27.2 GH families with a range of 19 to 31 GH families. Further, most genomes harbored multiple copies of each protein family. For example, a *Microbacterium* genome had as many as 24 copies of the GH13 family. Due to the multiple GH copies, genomes varied in the total number of GHs present (mean number of GHs=61.3), ranging from 3 GH proteins in a *Leucobacter* strain to 135 GH proteins in an *Agromyces* strain. On average, the *Curtobacterium* litter isolates encoded 82.1 GH proteins, almost 1.5 times the family average (Figure 4).

We examined the potential for each individual genome to target multiple polysaccharides. Almost all genomes within the family have the potential to process oligosaccharides or starch with the exception of 2 genomes, a *Leucobacter* and *Pseudoclavibacter* strain. Further, a majority of the genomes (103 genomes or 85.8%) within Microbacteriaceae are capable of processing at least one structural polysaccharide. Specifically, the frequency to be able to target cellulose, chitin, and xylan occurred in 64.2, 64.2, and 29.2% of the genomes, respectively.

DISCUSSION

In this study, we present the first global survey of *Curtobacterium* and show that it is ubiquitous in a variety of ecosystems (Figure 1) although it is most abundant in terrestrial ecosystems, and a majority of sequences are associated with plants and soil. This observation is in accordance with past studies of *Curtobacterium* that attribute its habitat to plants and the related phyllosphere (Komagata *et al.*, 1965; Behrendt, Ulrich, Schumann, Naumann, and K. I. Suzuki, 2002). However, *Curtobacterium* is primarily known as a plant pathogen and yet, the highest proportion of *Curtobacterium* strains resided in soil systems, suggesting that this genus may be capable of reproducing in soil.

We also provide a well-supported phylogeny of all known Microbacteriaceae genera. We built upon previous Microbacteriaceae phylogenetic analyses (see (Evtushenko and Takeuchi, 2006)) to incorporate all available Microbacteriaceae 16S rRNA sequences, providing the most comprehensive phylogenetic analysis of Microbacteriaceae to date (Figure 2). To explore diversity within *Curtobacterium*, we constructed a genus-specific tree to investigate the possibility of clade-specific habitat preference. Due to differences in sequencing platforms and targeted regions of the 16S rRNA gene, there may be habitat specialization at finer clade levels than we are able to differentiate here. In particular, the shorter sequenced reads (e.g., from the EMP dataset) are limited in their phylogenetic resolution and cannot resolve intrageneric patterns. Further, many GenBank sequences lacked metadata altogether or were limited in their details to allow for finer habitat designations (e.g., which part of the plant or the layer of soil from which a strain was isolated). Although, we did not detect any clade-specific patterns of

habitat preferences, most clades contained a majority of plant and soil isolated sequences (Figure 3), indicating that the genus as a whole may be adapted to plant or soil habitats.

Curtobacterium falls within the Actinobacteria phylum, which is known to play a crucial role in the recycling of organic material by decomposition and humus formation (Goodfellow and Williams, 1983). This characterization is supported by a comprehensive analysis into the distribution of GHs across all bacteria, which showed that Actinobacteria has the highest genomic potential for being cellulose degraders (Berlemont and Martiny, 2015). Therefore, we concentrated on these GH proteins, as they are responsible for the breakdown of large carbohydrates that may prove advantageous in decomposition of plant debris. For instance, an increase in diversity and abundance of GHs with the potential for cellulose utilization generally corresponds to better cellulose degradation (Fontes and Gilbert, 2010; Wilson, 2011; Berlemont and Martiny, 2015). Previously, *Curtobacterium* isolates collected from a neutral garden soil were shown to rapidly degrade cellulose fibers (Lednická *et al.*, 2000). Indeed, our results provide a genomic underpin for *Curtobacterium* to be a degrader. The genus has an elevated richness and abundance of GHs relative to other Microbacteriaceae genera. While there is large variation within the family with respect to GH richness and substrate degradation, *Curtobacterium* is one of only three genera with the potential ability to target all identified carbohydrate substrates. Moreover, out of these three genera, *Curtobacterium* has the highest abundance of GHs, suggesting an increased ability to utilize and degrade a wide range of carbohydrates. This variability in carbon usage within *Curtobacterium* suggests that alternative, intrageneric ecological roles have yet to be identified.

We conclude that *Curtobacterium* may be a dominant player in the functional breakdown of dead organic material in leaf litter communities based on its dominance in two grassland litter microbial communities, its high representation in soils, and its genomic potential for being a degrader. This work supports previous studies that show that *Curtobacterium* has the capability to survive on litter (Júnior Silva *et al.*, 2012) and thrive as a cellulytic bacterium (Lednická *et al.*, 2000). The conclusion also aligns with culture work that finds that coryneform bacterium, such as *Curtobacterium*, are in high abundance on grasses (Behrendt, Ulrich, Schumann, Naumann, and K. I. Suzuki, 2002). Despite the focus in the literature on its role as a crop plant pathogen, future research into the contribution of *Curtobacterium* to the recycling of nutrients in terrestrial ecosystems warrants further attention.

ACKNOWLEDGMENTS

We would like to thank Adam Martiny and Travis Huxman for their guidance and helpful comments on earlier revisions. We thank Kristin Dolan for use of the Loma Ridge isolates, Jeff Dukes for supplying litter from BACE, and Sean Gibbons and Jack Gilbert for help with the open reference database for the EMP. We also thank Richard Puxty, Claudia Weihe, and Michaeline Nelson for their input and assistance with labwork and computational methods. Funding was provided by the US Department of Energy, Office of Science, Office of Biological and Environmental Research (BER), under Award Number DE-PS02-09ER09-25 and by the U.S. National Science Foundation (DEB-1457160) to JBHM. This work was supported by the U.S. Department of Energy (DE-SC0008743) to MFP.

Tables and Figures

Table 2.1 General characteristics of the litter isolates. Strains originating from LRGCE are labeled as MMLR, while strains from BACE labeled as MCBA.

Genome ID	Taxonomy	# of contigs	Length (bp)	% GC	# of ORFs	Richness of GH Families	# of GHs
MCBA15_001	*Curtobacterium*	137	3808678	70.12	3940	28	81
MMLR14_002	*Curtobacterium*	40	3634776	71.39	4013	28	79
MCBA15_003	*Curtobacterium*	75	3648432	71.07	3743	29	90
MCBA15_004	*Curtobacterium*	91	3772244	69.38	3633	19	62
MCBA15_005	*Curtobacterium*	26	3601746	72.01	3941	27	77
MMLR14_006	*Curtobacterium*	87	3768639	69.80	3742	29	90
MCBA15_007	*Curtobacterium*	75	4023578	70.40	3888	28	77
MCBA15_008	*Curtobacterium*	41	3649950	71.44	4198	31	103
MCBA15_009	*Curtobacterium*	23	3476500	70.57	3759	29	90
MMLR14_010	*Curtobacterium*	112	3902159	70.56	4034	31	93
MMLR14_011	*Plantibacter*	104	4089281	69.16	4422	25	105
MCBA15_012	*Curtobacterium*	83	3616790	71.04	3638	19	62
MCBA15_013	*Curtobacterium*	86	3948212	69.88	4167	28	83
MMLR14_014	*Curtobacterium*	139	3822836	69.91	4017	27	78
MCBA15_016	*Curtobacterium*	77	3947873	69.07	4103	28	84
MCBA15_019	*Frigoribacterium*	27	3783004	70.04	3633	24	60

Table 2.2 Breakdown by genus of the distribution of GHs by targeted substrate.

SUBSTRATE		*Agreia*	*Agrococcus*	*Agromyces*	*Candidatus Aquiluna*	*Candidatus Rhodoluna*	*Clavibacter*	*Cryobacterium*	*Cryocola*	*Curtobacterium*	*Frigoribacterium*	*Glaciibacter*	*Gulosibacter*	*Herbiconix*	*Humibacter*	*Leifsonia*	*Leucobacter*	*Plantibacter*	*Microbacterium*	*Mycetocola*	*Pseudoclavibacter*	*Rathayibacter*	*Salinibacterium*	*Zimmermannella*	TOTAL
# of Genomes/Genus		1	2	4	1	1	8	2	1	20	2	1	1	1	1	6	10	1	49	2	1	3	1	1	120
Oligosaccharide	Richness of GHs	4.0	1.0	6.0	–	–	5.0	3.0	6.0	6.0	5.0	5.0	3.0	4.0	5.0	6.0	3.0	5.0	6.0	5.0	–	1.0	4.0	1.0	6.0
	Average # of GHs	4.8	0.2	3.5	–	–	3.5	1.0	5.7	3.3	2.2	5.0	0.7	3.0	5.7	4.3	0.4	7.5	3.8	4.0	–	0.2	2.5	0.2	3.1
	GHs per genome	29.0	1.0	21.3	–	–	20.8	6.0	34.0	20.9	13.0	30.0	4.0	18.0	34.0	26.0	2.1	45.0	22.6	24.0	–	1.0	15.0	1.0	18.6
Starch	Richness of GHs	3.0	3.0	3.0	1.0	2.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	2.0	3.0	3.0	2.0	3.0	5.0	1.0	–	3.0	2.0	3.0	5.0
	Average # of GHs	3.8	3.0	1.3	0.2	2.0	3.4	3.8	3.6	3.6	3.7	1.8	2.2	1.4	2.6	2.4	0.1	2.6	2.7	0.8	–	2.2	0.8	2.2	2.5
	GHs per genome	19.0	15.0	6.3	1.0	10.0	16.9	19.0	18.0	17.5	18.5	9.0	11.0	7.0	13.0	12.2	0.7	13.0	13.4	4.0	–	11.0	4.0	11.0	12.6
O.A.P.	Richness of GHs	2.0	–	4.0	–	–	1.0	3.0	1.0	2.0	1.0	2.0	–	1.0	1.0	3.0	–	–	4.0	1.0	–	1.0	1.0	–	4.0
	Average # of GHs	1.3	–	0.9	–	–	1.0	1.0	1.3	1.5	0.5	0.8	–	0.5	1.5	1.0	–	–	0.8	0.3	–	0.5	1.0	–	0.8
	GHs per genome	5.0	–	3.5	–	–	4.0	4.0	5.0	5.7	2.0	3.0	–	2.0	6.0	4.0	–	–	3.1	1.0	–	2.0	4.0	–	3.2
O.P.P.	Richness of GHs	3.0	–	7.0	–	–	6.0	4.0	5.0	7.0	7.0	3.0	–	1.0	3.0	6.0	2.0	3.0	9.0	4.0	–	1.0	1.0	2.0	10.0
	Average # of GHs	1.4	–	0.8	–	–	0.5	0.3	1.1	1.0	0.7	0.9	–	0.1	0.6	6.3	0.1	1.4	0.7	0.6	–	0.1	0.2	0.3	0.6
	GHs per genome	14.0	–	8.0	–	–	5.1	3.0	11.0	10.0	6.5	9.0	–	1.0	6.0	6.3	0.5	14.0	6.7	6.0	–	1.0	2.0	3.0	6.1
Mixed Polysacc.	Richness of GHs	4.0	1.0	7.0	–	–	3.0	2.0	3.0	6.0	2.0	6.0	1.0	5.0	3.0	6.0	1.0	4.0	11.0	2.0	1.0	1.0	1.0	1.0	11.0
	Average # of GHs	1.7	0.7	0.8	–	–	0.2	0.5	1.2	0.8	0.7	0.9	0.3	0.9	1.2	0.5	0.3	1.5	1.1	0.3	0.3	0.3	0.3	0.3	0.8
	GHs per genome	19.0	7.5	9.0	–	–	2.4	5.5	13.0	9.4	8.0	10.0	3.0	10.0	13.0	5.5	3.0	16.0	12.0	3.0	3.0	3.0	3.0	3.0	8.6
Cellulose	Richness of GHs	2.0	–	3.0	–	–	2.0	3.0	4.0	5.0	3.0	–	–	–	2.0	5.0	–	4.0	8.0	1.0	–	–	–	–	8.0
	Average # of GHs	0.4	–	0.3	–	–	0.3	0.3	0.8	0.5	0.6	–	–	–	0.5	0.4	–	0.6	0.3	0.3	–	–	–	–	0.3
	GHs per genome	3.0	–	2.0	–	–	2.8	2.0	6.0	4.1	5.0	–	–	–	4.0	3.2	–	5.0	2.1	2.0	–	–	–	–	2.2
Xylan	Richness of GHs	–	–	2.0	–	–	1.0	–	–	2.0	–	–	–	–	–	–	–	1.0	2.0	–	1.0	–	–	–	3.0
	Average # of GHs	–	–	0.3	–	–	1.0	–	–	0.2	–	–	–	–	–	–	–	0.3	0.1	–	0.3	–	–	–	0.2
	GHs per genome	–	–	1.0	–	–	2.9	–	–	0.7	–	–	–	–	–	–	–	1.0	0.3	–	1.0	–	–	–	0.5
Chitin	Richness of GHs	1.0	–	2.0	–	–	1.0	2.0	2.0	2.0	1.0	2.0	–	3.0	1.0	2.0	3.0	–	4.0	–	–	–	1.0	2.0	4.0
	Average # of GHs	0.5	–	1.0	–	–	0.2	1.4	1.5	0.8	0.3	0.8	–	3.5	0.8	0.7	0.4	–	0.3	–	–	–	0.5	0.8	0.5
	GHs per genome	2.0	–	4.0	–	–	0.8	5.5	6.0	3.1	1.0	3.0	–	14.0	3.0	2.8	1.7	–	1.3	–	–	–	2.0	3.0	1.9
Total	Richness of GHs	19.0	5.0	34.0	1.0	2.0	22.0	21.0	24.0	33.0	22.0	21.0	7.0	16.0	18.0	31.0	11.0	20.0	49.0	14.0	2.0	7.0	10.0	9.0	63.0
	Average # of GHs	22.0	6.0	18.0	2.0	4.0	21.1	18.5	28.0	26.9	24.0	24.0	9.0	17.0	21.0	20.7	4.5	21.2	18.0	25.0	3.0	8.0	11.0	10.0	19.4
	GHs per genome	97.0	24.5	62.0	5.0	17.0	22.0	51.0	114	80.5	60.0	74.0	22.0	57.0	92.0	74.3	8.7	69.1	51.0	105	5.0	19.0	34.0	22.0	61.3

O.A.P., other animal polysaccharides; O.P.P., other plant polysaccharides.

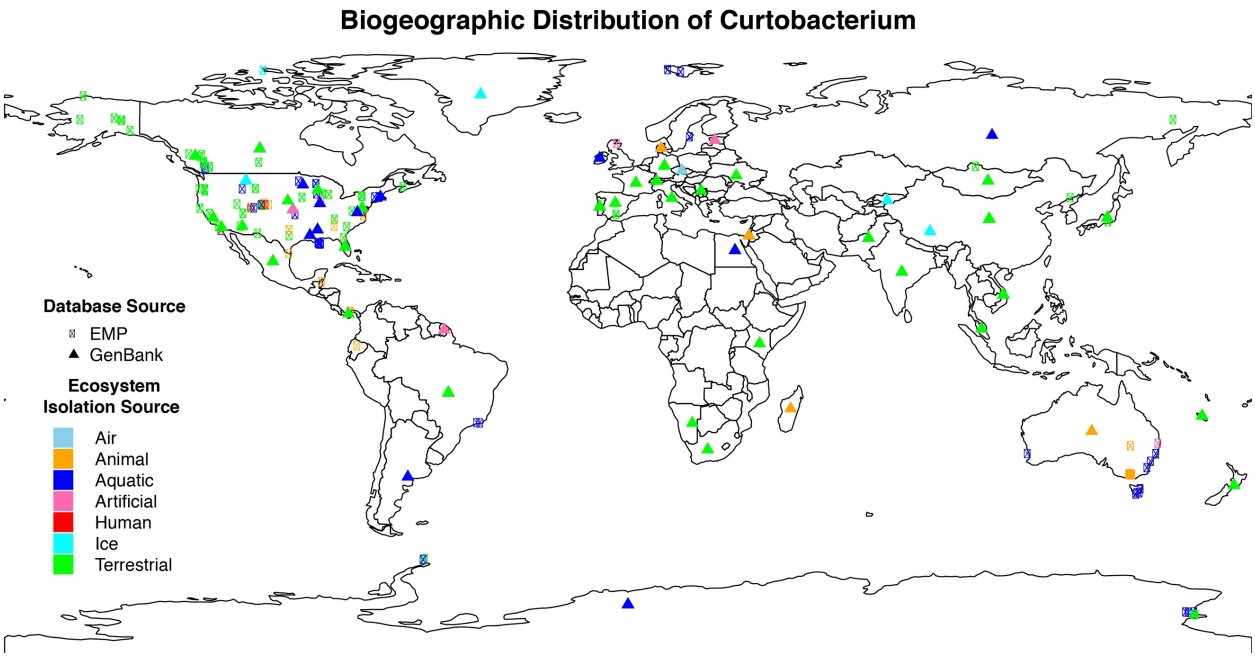


Figure 2.1 Geographic distribution of *Curtobacterium* compiled from various isolation sources. Colors indicate the different ecosystems from which the sequence was isolated. The symbol indicates the dataset from which the sequence originated. Sequences obtained from GenBank (indicated by the triangle symbol on the map) were mostly approximations as detailed GPS coordinates were not available.

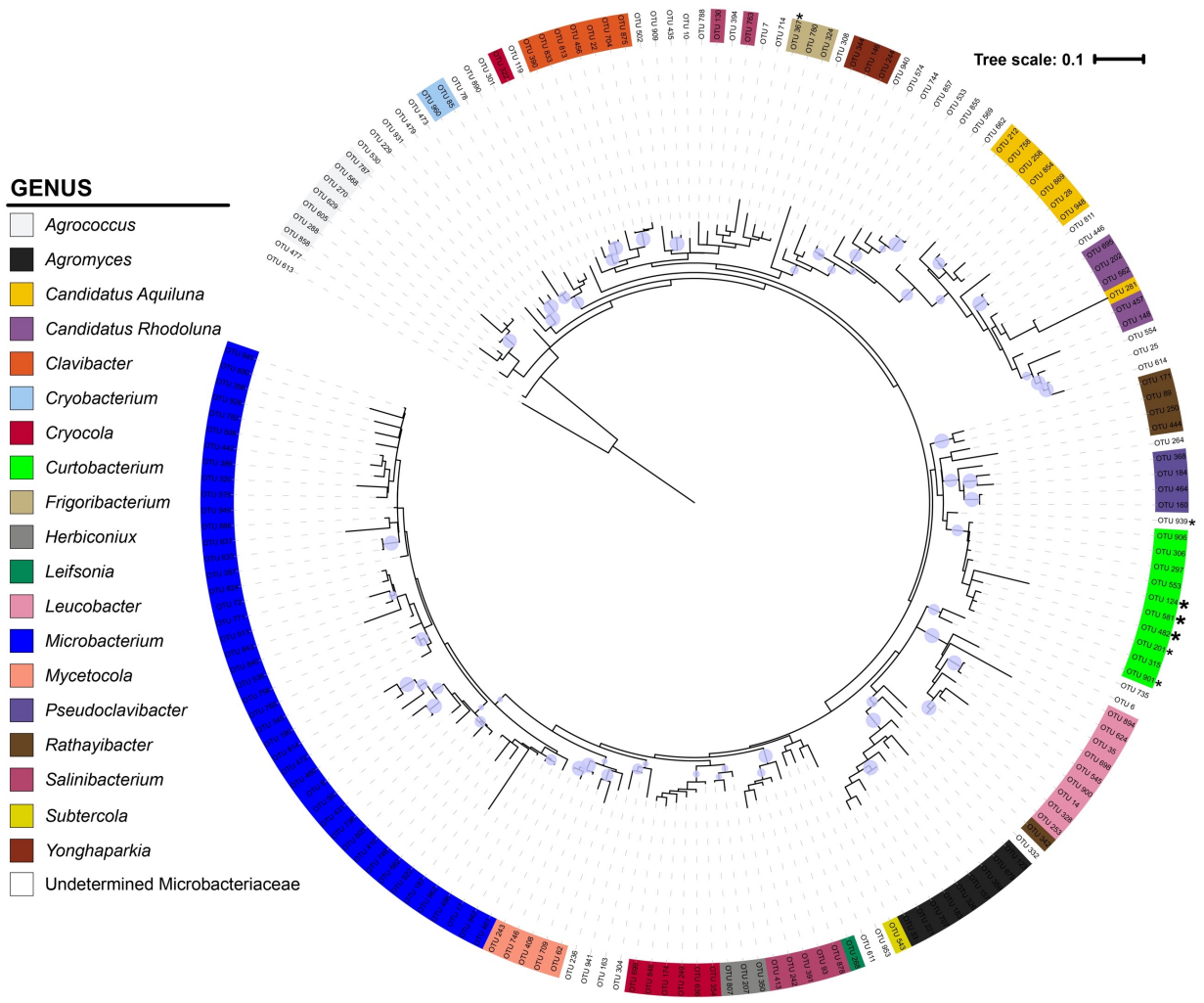


Figure 2.2 Phylogeny of Microbacteriaceae constructed from the 16S rRNA gene (maximum likelihood tree with 100 bootstraps and a GTR + GAMMA distribution). The tree is color-coded by genus using the taxonomic designation assigned from a combination of SILVA, BLAST, and RDP. The circles represent nodes with at least 70% support and the diameter of the circle represents the support level. * = 1 litter isolate within the OTU; * = >4 litter isolates in OTU.

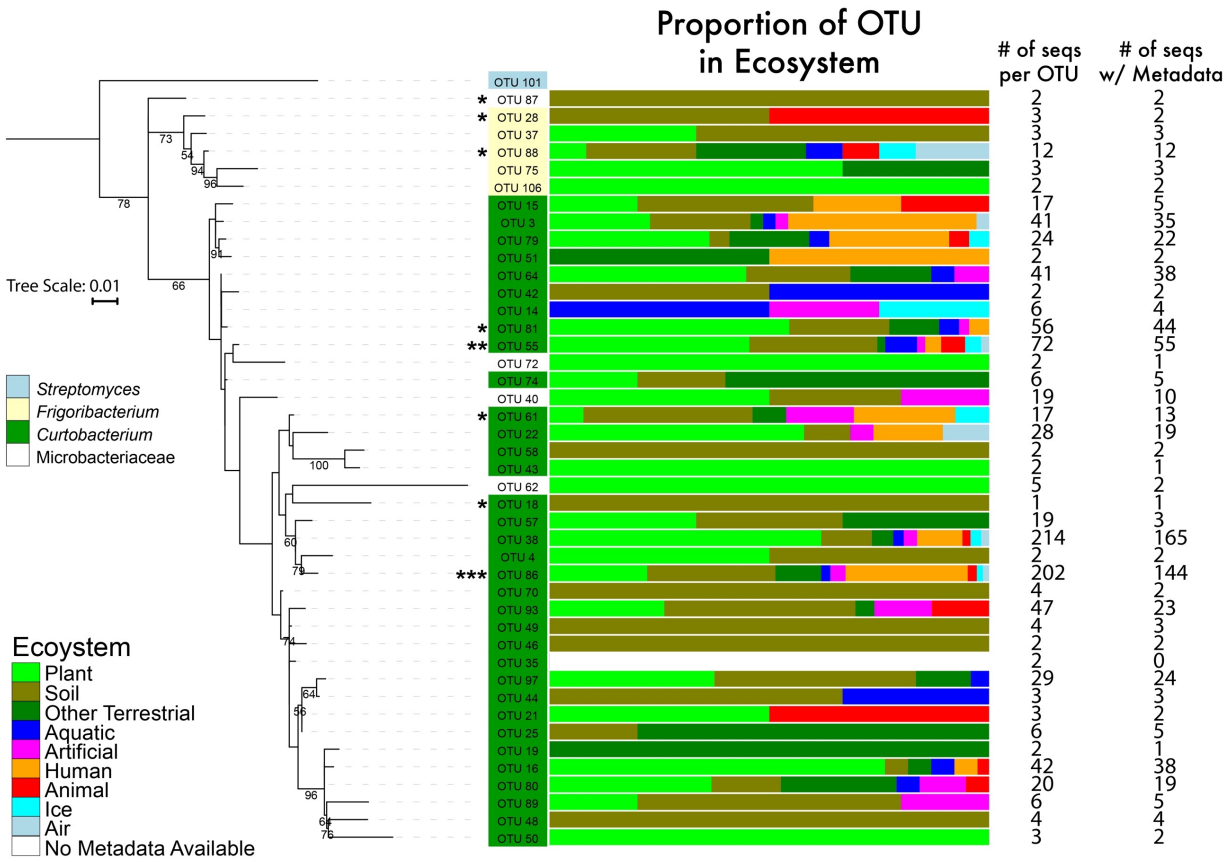


Figure 2.3 Phylogenetic tree of the genus *Curtobacterium* with 16S rRNA gene (maximum likelihood tree with 100 bootstraps and a GTR + GAMMA distribution). The numbers represent the support level of each node with at least 50% support. Bar graphs are color coded to show the percentages of the OTUs with sequences isolated from various ecosystems. Numbers in the columns represent the number of sequences incorporated into each branch for its respective OTU. * = 1-4 litter isolates within the OTU; ** = 5 isolates; *** = 10 isolates.

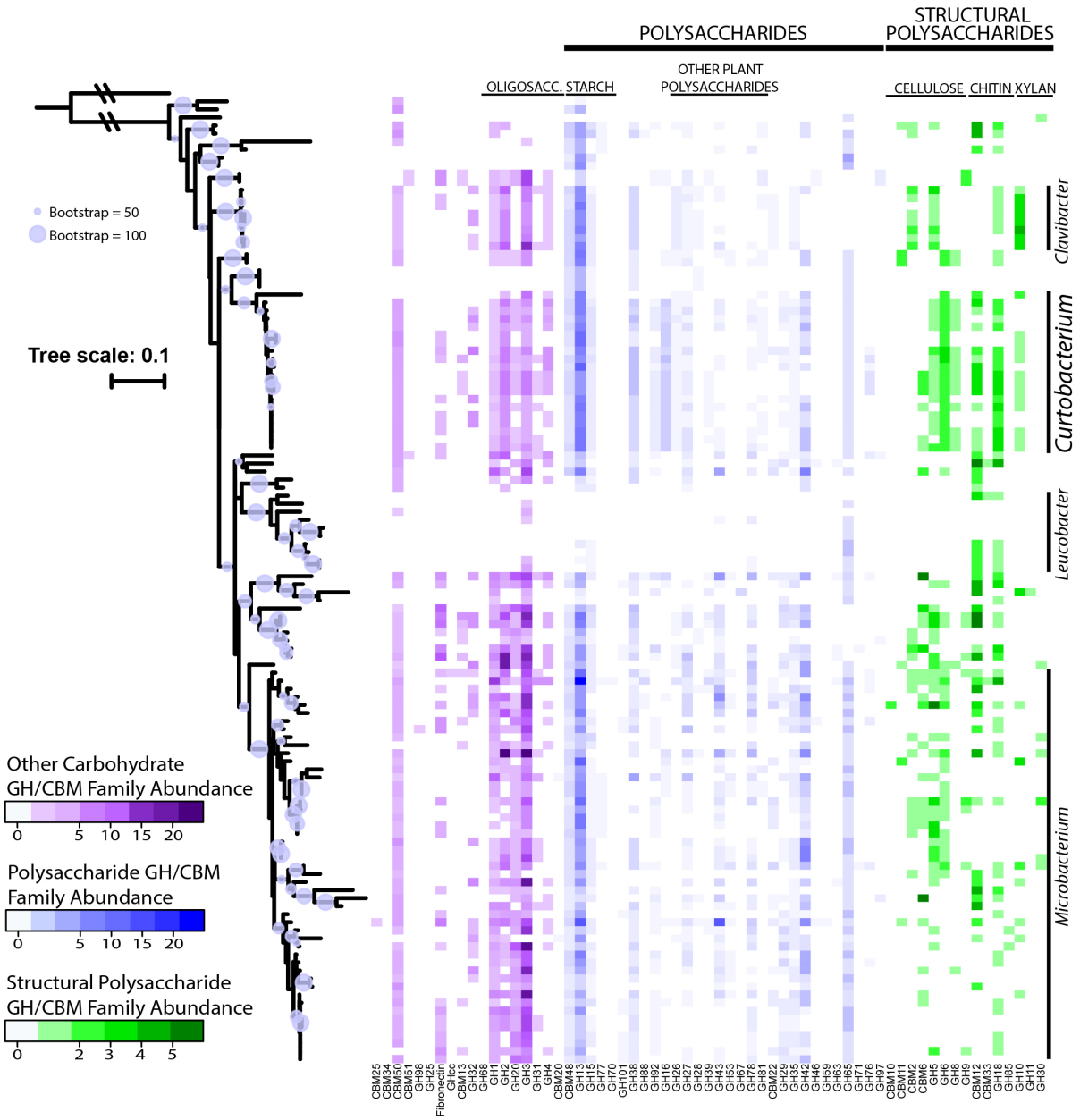


Figure 2.4 Abundance of GH and CBM families grouped and colored by substrate category across downloaded Microbacteriaceae genomes from the PATRIC database. Phylogenetic tree constructed from the 16S rRNA gene sequence (maximum likelihood tree with 100 bootstraps and a GTR + GAMMA distribution). The circles represent nodes with at least 50% support and the diameter of each circle represents the support level. Genera with more than 5 strains are denoted on the right.

CHAPTER 3

Microdiversity of an abundant terrestrial bacterium encompasses extensive variation in ecologically-relevant traits

ABSTRACT

Much genetic diversity within a bacterial community is likely obscured by microdiversity within OTUs (operational taxonomic units) defined by 16S rRNA gene sequences. However, it is unclear how variation within this microdiversity influences ecologically-relevant traits. Here, we employ a multi-faceted approach to investigate microdiversity within the dominant leaf litter bacterium, *Curtobacterium*, which comprises 7.8% of the bacterial community at a grassland site undergoing global change manipulations. We use cultured bacterial isolates to interpret metagenomic data, collected *in situ* over two years, together with lab-based physiological assays to determine the extent of trait variation within this abundant OTU. The response of *Curtobacterium* to seasonal variability and the global change manipulations, specifically an increase in relative abundance under decreased water availability, appeared to be conserved across six *Curtobacterium* lineages identified at this site. Genomic and physiological analyses in the lab revealed that degradation of abundant polymeric carbohydrates within leaf litter, cellulose and xylan, is nearly universal across the genus, which may contribute to its high abundance in grassland leaf litter. However, the degree of carbohydrate utilization and temperature preference for this degradation varied greatly among clades. Overall, we find that traits within *Curtobacterium* are conserved at varying phylogenetic depths. Similar to bacteria in marine systems, we speculate that microdiversity within this taxon may be structured into distinct ecotypes that are key to understanding *Curtobacterium* abundance and distribution in

the environment.

IMPORTANCE

Despite the plummeting costs of sequencing, characterizing the fine-scale genetic diversity of a microbial community – and interpreting its functional importance – remains a challenge. Indeed, most studies, particularly in soil, assess community composition at a broad genetic level by classifying diversity into taxa (OTUs) defined by 16S rRNA sequence similarity. However, these classifications potentially obscure variation in traits that result in fine-scale ecological differentiation among closely-related strains. Here, we investigated "microdiversity" in a highly diverse and poorly-characterized soil system (leaf litter in a southern Californian grassland). We focused on the most abundant bacterium, *Curtobacterium*, which by standard methods is grouped into only one OTU. We find that the degree of carbohydrate usage and temperature preference varies within the OTU, whereas its response to changes in precipitation are relatively uniform. These results suggest that microdiversity may be key to understanding how soil bacterial diversity is linked to ecosystem functioning.

INTRODUCTION

Currently, most studies assessing the response of bacterial communities to environmental change rely on broad taxonomic designations, for instance, by using operational taxonomic units (OTUs) based on the nucleotide sequence similarity of the 16S rRNA gene (Poretsky *et al.*, 2014). While this classification of bacterial diversity can capture broad taxonomic shifts, it provides limited genetic resolution at this loosely-defined species level (Konstantinidis and Tiedje, 2005; Cole *et al.*, 2010; Eren, Morrison, *et al.*, 2015) by obscuring important genetic diversity within the OTU (Acinas *et al.*, 2004; Thompson *et al.*, 2005; Konstantinidis and Tiedje, 2007) – so-called microdiversity (Moore *et al.*, 1998; Jaspers and Overmann, 2004). Given that most studies investigate microbial composition using 16S-defined OTUs (specifically, at the 97% level), a large gap in our understanding is the extent of microdiversity in natural communities and its relationship to variation in bacterial traits.

Growing evidence indicates that the genetic variation encompassed by bacterial microdiversity corresponds to variation in a wide range of functional traits (Larkin and Martiny, 2017). At fine genetic scales (Polz *et al.*, 2006; Eren, Sogin, *et al.*, 2015), microbes with distinct physiological traits may partition niche space within the environment (J. B. H. Martiny *et al.*, 2006; Martiny *et al.*, 2009). For example, extensive work in marine systems has demonstrated that microdiversity within a 16S-defined taxon encompasses distinct ecotypes, or lineages that respond differentially to variation in the environment over space and time (Johnson, Zinser, Coe, Mcnulty, *et al.*, 2006; Hunt *et al.*, 2008; Martiny *et al.*, 2009; Becraft *et al.*, 2011). However, our ability to characterize ecotypes at fine taxonomic levels is still largely dependent on cultured organisms because of the need to link genomic to phenotypic variation (Choi *et al.*,

2016). And while metagenomic sequencing has advanced the identification of uncultivated organisms (Simmons *et al.*, 2008), the functional role of microdiversity has rarely been considered in soils as we lack the cultured representatives of their abundant members.

Diverse bacterial and fungal communities on leaf litter, the top layer of soil, play a key role in the carbon cycle. Litter decomposition mediates the loss of carbon through respiration to the atmosphere or its storage as organic matter in soil (Adair *et al.*, 2008). The Loma Ridge Global Climate Experiment (LRGCE) in southern California was established to test how future changes in precipitation and nitrogen availability may alter semi-arid grassland and coastal sage scrub ecosystems. In grasslands at the LRGCE, the litter microbial community is dominated by bacteria (Alster *et al.*, 2013), suggesting that bacteria perform the bulk of grassland litter decomposition. Over a two-year period, the leaf litter community responded weakly, but significantly to treatment manipulations (Allison *et al.*, 2013; Matulich *et al.*, 2015). At the 97% OTU level, a *Curtobacterium* OTU (phylum: *Actinobacteria*, family: *Microbacteriaceae*) was the most abundant taxon within the bacterial community (Matulich *et al.*, 2015). An analysis of *Curtobacterium* sequences from around the globe revealed the genus to be a cosmopolitan terrestrial taxon, with isolates primarily derived from plant and soil habitats (Chase *et al.*, 2016). Further, genomic sequencing of *Curtobacterium* strains isolated from leaf litter indicated that the genus has a high genomic potential to decompose polymeric carbohydrates such as starch, cellulose, and xylan that are abundant in leaf litter (Chase *et al.*, 2016).

Our previous work demonstrated that isolates belonging to *Curtobacterium* harbored extensive genomic diversity despite being clustered within a single OTU as defined by 16S rRNA (Chase *et al.*, 2016). Here, we ask 1) What is the extent of *Curtobacterium* microdiversity in a

natural leaf litter bacterial community? and 2) Does this microdiversity encompass genetic and physiological variation in ecologically-relevant traits? To address these questions, we used a combination of environmental field data and physiological lab assays to assess the distribution of traits within *Curtobacterium* and their phylogenetic conservatism. First, we examined the response of *Curtobacterium* microdiversity to manipulations of precipitation and nitrogen availability by using cultured isolates to inform metagenomic data. Moisture limitation, in particular, is likely a major stressor on litter bacteria in Southern California, which experiences long dry seasons with little to no rainfall. Second, we assayed both the genomic potential and metabolic capacity of isolates to depolymerize cellulose and xylan. As leaf litter is primarily composed of these polysaccharides, access to this primary carbon supply in this environment may be a highly advantageous trait.

RESULTS

Curtobacterium Abundance and Microdiversity

We characterized *Curtobacterium* abundance and its microdiversity at the LRGCE using 48 metagenomic sequence libraries from litter samples collected over a two-year period. To estimate its relative abundance within the bacterial community, we created a custom pipeline using a curated reference database of 3,019 genomes representing 1,464 bacterial genera including 16 *Curtobacterium* genomes (Fig. S1). We calculated taxonomic abundance by using a phylogenetic classification of the metagenomic reads against the reference phylogeny, which we constructed using single-copy marker genes (Wu *et al.*, 2013). Using our pipeline, we identified *Actinobacteria* and *Microbacteriaceae* as the most abundant phylum (46.3%) and family (28.2%), respectively. We detected similar relative abundances using MG-RAST

annotations of the marker genes (see Supplementary Information; Table S1), but this approach did not detect any *Curtobacterium*. Therefore, we used the new pipeline to investigate finer taxonomic levels. This analysis revealed that *Curtobacterium* was the most abundant genus observed over the two-year period within the leaf litter community, comprising an average of 7.8% of the bacterial community. However, even with our pipeline, we were unable to characterize 31.2% of the marker genes at or below the genus level.

Based on the full-length 16S rRNA gene, the *Curtobacterium* genomes (14 of which were cultured isolates from leaf litter (Chase *et al.*, 2016) and two other publicly available genomes), clustered into the same OTU defined at the 97% sequence identity level. We therefore identified genomic clusters within the *Curtobacterium* OTU using a phylogenetic analysis of 29 single-copy marker genes and designated the isolates into six well-supported clades (Fig. 1A). These clades were supported by nucleotide (ANI) and amino acid (AAI) similarity (Table S2). Specifically, isolates shared >97% AAI within clades for the 29 marker genes. Across the whole genome, isolates within clades were more similar in ANI and AAI than with isolates between clades, which had a minimum pairwise similarity of 83.2% ANI and 78.9% AAI across all *Curtobacterium* isolates.

We then classified the metagenomic marker gene reads assigned to *Curtobacterium* in the taxonomic analysis onto the six identified clades. Only a tiny fraction (0.27% of the total bacterial community) identified as *Curtobacterium* but failed to classify into one of the six clades, suggesting that our isolates encompassed most of the genomic diversity of *Curtobacterium* at the LRGCE. Across all samples, *Curtobacterium* was dominated by two clades (Table S1); Clades IA and III averaged 3.0% and 2.4% of the marker gene sequences, respectively

(Fig. 2C). Together, the remaining *Curtobacterium* clades (Clades IB, IC, IIA, and IIB) composed >2% of the bacterial community, but, separately, each of the four clades represented <0.6% of the bacterial community (Table S1).

Response to the Global Change Treatments

Within the global change experiment, the composition of the microbial community varied seasonally by sampling date such that some bacterial phyla, including *Actinobacteria*, were strongly correlated with background precipitation (Fig. S2A) as previously reported (Berlemont *et al.*, 2014). Indeed, at the phylum level (Fig. 2A), bacterial composition varied significantly with time (Bray-Curtis similarity; PERMANOVA: $p<0.002$) and responded marginally to the global change treatments of reduced precipitation (drought) and added nitrogen ($p=0.061$), with no significant interaction between the two factors. However, the global change treatments explained only 1.9% of the variation in phylum composition, whereas time (date of collection over a two-year period) explained 65.1%. In particular, during a prolonged hot, dry season in year two (Fig. 2B), the bacterial community became dominated by *Actinobacteria* (Fig. 2A).

Much of the response of *Actinobacteria* to the global change treatments was due to *Curtobacterium*. The relative abundance of all *Curtobacterium* increased by 20.2% in the drought treatment and decreased by 17.2% in the nitrogen treatment relative to the control plots (PERMANOVA: $p<0.05$; Table S3). Similar to the phylum-level response, time of sampling explained the greatest amount of variation in *Curtobacterium* abundance, accounting for 52.6%, while treatment contributed only 5.0%. *Curtobacterium* abundance was strongly associated with seasonal precipitation (Fig. S3A), increasing in relative abundance during the

dry seasons and accounting for over 10% of all leaf litter bacteria in the second, drier year of the study (Figs. 2B and 2C). *Curtobacterium* abundance, however, was not correlated with the mean temperature in the field two weeks prior to sampling (Fig. S3B).

We next tested whether microdiversity within *Curtobacterium* (and in particular, the six identified clades) varied in its response to the global change treatments. All *Curtobacterium* clades responded similarly to drought, increasing in abundance relative to the control and, with the exception of Clade IC, responded negatively to the increased nitrogen treatment (Table S3). Furthermore, *Curtobacterium* clade composition varied significantly over time (PERMANOVA: $p<0.001$; Fig. S2B), with Clades IA and III increasing in relative abundance during the drier, second year of the study (Fig. 2C).

Carbohydrate Degradation Traits within *Curtobacterium*

Genomic Characterization

To analyze the genomic potential for carbohydrate degradation, we characterized the glycoside hydrolase (GH) and carbohydrate binding module (CBM) protein families within and among *Curtobacterium* clades. The abundance of total GH/CBM genes varied among all genomes, ranging from 58-98 GH/CBM copies. The total distribution of GH/CBM genes varied significantly with phylogenetic distance such that more closely related genomes carried more similar copy numbers (RELATE; $\rho=0.45$, $p<0.05$; Fig. S4). Clades IA, IIB, and III encoded the highest abundance of GH/CBM genes (an average of 86, 87.7, and 84.5 genes), which differed significantly (ANOVA, $F_{(5,10)}=5.3027$; $p<0.05$) from Clade IIA (65 genes), whereas Clades IB and IC encoded an intermediate number (76.5 and 78.3, respectively; Fig. 1B).

Next, we considered the GH/CBM gene diversity that is thought to be responsible for degradation of the most abundant carbohydrates in the leaf litter at the LRGCE, cellulose and hemicellulose (specifically, xylan) (Allison *et al.*, 2013). Overall, the number of both cellulose- and xylan-related GH/CBMs were significantly correlated with phylogenetic distance (RELATE; $\rho=0.57$, $p<0.01$ and $\rho=0.26$, $p<0.05$, respectively). All *Curtobacterium* genomes contained at least one copy of a GH or CBM protein family that targeted either cellulose or xylan. However, some strains (e.g. MCBA15013 and MCBA15016, both from Clade IIB) had an elevated abundance of GH/CBM genes targeting cellulose, with an apparent absence of genes targeting xylan. Clades IA and IB were the only clades to contain both GH and CBM genes targeting each substrate (Fig. 1B).

Phenotypic Characterization

The presence of GH/CBM genes within a genome only suggests the potential for substrate utilization. Therefore, we conducted substrate assays in the lab to confirm each isolate's ability to degrade cellulose and xylan at 22°C. We performed these assays at this temperature as the optimum growth for the genus is thought to range from 20-26°C (Evtushenko and Takeuchi, 2006; Whitman *et al.*, 2012). All but one of the strains (MCBA15001) degraded both cellulose and xylan over a four-day period, including the four isolates that did not encode known xylan-targeting genes (Fig. 1B). Indeed, the size of an isolate's zone of depolymerization was not correlated with the abundance of either cellulose- (independent phylogenetic contrasts: PIC, $F_{(1,14)} = 1.24$; $p>0.05$) or xylan- (PIC, $F_{(1,14)} = 0.15$; $p>0.05$) targeting genes.

The degradation patterns of the *Curtobacterium* strains also depended greatly on the temperature of the assay. When isolates were assayed at 37°C, the expected maximum temperature for growth in *Curtobacterium* (Evtushenko and Takeuchi, 2006), four strains saw an increase in degradation capability, including two strains from Clade IA, while three strains were unable to degrade either substrate at 37°C (Fig. 1B). The total area of the zone of depolymerization varied significantly by temperature (ANCOVA, $F_{(1,43)} = 4.67$; $p<0.05$) and clade ($F_{(5,43)} = 4.74$; $p<0.01$), with a significant interaction between them ($F_{(5,43)} = 2.46$; $p<0.05$), whereas the substrate of the assay had no effect on depolymerization area ($F_{(1,43)} = 0.95$; $p>0.05$). When averaged across *Curtobacterium* clades, only Clade IA saw an average increase in depolymerization area when strains were grown at 37°C when compared to 22°C (Fig. 1C). Most clades maintained some level of degradation capability at the higher temperature except for Clade III, which failed to depolymerize either cellulose or xylan at 37°C (Fig. 1C).

DISCUSSION

In this study, we investigated the extent of genomic microdiversity of *Curtobacterium* in the field and the relationship between this diversity and the bacterium's functional traits. To our knowledge, this study is the first to do so in a dominant soil bacterium. As in aquatic and host-associated ecosystems (Moore *et al.*, 1998; Thompson *et al.*, 2005; Martiny *et al.*, 2009; Frese *et al.*, 2011; Morrison *et al.*, 2012; Williams *et al.*, 2014), microdiversity within this abundant bacterium is extensive. Co-occurring strains within the same *Curtobacterium* OTU share as low as an 83% average nucleotide identity (ANI), far below the traditional species boundary (Richter and Rosselló-Móra, 2009). Our results support the growing understanding that traditional taxonomic assignments (i.e. OTUs) are insufficient to resolve ecologically

distinct microorganisms into their correct taxonomic assignments (Rodriguez-R and Konstantinidis, 2014; Varghese *et al.*, 2015). Indeed, extensive *Curtobacterium* microdiversity persists in grassland leaf litter and encompasses variation in several ecologically-relevant traits, including its ability to degrade abundant carbohydrates as well as temperature preferences for this degradation. Thus, binning of 16S rRNA sequences obscures detection and interpretation of ecologically important trait variance.

Trait variability within soil bacterial OTUs has been described previously, suggesting that local adaptation and coexistence are probable among closely related strains (Schloter *et al.*, 2000; Wielbo *et al.*, 2007; Choudhary and Johri, 2011; Schlatter and Kinkel, 2014). However, the combination of lab assays on cultured representative isolates in conjunction with metagenomic data allowed us to compare the physiological findings to their representation in the environment, as well as test the response to environmental change in the context of the whole community. Further, this combination enabled us to quantify and interpret metagenomic data of ecologically-relevant microdiversity that would otherwise be undetectable (Table S1) due to lack of genomic representation in public databases. Indeed, particularly for terrestrial soil systems, the genomic reference databases often lack the resolution to detect fine-scale taxonomic groups, as defined as >95% ANI (Nayfach *et al.*, 2016), or result in mischaracterization of taxonomic groups altogether (Gonzalez *et al.*, 2016).

The results of this study are also consistent with the idea that bacterial traits are often conserved at varying phylogenetic depths (Martiny *et al.*, 2009, 2015). Complex quantitative traits like an organism's response to drought have been proposed to be more phylogenetically conserved (Martiny *et al.*, 2015; Larkin and Martiny, 2017). Here, we observed that a response

to dry conditions (both in the drought treatment and the dry seasons) appears to be generally consistent among *Curtobacterium* clades, suggesting that biological and physiological traits responsible for moisture response are ecologically cohesive (Philippot *et al.*, 2010) within this taxon. Thus, the response of *Curtobacterium* to future drought would likely be apparent at the OTU level, although certain clades may be more abundant than others. However, given that some of the clades were relatively rare within the community, further investigation is still needed to confirm this interpretation.

In contrast, traits that rely on one or a couple genes such as carbon utilization are thought to be more shallowly conserved (Martiny *et al.*, 2013) as they may be more prone to horizontal gene transfer. Using physiological assays, we confirmed the genomic potential for *Curtobacterium* to degrade polymeric carbohydrates, which are likely central to their success within the leaf litter community. Although all *Curtobacterium* clades could depolymerize both xylan and cellulose, the degree of carbohydrate utilization varied among and within clades, suggesting the carbohydrate utilization is finely conserved. Such intricate differences in carbohydrate degradation traits among *Curtobacterium* may contribute to the persistence of this microdiversity within the leaf litter community. However, the genomic potential for carbohydrate utilization (number and composition of GH/CBMs) did not predict observed phenotypic variation in the lab, highlighting the difficulty in using gene annotations to predict ecological roles.

Carbohydrate degradation was also temperature dependent, regardless of the substrate. Further, this dependency varied among clades, revealing that *Curtobacterium* microdiversity also incorporates variation in temperature preference. Broadly, this result

supports the idea that bacterial temperature preference can be relatively finely conserved (Martiny *et al.*, 2015), despite typically being viewed as an adaptive response (Bennett *et al.*, 1992). More specifically, it suggests differential physiological tradeoffs between temperature and carbohydrate utilization (Schimel *et al.*, 2007) among clades. Such variation in this tradeoff might explain the coexistence of these closely-related clades, particularly for the two most abundant clades, Clades IA and III. Despite similar environmental responses to drought and seasonal fluctuations, these clades exhibited opposite responses to temperature with respect to carbohydrate utilization (Fig. 1C). While temperature preference has previously been shown to drive shifts in ecotype abundance within marine systems (Johnson, Zinser, Coe, Mcnulty, *et al.*, 2006), we did not observe a correlation between clade abundance and temperature at this one site. However, further investigation is needed across a wider temperature range to test whether temperature drives the geographic distribution of *Curtobacterium* clades.

In conclusion, the microdiversity within a single *Curtobacterium* OTU in this grassland leaf litter encompasses variation in traits involved in carbon degradation and temperature preference. Classic ecological theory would suggest that this trait variation allows microdiversity to occupy distinct ecological niches (Chase and Leibold, 2003), although further work is needed to identify distinct *Curtobacterium* ecotypes in the environment. At the same time, *Curtobacterium* appears to be consistent in its response to changes in precipitation, suggesting that variability in moisture conditions are unlikely to explain the maintenance of this microdiversity. Thus, similar to marine bacteria (Moore *et al.*, 1998; Jaspers and Overmann, 2004), our work highlights that the depth of trait conservatism (Martiny *et al.*, 2009) may help to understand the response of soil bacteria to changing environments.

METHODS AND MATERIALS

Field Site

The Loma Ridge Global Change Experiment (in Irvine, California, USA [33° 44′ N, 117° 42′ W]; (Potts *et al.*, 2012)) was established in 2007 with precipitation and nitrogen manipulations in areas of deciduous shrubland (coastal sage scrub) and annual grasses. For this study, we sampled only in the grassland plots, which are dominated by *Avena*, *Bromus*, and *Lolium* (Allison *et al.*, 2013). We used a subset of the plots that included reduced precipitation treatment (-50% reduction in annual precipitation), added N treatment (20-40 kg N/ha), and a control treatment, as previously described (Allison *et al.*, 2013).

We collected leaf litter from these plots by sampling each season from May 2010 – March 2012 across three treatments: control, reduced precipitation (drought), and added nitrogen (8 time points x 3 treatments x 2 replicates). As described previously, metagenomic libraries were created from these samples by extracting DNA from ground litter, prepared using a TruSeq library kit (Illumina, San Diego, CA, USA), and sequenced on an Illumina HiSeq2000. Samples were pooled from 8 plots from each treatment to form the 2 replicate libraries at each time point (for more information, see (Berlemont *et al.*, 2014)). The sequences libraries are available on MG-RAST under the project IDs 4511045-4511050, 4511060-4511065, 4511111-4511116, 4511134-4511153, 4511178-4511193. We excluded two libraries (Drought April 2010 and Nitrogen August 2010) due to low sequence count. Temperature and precipitation data was recorded at a nearby weather tower (Allison *et al.*, 2013).

Curated Marker Gene Reference Database

We developed a reference genomic database to characterize phylogenetic marker genes from the metagenomic sequences of the microbial community. This approach is similar to PhyloSift (Darling *et al.*, 2014), except we performed a more robust search to compensate for the lack of genomic references to characterize soil microbial communities. We downloaded 79,838 genomes from the PATRIC database (Wattam *et al.*, 2014) with RAST (Aziz *et al.*, 2008) annotations on December 9th, 2016. We screened all genomes for annotations of 29 conserved, single-copy phylogenetic marker genes (Wu *et al.*, 2013) and discarded failed genomes, most of which were draft genomes with >1000 contigs. Remaining genomes were manually curated by assigned nomenclature to include two genomes per genus. When available, we prioritized complete genomes and genomes isolated from soil ecosystems. The 3,159 resulting genomes were combined with 14 *Curtobacterium* genome sequences isolated from two grassland leaf litter sites (Chase *et al.*, 2016), including four strains isolated during the time of metagenomic sampling from the LRGCE.

We curated the downloaded genomes to ensure all genomes were properly assigned to the correct taxonomy. Individual marker genes from each genome were aligned using ClustalO v1.2.0 (Sievers *et al.*, 2011) and used to construct a 15,963 bp concatenated alignment for phylogenetic analysis using FastTree2 (Price *et al.*, 2010). The resulting reference phylogeny guided the construction of each individual marker gene tree to maintain relative node structure across trees. For each marker gene tree, we performed a maximum likelihood bootstrap analyses using RAxML v8.0.0 (Stamatakis, 2014) under the PROTGAMMAWAGF model for 100 replicates. If a genome was named incorrectly or showed a problematic alignment for any of

the individual marker gene trees (i.e. genome terminal branch length was >5), the entire genome was removed (total of 154 genomes were removed) and all trees were re-generated.

The NCBI taxonomy database (Federhen, 2012) was downloaded on June 17, 2016. The taxonomic information of the remaining 3,019 genomes were added locally to the NCBI database using the PATRIC genome IDs. The individual marker gene trees and taxonomic information were all used to generate reference packages for the program PPlacer v1.1.alpha17 (Matsen *et al.*, 2010). Reference packages were subsequently used to characterize the microbial community (available at https://github.com/alex-b-chase/LRGCE).

Metagenomic Analyses

To evaluate the taxonomic diversity of the bacterial community as well as finer-scale diversity within *Curtobacterium* at the LRGCE, we re-analyzed the metagenomic libraries previously described (Berlemont *et al.*, 2014). Metagenomes were retrieved from the metagenomics analysis server (MG-RAST) (Meyer *et al.*, 2008) after sequences had been processed for quality control and coding regions were predicted by FragGeneScan (Rho *et al.*, 2010). We performed an initial filter using BLASTP (Altschul *et al.*, 1997) against our custom database with an e-value of $1x10^{-5}$. We applied a secondary filter using HMMer v3.1b2 (Finn *et al.*, 2011) with an e-value of $1x10^{-10}$ to achieve a higher specificity. We grouped the filtered reads for each library by each marker gene and aligned them using ClustalO v1.2.0 (Sievers *et al.*, 2011) to the corresponding marker gene reference package (see above). Aligned metagenomic reads were "placed" onto the reference phylogeny using PPlacer v1.1.alpha17 (Matsen *et al.*, 2010), keeping at most 20 placements, and a posterior probability for final placement on the reference tree was calculated. Finally, we created single branch abundance

matrices yielding an abundance distribution ranging from phyla to individual genomes. All abundances were normalized by the total number of marker genes present.

Comparison of Curated Pipeline to other Methods

To validate the taxonomic results generated by our custom pipeline (see Supplementary Information), we compared our taxonomic abundances using two alternative approaches: 1) the MG-RAST pipeline using a read-based analysis and 2) a *de novo* co-assembly of all metagenomic libraries using the paired-end reads.

First, to generate the MG-RAST taxonomic profiles, we downloaded the KEGG database annotations for each library from the MG-RAST API (Meyer *et al.*, 2008) and calculated relative abundances across all annotated reads. Next, we standardized the MG-RAST output by filtering the MD5 IDs corresponding to the 29 marker genes and regenerated standardized taxonomic abundance profiles. All gene sequences retrieved from MG-RAST were assigned to the closest hit genus in the MG-RAST database using an e-value of $1x10^{-5}$.

Second, we conducted a genome-centric analysis by performing a *de novo* co-assembly of all of the paired-end shotgun metagenomic libraries using MEGAHIT (Li *et al.*, 2014). We used an iterative k-step ranging from k=27-111 and discarded all assembled contigs <3000bp. Read coverage for each assembled contig was calculated using bbwrap.sh within the suite of tools available via BBMap v35.66 (Bushnell, 2016). Taxonomic assignments for all assembled contigs were generated using MegaBLAST against the NCBI nt database (January 2015 version) with an e-value of $1x10^{-5}$.

Genomic Comparisons of the Isolates

To validate that all *Curtobacterium* genomes, including two publically available *Curtobacterium* genomes, clustered within the same OTU, we used Barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml) to predict rRNA genes and clustered the 16S rRNA gene using UCLUST (Edgar, 2010). We then examined the relationship among all 16 *Curtobacterium* genomes using 29 single-copy phylogenetic marker genes (Wu *et al.*, 2013). Each conserved gene was independently aligned using ClustalO v1.2.0 (Sievers *et al.*, 2011) and used to create a concatenated alignment for phylogenetic analyses. We constructed a maximum likelihood phylogenetic tree using RAxML v8.0.0 (Stamatakis, 2014) under the PROTGAMMAWAGF model for 100 replicates. For convenience, we designated six monophyletic clades based on the results from the phylogenetic analyses. To confirm these designations, we calculated pairwise average amino acid identity (AAI) across the 29 marker genes across all genomes.

Next, we confirmed that our clade designations were in accordance with additional genomic characterizations. Specifically, we calculated pairwise whole genome average nucleotide identity (ANI) and AAI (Rodriguez-R and Konstantinidis, 2016), and computed the core genome for within each clade by generating groups of orthologous proteins with MCL (Enright *et al.*, 2002). Genes identified as orthologous groups within clades were subsequently used to calculate AAI of all clade-specific core genes. All genomic analyses were performed using the suite of tools available in the Microbial Genomes Atlas (MiGA; https://github.com/bio-miga/miga).

To analyze each genome for its potential to degrade carbohydrates, genomic ORFs were generated by the RAST annotation pipeline (Aziz *et al.*, 2008) and searched using HMMer

against the Pfam-A v30.0 database (Finn *et al.*, 2016). We then used a subset of identified protein families, representing glycoside hydrolase (GH) and carbohydrate-binding module (CBM) proteins to identify the genomic potential to degrade carbohydrates of each isolate (Berlemont and Martiny, 2013; Chase *et al.*, 2016). GH/CBM gene composition profiles for each isolate were subsequently used to generate a Bray-Curtis similarity matrix to produce a non-metric multi-dimensional scaling (MDS) ordination plot.

Physiological Analyses of the Isolates

In the laboratory, we characterized the 14 *Curtobacterium* isolates for their ability to utilize two polysaccharides, cellulose and xylan, at two temperatures. All isolates were grown from -80°C freezer stocks for 24-48 h in LB liquid media at room temperature (22°C). Isolates were spun down at 13,500 RCF for 4 minutes with LB supernatant being discarded. Pelleted cultures were washed with 0.9% saline solution three times and re-suspended in 10 mL of M63 minimal media with 0.5% w/v dextrose and allowed to grow for 24 h. All cultures were then diluted to OD_{600} = 0.1 to ensure equal cell density across isolates. We used 10 µL of grown cultures (in triplicates) to inoculate onto solid M63 media containing 0.5% w/v carboxymethyl cellulose (CMC; MPBio 150560) or xylan (Sigma X0502) and were placed at 22°C (optimum temperature for growth (Evtushenko and Takeuchi, 2006; Whitman *et al.*, 2012)) and 37°C (maximum temperature for growth (Evtushenko and Takeuchi, 2006)). Depolymerization of each substrate was classified after 4 days by measuring the zones of transparent growth around the inoculum as previously described (Pold *et al.*, 2016) with Gram's iodine stain (Kasana *et al.*, 2008). We analyzed the zones of depolymerization around inoculated colonies on ImageJ

(https://imagej.nih.gov/ij/) to calculate the total area of carbohydrate degradation. An *E. coli* strain was included as a negative control for all physiological assays.

Statistical Analyses

To test the effects of environmental treatment manipulations on the distribution of bacterial communities and *Curtobacterium* clade composition, we used a permutational multivariate analysis of variance (PERMANOVA) (Clarke, 1993). The statistical model included plot treatment (control, drought, or N addition) and date of collection as fixed effects. We tested the effects of time and treatment by generating Bray-Curtis similarity matrices at the phyla and clade taxonomic levels. Subsequent PERMANOVA analyses used a type III partial sum of squares for 999 permutations of residuals under a reduced model. Similarity matrices were also used to generate non-metric multi-dimensional scaling (MDS) ordination plots. All multivariate statistical analyses were conducted using PRIMER6 with the PERMANOVA+ function (Primer-E Ltd, Ivybridge, UK).

We analyzed the distribution of GH/CBM genes within and among *Curtobacterium* clades. To test for differences in the total abundance of GH/CBM proteins across clades, we used a one-way analysis of variance (ANOVA). For the ANOVA analysis, we used a Tukey "Honest Significance Difference" to detect the difference in total abundance of GH/CBM genes across clades. To test for correlations between the abundance of GH/CBM proteins, with respect to cellulose and xylan, and phylogenetic distance, we calculated a Spearman's rank correlation coefficient using a RELATE test. Further, we performed a phylogenetic independent contrast (PIC) analysis to test whether the abundance of GH/CBM genes related to an isolate's phenotypic ability to degrade cellulose or xylan in the laboratory. Finally, to determine the

factors driving degradation, we used a multiple regression model including the following variables: temperature, clade designation, and carbon substrate. Starting with a three-way ANCOVA, we implemented a backward selection process (Mac Nally, 2002). If the model returned non-significant interactions, the interaction was removed and the model was regenerated to decrease the chance of spurious relationships (Harrell, 2015). All analyses were performed in the R software environment.

ACKNOWLEDGEMENTS

We thank Michaeline Nelson and Stephen Allison for comments on earlier versions; Alberto Lopez for assistance in physiological assays; Kristin Dolan and Renaud Berlemont for their use of isolates and metagenomic data, respectively; and Michael Goulden for establishing and maintaining the LRGCE. Support for this project was provided by a U.S. Department of Education Graduate Assistance of National Need (GAANN) fellowship (ABC), National Science Foundation (DEB-1457160), and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC-SC0016410 and DE-PS02-09ER09-25. Work at Lawrence Berkeley National Laboratory (ELB, UK) was supported by the Department of Energy under contract DE-AC02-05CH11231 with the University of California.

Figures

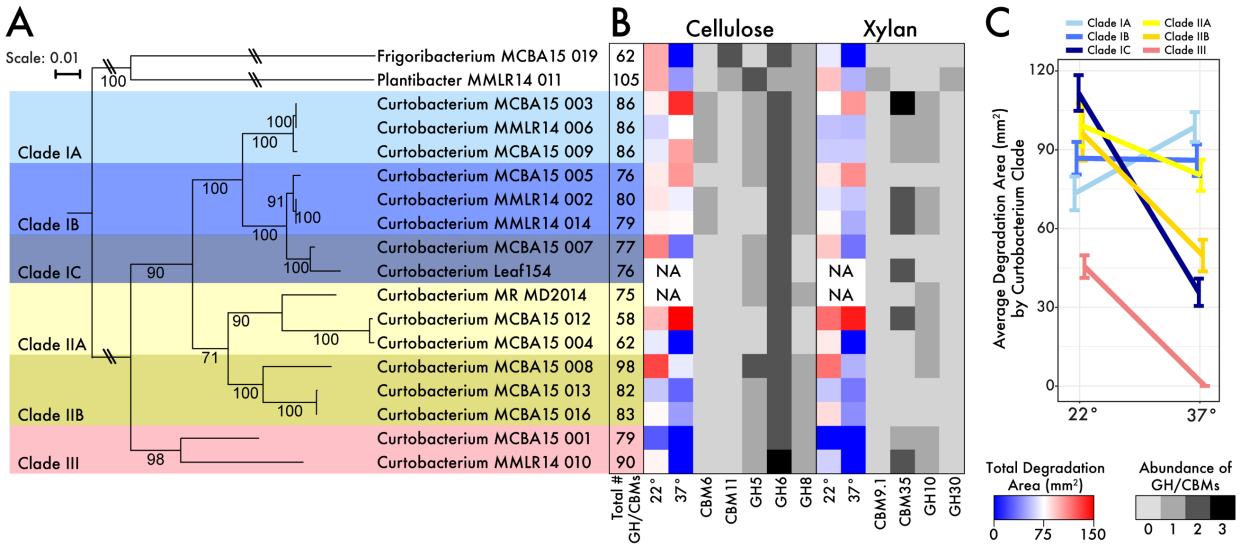


Figure 3.1 Phylogeny and traits of *Curtobacterium* strains. **A)** Multilocus phylogenetic analysis using a concatenated alignment of 29 single-copy marker genes. **B)** Genomic and physiological metrics of carbohydrate utilization. The total number of GH/CBM families targeting all potential carbohydrate substrates is shown in the first column. Physiological ability to degrade cellulose and xylan is shown in blue/red while the genomic potential (presence of GH/CBM families) to degrade either cellulose or xylan are represented in grey/black. Strains designated as "NA" were not assayed for carbon degradation. **C)** The average degradation area (± 1 SD) of the substrates by *Curtobacterium* clade at each temperature.

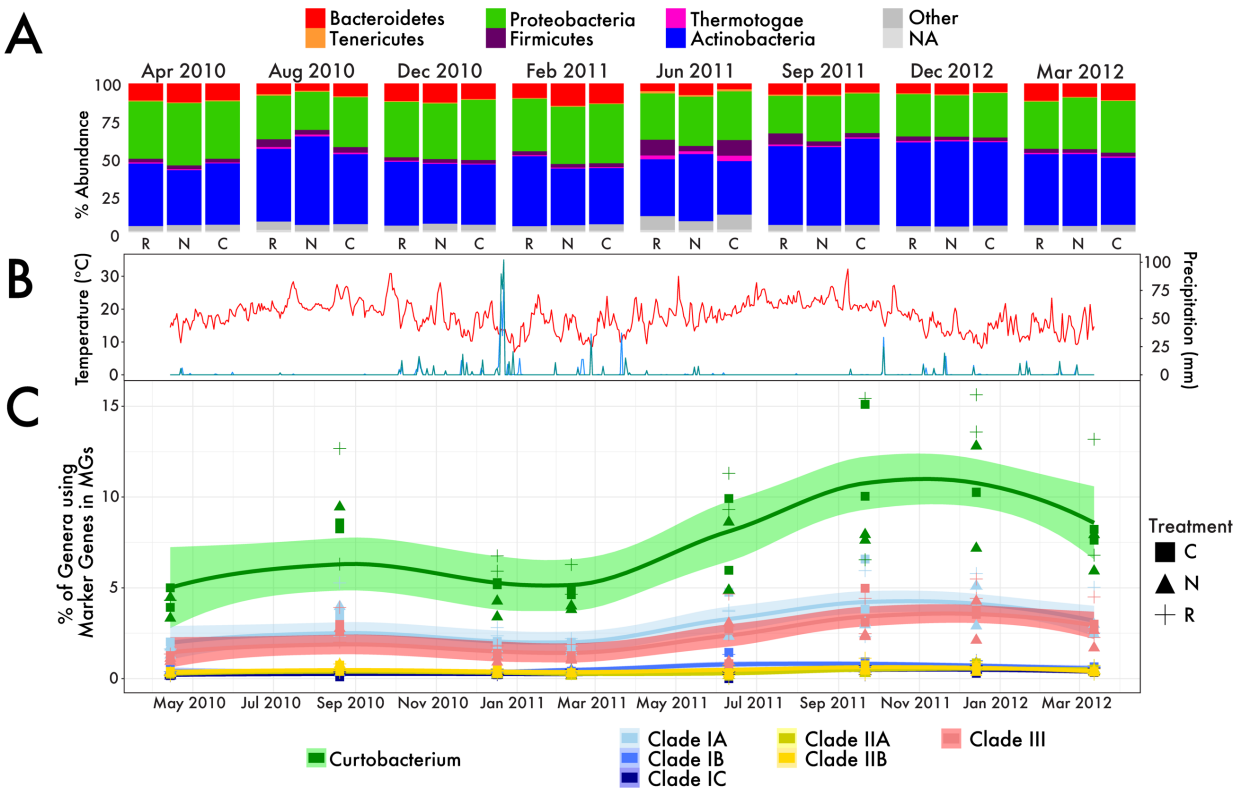


Figure 3.2 Bacterial community composition in Loma Ridge field site over two years. **A)** Relative abundances of the six most abundant phyla; replicates were averaged by each treatment (N = +nitrogen, R = reduced precipitation, C = control) and time point. **B)** Temperature and precipitation at Loma Ridge collected from May 2010 to March 2012. **C)** Relative abundance of total *Curtobacterium* (green line) and each individual clade over time and by treatment. Smoothed averages were calculated from locally weighted smoothing (LOESS) with confidence intervals.

Supporting Information

This chapter contains supporting information that can be found online at https://mbio.asm.org/content/8/6/e01809-17.full (DOI: 10.1128/mBio.01809-17).

CHAPTER 4

Emergence of soil bacterial ecotypes along a climate gradient

ORIGINALITY-SIGNIFICANCE STATEMENT

Microbial community analyses typically rely on delineating operational taxonomic units; however, a great deal of genomic and phenotypic diversity occurs within these taxonomic groupings. Previous work in closely-related marine bacteria demonstrates that fine-scale genetic variation is linked to variation in ecological niches. In this study, we present similar evidence for soil bacteria. We find that an abundant and widespread soil taxon encompasses distinct ecological populations, or ecotypes, as defined by their phenotypic traits. We further validated that differences in these soil ecotypes correspond to variation in their distribution across a regional climate gradient. Thus, there exists genomic and phenotypic diversity within this soil taxon that contributes to niche differentiation. The study further highlights the need to link fine-scale genomic diversity with trait variation to investigate the ecological and evolutionary processes governing bacterial diversity.

SUMMARY

The high diversity of soil bacteria is attributed to the spatial complexity of soil systems, where habitat heterogeneity promotes niche partitioning among bacterial taxa. This premise remains challenging to test, however, as it requires quantifying the traits of closely-related soil bacteria and relating these traits to bacterial abundances and geographic distributions. Here, we sought to investigate whether the widespread soil taxon *Curtobacterium* consists of multiple coexisting ecotypes with differential geographic distributions. We isolated *Curtobacterium* strains from six sites along a climate gradient and assayed four functional traits that may

contribute to niche partitioning in leaf litter, the top layer of soil. Our results revealed that cultured isolates separated into fine-scale genetic clusters that reflected distinct suites of phenotypic traits, denoting the existence of multiple ecotypes. We then quantified the distribution of *Curtobacterium* by analyzing metagenomic data collected across the gradient over 18 months. Six abundant ecotypes were observed with differential abundances along the gradient, suggesting fine-scale niche partitioning. However, we could not clearly explain observed geographic distributions of ecotypes by relating their traits to environmental variables. Thus, while we can resolve soil bacterial ecotypes, the traits delineating their distinct niches in the environment remains unclear.

INTRODUCTION

A focus on traits can provide a mechanistic understanding of an organism's geographic distribution (McGill, Enquist, *et al.*, 2006; Litchman and Klausmeier, 2008; Allison, 2012; Edwards *et al.*, 2013) as traits underlie an organism's response to abiotic and biotic conditions (Lavorel and Garnier, 2002). In microbial communities, the link between traits and the distribution of microbial taxa remains poorly understood (Green *et al.*, 2008; Litchman *et al.*, 2015). While whole-genome and metagenomic data provide a sense for the potential types of traits of the microorganisms within an environmental sample (Raes *et al.*, 2011), it is unclear how well this potential translates to actual phenotypic differences (Chase and Martiny, 2018). However, as in macroorganisms, the functional assessment of an ecosystem's abundant taxa is important to developing trait-based approaches to predict community and ecosystem dynamics (Lavorel and Garnier, 2002; Enquist *et al.*, 2015).

Soil systems harbor incredible microbial diversity where high habitat heterogeneity promotes niche partitioning among bacterial taxa (Ranjard and Richaume, 2001; Nannipieri *et al.*, 2003). Indeed, the biogeographic distributions of soil bacterial communities are correlated with environmental variables (e.g. pH (Fierer and Jackson, 2006; Yao *et al.*, 2011) and nutrients (Leff *et al.*, 2015)) suggesting that traits related to the response of to these variables underlie bacterial distributions. However, most studies consider these patterns at a fairly broad genetic resolution, lumping taxa based on the sequence similarity of a highly conserved 16S rRNA region, and, subsequently, mask a high degree of trait variation among distinct bacterial taxa (Chase *et al.*, 2017; Larkin and Martiny, 2017). Such variation may be important for explaining the distribution of bacterial diversity in soil. In particular, some ecologically relevant traits,

including a response to drought conditions may be shared at broad taxonomic levels (Amend *et al.*, 2016), while others, such as temperature preference (Martiny *et al.*, 2015), might be more variable. Thus, the ability to resolve the degree of trait variation requires linking genetic information with relevant phenotypic variation (McLaren and Callahan, 2018) to distinguish fine-scale niche partitioning contributing to the distribution and diversity of soil bacteria.

To illustrate the importance of trait variation among very closely related bacterial strains, we can consider the distribution of the abundant marine phototroph, *Prochlorococcus*. Strains of *Prochlorococcus* cluster into genetically distinct clades that share physiological traits, including light preference and nutrient utilization (Moore and Chisholm, 1999; Moore *et al.*, 2002). Distinct fine-scale genetic clusters corresponding to ecologically relevant phenotypes have been defined as ecotypes (Rocap *et al.*, 2003), where all strains within an ecotype occupy the same ecological niche (Cohan, 2001). Co-existing, but distinct *Prochlorococcus* ecotypes exhibit differential geographic distributions that are highly correlated with environmental variables, suggesting fine-scale niche partitioning (Moore *et al.*, 1998; Johnson, Zinser, Coe, McNulty, *et al.*, 2006). Further, the traits that underlie these correlations are relatively clear; for instance, the optimal temperature of an ecotype's growth in the laboratory corresponds to its distribution across oceanic temperature gradients (Johnson, Zinser, Coe, McNulty, *et al.*, 2006). Thus, coexisting individuals can be clustered based on genotypic and phenotypic information to provide a functional basis in driving ecotype differentiation and niche partitioning (Kent *et al.*, 2016; Delmont and Eren, 2018).

Here, we sought to determine whether soil bacteria – like marine bacteria – form distinct ecotypes that are differentially distributed. To test this idea, we focused on the

widespread soil taxon *Curtobacterium* (Chase *et al.*, 2016). In a southern California grassland, *Curtobacterium* is highly abundant in leaf litter, the top layer of soil, suggesting a potential functional role in plant decomposition and, therefore, the carbon cycle (Matulich *et al.*, 2015). In previous work, we identified the co-occurrence of multiple *Curtobacterium* clades in leaf litter, and we hypothesized that thermal adaptation might be contributing to ecological differences of co-occurring clades (Chase *et al.*, 2017). Indeed, a recent study found that northern and southern *Streptomyces* lineages differed in their thermal tolerance across a latitudinal gradient (Choudoir and Buckley, 2018). However, it remains unclear what traits differentiate *Curtobacterium* clades and if this diversity is associated with niche partitioning in the environment. Therefore, in this study, we isolated *Curtobacterium* strains from leaf litter across six locations spanning a climate gradient and assayed four functional traits (growth, biofilm formation, and depolymerization of xylan and cellulose). We further investigated the biogeographic distributions of *Curtobacterium* clades using metagenomic data collected from leaf litter across the gradient over 18 months. We hypothesized that (i) genetically-distinct *Curtobacterium* lineages share functional traits, forming distinct ecotypes; (ii) these ecotypes have differential geographic distributions across the gradient; and (iii) the distribution of ecotypes is correlated with environmental variation (e.g. soil temperature).

RESULTS

Identification of *Curtobacterium* Ecotypes

We established six sites along an elevation gradient that co-varied in precipitation and temperature (Supporting Information Table S1). We isolated *Curtobacterium* strains from leaf litter (decaying leaves that make up the topmost layer of soil) at all sites along this climate

gradient except at the coldest and wettest Subalpine site. We sequenced 56 new *Curtobacterium* genomes with an average size of 3.6 Mbp and 70.7% GC content (Supporting Information Table S2). Incorporating all known *Curtobacterium* genomic diversity into a phylogenetic analysis, we established five major clades (I – V) delineating the genus (Fig. 1). These genomes were highly diverse, sharing as low as 79.2% average nucleotide identity (ANI) and 68.9% average amino acid identity (AAI). Most (93%) of the isolates from our sites fell within Clades I, IV, and V. The major clades further diverged into finer genomic clusters (assigned to subclade designation at ≥90% AAI; Fig. 1). Each site along the gradient exhibited vast genomic diversity, with most of the subclades including isolates from multiple sites. For instance, 25 of the isolates from 4 of the sites (excluding the higher elevation sites, Pine-Oak and Subalpine) fell into Subclade IB/C. In contrast, Subclade VA was only isolated from the Desert site.

Our analyses revealed extensive genomic diversity that would otherwise be masked using traditional genetic characterizations (i.e. 16S rRNA gene similarity) of bacterial operational taxonomic units (OTUs). Analyzing the hypervariable V4/V5 region of the 16S rRNA region at 100% sequence similarity, we identified only four OTUs. All eight isolates from Subclade VA grouped together, while the majority of strains (45 of the 56 climate gradient strains), irrespective of phylogenetic relatedness, shared identical V4/V5 regions. An alignment of the full-length 16S rRNA gene region revealed some congruence between subclade diversity, but clustering into OTUs at 99% similarity (as recommended for full length sequences (Schloss, 2010; Edgar, 2018)), still revealed only two OTUs (Supporting Information Fig. S1).

We next asked whether this genomic variation corresponded to phenotypic diversity by assaying a subset of isolates for four functional traits (growth, biofilm formation, and depolymerization of xylan and cellulose). Given the stark gradient in temperature across our sites, we measured each trait at a range of temperatures to consider the response of traits to environmental variation (McGill, Enquist, *et al.*, 2006). Indeed, across all assays, temperature explained 15-20% of the trait variation observed in the lab assays (linear regressions; all $p < 0.0001$; reporting adjusted R^2). Isolates depolymerized both cellulose and xylan at varying efficiencies but strains did not discriminate between carbon substrate utilization (linear regression; $p > 0.05$). Additionally, the ability to depolymerize specific carbon substrates was not a simple product of increased growth across temperatures (Supporting Information Fig. S2), underlining the differences in carbon utilization ability across strains. Total carbon depolymerization (combining cellulose and xylan assays) varied significantly by subclade designation (analysis of covariance (ANCOVA); $F_{5,195} = 175.7$, $p < 0.0001$) with a significant interaction between temperature ($F_{15,195} = 13.6$, $p < 0.0001$; Fig. 2). Maximum growth rate (μ_{max}) and biofilm formation followed similar statistical trends (Supporting Information Table S3) with subclade and temperature significantly explaining trait performance (Supporting Information Fig. S3A and S3B, respectively). However, temperature affected traits differently; for example, carbon depolymerization and maximum growth rate peaked at 28°C while biofilm formation decreased linearly with increasing temperature (Fig. 2). Across all assays, trait performance was strongly influenced by temperature amplifying the degree of phenotypic variation among *Curtobacterium* isolates.

Together, the trait assays indicated that despite highly differential trait responses within *Curtobacterium,* isolates within the same subclade reflected similar phenotypic traits. By combining all observed trait variation, isolates clustered significantly by subclade, such that strains within the same genetic subclade shared more similar traits (analysis of similarities (ANOSIM); $R = 0.69$, $p < 0.001$; Fig. 3). Carbon depolymerization abilities and growth parameters (maximum absorbance (A_{max}), μ_{max}, and lag phase) across temperatures best distinguish these strains while biofilm formation remained highly variable (Fig. 3). All isolates could depolymerize carbon, but the efficiency of carbon depolymerization, especially at higher temperatures, strongly differentiated subclades. For example, Subclade IVA was unable to degrade either cellulose or xylan at high temperatures, whereas Subclade IVB was generally the best degrader (Fig. 2). In contrast, the broader genetic clades (i.e. Clades I – V) were indistinguishable by the measured traits at varying temperatures (ANOSIM; $R = 0.14$, $p > 0.05$; Fig. 3). Thus, the degree of trait variation within subclades was highly correlated with fine-scale genetic clusters within *Curtobacterium*, denoting the existence of distinct *Curtobacterium* ecotypes.

To consider whether there was evidence for adaptation to the site from which the strains were isolated, we also tested whether the isolation site influenced trait performance. When we considered both subclade designation and the site of isolation along with the temperature of the assay, we accounted for 52% and 87% of the variation in maximum growth rate and carbon depolymerization, respectively (ANCOVA; all $p < 0.01$; reporting adjusted R^2; Supporting Information Table S3). However, subclade designation explained more trait variation than site effects for both carbon depolymerization (ANCOVA; $\Omega^2 = 0.50$ vs. 0.02, respectively) and μ_{max} (ANCOVA; $\Omega^2 = 0.14$ vs. 0.06, respectively) across the temperature

gradient (Fig. 2; Supporting Information Table S3). Biofilm formation, conversely, was only related to subclade and temperature (ANCOVA; adjusted $R^2 = 0.29$, $p < 0.001$) despite being highly variable across strains. For example, strains from both Subclades IA and IB/C were among the best biofilm producers at lower temperatures, while other strains from the same subclade and site of isolation produced minimal biofilms (Supporting Information Fig. S3B). Although site effects contributed to observed trait variation among isolates, site effects explained little additional variation beyond subclade designations to accurately distinguish ecotypes.

Biogeography of *Curtobacterium* and its Ecotypes

To evaluate the biogeographic distribution of *Curtobacterium* ecotypes, we characterized the litter bacterial community across the climate gradient using 91 metagenomic libraries collected over 18 months. Total bacterial composition varied across the climate gradient, with Desert, Grassland, and Scrubland communities more similar in composition to one another than to Pine and Subalpine communities (Supporting Information Fig. S4A). Notably, *Acidobacteria* were common in the colder, wetter sites (Pine-Oak and Subalpine), while *Actinobacteria* dominated in the hotter, drier sites (Supporting Information Fig. S4B). Salton Sea bacterial composition was distinct from all other sites being dominated by *Proteobacteria* and the genus *Halomonas* (Supporting Information Fig. S4C). *Curtobacterium* (phylum: *Actinobacteria*) was the 6th most abundant genus across all sites and time points with an average relative abundance of 1.6% (Supporting Information Fig. S4C). Total *Curtobacterium* abundance was highest in the Grassland and decreased towards the extreme ends of the climate gradient (top line in Fig. 4A).

The geographic distribution of subclades within *Curtobacterium* – the genetic resolution at which we could distinguish ecotypes – also varied along the climate gradient (Fig. 4A). To identify subclade sequences in the metagenomic data, we extracted 830 orthologous protein groups belonging to *Curtobacterium* and *Frigoribacterium* (a sister genus) identified from the full genome sequences (see Experimental Procedures). The overall abundance of *Curtobacterium* was represented by multiple subclades, comprised primarily of six abundant ecotypes spanning the climate gradient (Fig. 4A). Nevertheless, subclade composition varied significantly by site (permutational multivariate analysis of variance (PERMANOVA; $p < 0.001$) such that some subclades were strongly correlated with site location (Fig. 4B). For example, Subclade IVB was the dominant *Curtobacterium* ecotype in the hot, dry Desert site and the cold, wet Pine-Oak and Subalpine sites; whereas, at the intermediate climate sites, Scrubland and Grassland, Subclade IB/C was the dominant ecotype (Fig. 4A). The less abundant ecotypes also exhibited differential distributions with Subclades IVB and IVC being more pronounced in the Desert and Subclade IA peaking in the Grassland (Fig. 4A,B). Along the climate gradient, we identified six abundant ecotypes co-occurring at each site with each ecotype exhibiting preferential distributions.

The relative abundance of the ecotypes remained relatively constant over the year and a half (Supporting Information Fig. S5A). Overall, the temporal effects were less pronounced than the site effects (PERMANOVA; $p > 0.05$), and accounted for only 0.8% of the observed variation in subclade composition across sites (as compared to 54.2% attributed to site effects). Therefore, *Curtobacterium* composition along the climate gradient was temporally stable over the course of the study.

Ecotype – Environmental Relationships

Since *Curtobacterium* ecotypes clearly differed in their geographic distributions, we next asked how their abundances were correlated with environmental variation. Ecotype composition varied over time, but this shift in composition was minimal relative to the site effects (Fig. 5). Indeed, the environmental factors measured at each site (leaf litter chemistry and abiotic parameters) largely explained the observed ecotype composition (distance based linear model (distLM); adjusted $R^2 = 0.91$). In particular, the proportion of hemicelluloses (e.g., xylan) in the leaf litter explained 41.5% of the variation in ecotype composition (distLM; $p < 0.05$; Supporting Information Table S4). The measured abiotic factors (precipitation and soil surface day- and night-time temperatures; Supporting Information Fig. S5B,C) explained an additional 38% of ecotype variation between sites and across seasons.

Despite identifying environmental factors related to ecotype composition, we were unable to link these patterns to the trait measurements. For instance, leaf litter from the Grassland and Scrubland sites contained the highest proportion of polymeric carbohydrates (cellulose and hemicelluloses; Supporting Information Fig. S5D). However, Subclade IVB, the most efficient degrader of cellulose and xylan (Fig. 2), was not the dominant ecotype at these sites; instead, Subclade IB/C was nearly twice as abundant as Subclade IVB (Fig. 4A). Further, subclades whose trait performance peaked at warmer temperatures in the lab assays (e.g., Subclade IA in carbon depolymerization and Subclade VA in growth parameters; Fig. 2) were more abundant at the Grassland site rather than the warmer Desert site. Thus, while the phenotypic trait measurements strongly differentiated strains into ecotypes, these traits did not explain ecotype distribution across the sites.

Finally, given the variability in litter chemistry among sites (Supporting Information Fig. S4D), we considered whether the genetic potential to utilize a range of carbohydrates might be correlated with the distribution of *Curtobacterium* ecotypes. Specifically, we targeted the genomic diversity of glycoside hydrolase (GH) and carbohydrate binding module (CBM) proteins that potentially contribute to degradation of various carbon substrates in leaf litter. Overall, the composition of GH and CBM genes was correlated with phylogenetic distance between *Curtobacterium* strains (RELATE test; $\rho = 0.43$, $p < 0.0001$) such that more phylogenetically similar genomes encoded similar GH-CBM profiles. Further, the genomic potential to degrade complex polymeric carbohydrates common in leaf litter (i.e. cellulose, chitin, and xylan) differed significantly between subclades (Kruskal-Wallis test, $p < 0.0001$; Supporting Information Fig. S6A). However, the total abundance of polymeric GH/CBM did not clearly predict ecotype distribution along the gradient. We predicted that ecotypes with higher numbers of polymeric GH-CBMs would be more abundant on leaf litter; however, two of the rarer ecotypes, Subclades IA and IVA, contained the highest total number of polymeric GH-CBMs (Supporting Information Fig. S6A). Similarly, total GH and CBM composition also varied significantly by subclade (PERMANOVA; $p < 0.001$; Supporting Information Fig. S6B), but ecotypes with highly similar GH-CBM compositions, such as Subclades IVA and IVB, differed strikingly in their association with different sites (Fig. 4B). Therefore, while the abundance and composition of GH-CBMs in *Curtobacterium* genomes supported our ecotype designations (Supporting Information Fig. S6B), variation in these functional genes did not elucidate ecotype distributions along the climate gradient.

DISCUSSION

In this study, we applied a trait-based framework (Diaz *et al.*, 1998; Dıaz and Cabido, 2001; Cadotte *et al.*, 2015) to identify ecological populations, or ecotypes, in a terrestrial bacterium and investigated the drivers of their biogeographic distribution. By sampling across a climate gradient varying in temperature and precipitation, we identified highly similar genomic clusters within *Curtobacterium* that corresponded to distinct phenotypes denoting the existence of bacterial ecotypes. These results contribute to the growing understanding that traditional taxonomic assignments (i.e. OTUs) mask bacterial "microdiversity" that contributes to ecological differentiation (Jaspers and Overmann, 2004; Larkin and Martiny, 2017). More broadly, our study highlights the application of a trait-based approach to microbial systems to assess the ecological and evolutionary mechanisms contributing to community assembly (McGill, Enquist, *et al.*, 2006; Nemergut *et al.*, 2013; Enquist *et al.*, 2015).

A growing number of studies demonstrate that the fine-scale genomic structure of bacterial diversity reflects diverged populations (Polz *et al.*, 2006; Connor *et al.*, 2010) occupying separate ecological niches (Johnson, Zinser, Coe, McNulty, *et al.*, 2006; Hunt *et al.*, 2008). Indeed, it appears that the total abundance of typically-defined taxa (i.e. OTUs based on 16S rRNA sequence similarity) may often be comprised of distinct ecological populations that vary over a range of environments (Moore *et al.*, 1998; Thompson *et al.*, 2005). This emerging pattern has implications for how we interpret biogeographic patterns of bacterial diversity. In particular, phenotypic differences among ecotypes can permit the coexistence of fine-scale genetic diversity within an environment. In this study, we identified and observed six abundant *Curtobacterium* ecotypes at all sites along our gradient, suggesting fine-scale niche partitioning

of environmental resources. In addition, ecotypic diversity may allow a taxon to persist in a broader range of environments than would be expected based on the phenotype of a single representative (Moore *et al.*, 1998; Partensky *et al.*, 1999). Therefore, the biogeographic distribution of a typical OTU or a representative strain may be not be indicative of the range of genetic and phenotypic diversity encoded at finer taxonomic levels.

Whether an isolate is representative of its broader taxon will depend on the particular trait of interest (McLaren and Callahan, 2018), as different traits vary in the degree to which they are phylogenetically conserved (Martiny *et al.*, 2013). Many traits are conserved across all *Curtobacterium* diversity including those contributing to its dominance as a leaf litter bacterium (Chase *et al.*, 2016). For instance, all strains in this study shared the ability to degrade polymeric carbohydrates common in leaf litter, cellulose and xylan, and, relative to other genera in the Microbacteriaceae family, *Curtobacterium* has a high genomic potential for carbohydrate degradation (as assessed by the total number of GH-CBM genes) (Chase *et al.*, 2016). Additionally, the taxon generally appears to prefer relatively dry surface soil conditions (Lennon *et al.*, 2012) as the relative abundance of *Curtobacterium* as a whole tends to increase in drier seasons (Chase *et al.*, 2017). In contrast, traits that vary within the genus will contribute to ecological differences amongst ecotypes. Such fine scale trait variation may often result in quantitative rather that qualitative differences. For example, *Curtobacterium* ecotypes varied in their growth rates at different temperatures. And while all *Curtobacterium* ecotypes could degrade cellulose and xylan, the degree to which they degraded these compounds in the lab varied significantly.

Although we were able to identify correlations between ecotype composition and environmental factors across the sites, it was not clear which traits underlie *Curtobacterium* ecotype distributions as has been resolved for marine bacteria (Johnson, Zinser, Coe, McNulty, *et al.*, 2006; Martiny *et al.*, 2009). There are several possible reasons for this disconnect. One possibility is that we did not measure the correct traits. For example, we hypothesized that the ability to form biofilms might be important because biofilms can protect bacteria from desiccation and fluctuations in water potential (Hartel and Alexander, 1986; Roberson and Firestone, 1992) and are correlated with soil moisture adaptation (Lennon *et al.*, 2012). Thus, we expected biofilm formation to be prevalent across all *Curtobacterium* strains, especially with higher production in strains abundant at drier sites. However, biofilm formation was highly variable among strains, so much so that subclade differences explained little observed variation and there was no effect from the site of isolation. Of course, biofilm formation is just one trait that might contribute to moisture adaptation (Potts, 1994) and other traits related to moisture preference might be more predictive for assessing fine-scale niche partitioning. We also did not measure a variety of traits that known to be important to soil bacteria including nutrient uptake abilities and pH preferences (Fierer and Jackson, 2006; Leff *et al.*, 2015). Environmental constraints clearly contribute to the distribution of soil bacterial taxon, however, the traits delineating these biogeographic patterns requires further investigation.

A second reason that we may have missed ecotype-environment correlations is that we are not measuring the environment at the correct spatiotemporal scale. Soils are highly heterogeneous and differences in soil microhabitats are thought to contribute to the maintenance of soil diversity (Ranjard and Richaume, 2001; Nannipieri *et al.*, 2003).

Consequently, soil ecotypes are likely to respond to environmental variation at very small spatial scales. The existence of multiple *Curtobacterium* ecotypes co-occurring within a given site suggest that fine-scale environmental variation is contributing to niche partitioning in leaf litter. Indeed, even in the marine water column, which is thought to be more homogeneous than soil, strains of *Vibrio splendidus* partition resources to differentiate between particle-associated or free-living habitats (Hunt *et al.*, 2008). On a similar spatial scale, variation in hemicellulose availability or temperature within a decomposing leaf may explain the coexistence of multiple *Curtobacterium* ecotypes. Thus, by sampling across a regional climate gradient, we may have masked much of the within-site environmental variation that contributes to soil ecotype distributions. A further possibility is that *Curtobacterium* diversity is not at equilibrium in the sampled communities. Maladapted strains may be present and even abundant if environmental selection is weak and/or dispersal is high (Lenormand, 2002). Much more work is needed to understand the spatiotemporal scales of these mechanisms for soil bacterial diversity.

In sum, our study presents evidence that the genomic diversity within an abundant terrestrial bacterial taxon can be classified into ecotypes that vary in their biogeographic distribution across a climate gradient. Especially for terrestrial soil communities, we lack an understanding of the ecological and evolutionary processes governing the distribution and functioning of bacterial diversity. The results presented here are consistent with the growing understanding that fine-scale genomic diversity, and the traits encoded by this variation, is key to microbial biogeography. However, identifying and measuring relevant traits remains a distinct challenge for the application of trait-based frameworks to microbial communities.

EXPERIMENTAL PROCEDURES

Field Sites

We characterized the microbial community on leaf litter by establishing four replicate plots (1 m^2) at six sites across a climate gradient in southern California from October 2015 to April 2017 (Glassman et al. In prep.). The five sites (from lowest to highest elevation) include the Sonoran desert (33.652 N, 116.372 W), pinyon-juniper scrubland (33.605 N, 116.455 W), coastal grassland (33.737 N, 117.695 W), pine-oak forest (33.808 N, 116.772 W), and subalpine forest (33.824 N, 116.755 W) as previously described in (Baker and Allison, 2017). In addition, we sampled leaf litter near the highly-saline Salton Sea (33.518 N, 115.938 W) to extend the climate gradient further (Supporting Information Table S1). Sites are hereafter referred to as Desert, Scrubland, Grassland, Pine-Oak, Subalpine, and Salton Sea, respectively. All sites experience Mediterranean climate patterns with a hot, dry summer and a cool, wet winter. The sites range in mean annual air temperature (MAT) from 10.3-24.6°C and precipitation (MAP) from 80-400mm. To characterize climate at the sites during the experiment, we collated precipitation data from nearby weather stations and collected surface soil temperature at 90 min intervals using two iButton temperature sensors (Maxim Integrated) from April 4th, 2016 to April 20th, 2017 at five of the sites (excluding Salton Sea) (Glassman et al. In prep.). In addition to changes in temperature and precipitation, the sites differed greatly in the plant communities present and, therefore, the litter chemistry. Leaf litter chemistry was determined from samples in both the dry (June) and wet (December) seasons in 2015 using near-IR spectroscopy, as previously described (Baker and Allison, 2017).

Isolation and Genomic Characterization of *Curtobacterium*

To isolate *Curtobacterium* strains, we collected fresh leaf litter from the perimeter of the four plots at each site on June 14th, 2016 to create a homogenized batch of litter from each site. We ground the litter in a sterile coffee grinder and vortexed 0.2 g of homogenized litter in 5 mL of 0.9% saline (NaCl) solution for 5 min. Samples were serially diluted and plated on grassland leaf litter leachate media (Chase *et al.*, 2016). Colonies were visually screened for phenotypic characteristics ascribed to *Curtobacterium* (Evtushenko and Takeuchi, 2006), streaked on Luria Broth (LB) media agar plates, transferred three times, and stored in glycerol solution at -80°C. We identified each cultured isolate by PCR amplification and Sanger sequencing of a 1500 bp region of the 16S rRNA region. For each isolate, we used DNA extracted from a single colony that we added to a PCR cocktail containing 0.3 µL HotMaster Taq polymerase (5 units/µL), 15 µL 2x Premix F (Epicentre; Madison, WI), and 0.2 µL of 50 µM of each primer, pA (5'-AGAGTTTGATCCTGGCTCAG-3') and pH' (5'-AAGGAGGTGATCCAGCCGCA-3'), under identical PCR conditions (Chase *et al.*, 2016). The 16S rRNA sequence of each isolate was used to identify taxonomy using the Ribosomal Database Project (RDP) database (Wang *et al.*, 2007).

Identified *Curtobacterium* isolates were selected for whole-genome sequencing and grown on LB plates for 48-72 hrs. A single colony from each plate was transferred to 10 mL liquid LB media to grow for an additional 48 hrs. Genomic DNA extraction was performed using the Wizard Genomic DNA Purification Kit (Promega; Madison, WI) with the additional step of adding lysozyme for Gram-positive bacteria. Extracted DNA was quantified on the Qubit (BioTek; Winooski, VT), quality assessed on the Nanodrop (Thermo Fisher; Waltham, MA), and diluted to 0.5 ng/µL for library preparation. Next, we followed the protocol for the Nextera XT

Library Preparation kit (Illumina Inc., San Diego, CA, USA). Samples were pooled in equimolar portions and assessed using the High Sensitivity Bioanalyzer. The pooled library was sequenced using an Illumina HiSeq4000 instrument (Illumina Inc., San Diego, CA, USA) with 150 bp paired-end reads. Demultiplexed sequence data were assembled using the SPAdes genome assembler (Bankevich *et al.*, 2012) with a "careful" iterative k-step ranging from k=31 to 111. We assessed the quality of the assemblies by creating taxon-annotated-GC-coverage (TAGC) plots. Specifically, we calculated coverage for each contig by mapping back the raw sequence data to assembled contigs using Bowtie2 (Langmead and Salzberg, 2012) and taxonomic assignments were assigned using MegaBLAST against the NCBI nucleotide database (Federhen, 2012) with an E value of 1×10^{-5}. Based on the results from the TAGC-plots, we discarded all contigs with coverage <30, length <500 bp, and GC% <55%. In total, we identified 56 high-quality *Curtobacterium* genomes to be included in this study, which are deposited at GenBank under BioProject PRJNA391502 with biosamples SAMN09009025 – SAMN09009080.

We created a *Curtobacterium* phylogeny using a multi-locus sequence alignment (MLSA) of 21 single-copy marker genes (Wu *et al.*, 2013). For comparison of the climate gradient genomes (N=56), we downloaded all publicly available *Curtobacterium* genomes (N=30) and a *Frigoribacterium* genome (to serve as an outgroup), which included 14 previously identified *Curtobacterium* isolates from our previous work in leaf litter (Chase *et al.*, 2016, 2017). Each of the 87 genomes were translated using Prodigal (Hyatt *et al.*, 2010) and screened for the presence of the 21 marker genes using HMMER v3.1b2 (Finn *et al.*, 2011) with an E value of 1×10^{-10}. Each marker gene was independently aligned using ClustalO v1.2.0 (Sievers *et al.*, 2011) to create a concatenated protein alignment consisting of 3947 amino acids for phylogenetic

analysis using RAxML v8.0.0 (Stamatakis, 2014) under the PROTGAMMAWAG model for 100 replicates. We designated the major branching points in the resulting phylogeny into five distinct clades. To identify finer taxonomic groupings, we calculated pairwise average amino acid identity (AAI) and nucleotide identity (ANI) across all 87 genomes using the enveomics package (Rodriguez-R and Konstantinidis, 2016). Genomes that clustered at ≥90% AAI at the whole genome level, the suggested boundaries for bacterial species groupings (Richter and Rosselló-Móra, 2009), were further designated into subclades. Subclade designations were also supported by the phylogeny.

To cluster genomes into operational taxonomic units (OTUs), we extracted 16S rRNA gene sequences from the full genomes using Barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml) and conducted two analyses recommended for optimal assessment of taxnomomic units (Edgar, 2018). First, we extracted the hypervariable V4/V5 region of the 16S rRNA gene and defined OTUs at 100% gene similarity with UCLUST (Edgar, 2010), also termed zero-radius OTUs (zOTUs) or exact sequence variants (ESVs). To include effects of alignment quality (Schloss, 2010), we aligned the full-length 16S rRNA gene region with SINA (Pruesse *et al.*, 2012) then clustered at 99% gene similarity with mother (Schloss *et al.*, 2009). We conducted a phylogenetic analysis of the full-length, aligned 16S rRNA gene region using RAxML v8.0.0 (Stamatakis, 2014) under the GTRGAMMA model for 100 replicates.

We characterized the functional potential to degrade carbohydrates (glycoside hydrolase (GH) and carbohydrate binding module (CBM) proteins) within all *Curtobacterium* genomes. The predicted open reading frames generated from Prodigal were searched using

HMMER against the Pfam-A v30.0 database (Finn *et al.*, 2016). GH and CBM genes and their targeted substrate were identified according to the Pfam identifiers as stated in (Chase *et al.*, 2016). Total GH and CBM gene composition profiles for each genome were normalized and used to construct a Euclidean distance matrix for producing an ordination plot.

Characterization of *Curtobacterium* Traits

In the laboratory, we characterized the traits of a subset of the *Curtobacterium* isolates spanning across the climate gradient and phylogenetic clades. Specifically, we sought to measure four functional traits (growth, biofilm formation, and depolymerization of cellulose and xylan) across a temperature range (15-42°C) experienced along the climate gradient. We selected these traits because we speculated that they would influence competitive dynamics in the leaf litter community. The ability to degrade polymeric carbohydrates and, specifically, an increased degradation efficiency should provide a competitive advantage as the primary carbon supply in leaf litter is in the form of celluloses and hemicelluloses (e.g. xylan) (Baker and Allison, 2017). Our sites experience long periods without precipitation and, therefore, the ability to form biofilms may prevent desiccation from water stress (Lennon *et al.*, 2012). Increased growth, both in response (lag phase) and rate (μ_{max}), could allow for competitive exclusion of other organisms. Traits were assayed along the temperature gradient to simulate abiotic conditions from the climate gradient.

For all assays, a subset of *Curtobacterium* strains and one *Escherichia coli* strain (as a control) were grown from -80°C freezer stocks for 48 hrs in liquid LB media at 22°C. Isolates were pelleted by spinning down at 4500 RPM for 10 min, washed three times with 0.9% saline solution to remove residual media, and resuspended in 10 mL of M63 minimal media

(supplemented with 0.1% peptone and 1 µg/mL thiamine) with 0.5% (wt/vol) dextrose as the sole carbon source. After 24 hrs, isolates were washed again under identical conditions and diluted to an optical density of 0.1 OD_{600} to ensure equal cell density across all isolates.

For the growth rate and biofilm assays, we inoculated 10 µL of diluted isolates (N=29 *Curtobacterium* isolates) into 96-well plates containing 190 µL of M63 media with 0.5% (wt/vol) dextrose. Each strain was grown in triplicate on each plate for each assay. The inoculated plates for the growth rate assays were shaken at 200 RPM at four temperatures (15, 25, 28, and 37°C) with OD_{600} being measured every 1-2 hrs for the first 48 hrs and every 4 hrs thereafter. Sampling was terminated if any of the six negative controls in any plate increased in OD_{600} measurements over the course of the experiment. To estimate growth parameters (max absorbance (A_{max}), max growth rate (μ_{max}), and lag phase), we fit OD_{600} measurements to either a logistic, gompertz, or a locally weighted scatterplot (LOESS) regression model using the "growthcurve" package in the R software environment (Pinheiro *et al.*, 2011). For biofilm assays, inoculated plates were sealed and placed in incubators at six temperatures (15, 22, 25, 28, 34, and 37°C) without shaking. After 4 days, we removed residual cells and media by submerging the microplates in deionized water. We then added 125 µL of 0.1% crystal violet solution to each well and incubated the plates for 15 min at room temperature. Plates were re-submerged in water and vigorously shaken to remove residual liquid (repeated 4x). We dried each plate for 2 hrs and added 125 µL of 30% acetic acid to solubilize the crystal violet. Plates were incubated for 15 min and absorbance was measured at OD_{550} for biofilm production (O'Toole, 2011).

We selected a subset of *Curtobacterium* isolates (N=18) from various sites along the climate gradient to assess their ability to depolymerize cellulose and xylan as previously described (Chase *et al.*, 2017). Briefly, we inoculated 10 µL of washed 0.1 OD_{600} cultures, in triplicate, onto solid M63 media with 0.5% (wt/vol) of either carboxymethyl cellulose (CMC) (catalog no. 150560; MP Biomedicals) or xylan (catalog no. X0502; Sigma) and placed inoculated plates in incubators at seven temperatures (15, 22, 25, 28, 34, 37, and 42°C). We classified the zones of depolymerization after 4 days using ImageJ (https://imagej.nih.gov/ij/) by subtracting the original colony area from the total area of carbohydrate degradation. An *E. coli* strain was included as a negative control and did not depolymerize either substrate at any temperature. No strains could depolymerize either substrate at 42°C and, therefore, were removed from statistical analyses.

Metagenomic Sequencing and Analysis

Metagenomic Samples

We sampled leaf litter from the 4 replicate plots at each site every 6 months until April 20th, 2017 (6 sites x 4 time points x 4 replicate plots). We extracted DNA from 0.05 g of ground leaf litter using the FastDNA SPIN Kit for Soil (Mo Bio; Carlsbad, CA) and cleaned the DNA with the Genomic DNA Clean and Concentrator kit (Zymo Research; Irvine, CA). Cleaned samples were diluted to 0.5 ng/µL and 1 ng of DNA was used for input for the Nextera XT library Prep kit for sequencing on the Illumina HiSeq 4000 instrument with 150 bp paired end reads. Due to low quality sequence data, we excluded 5 libraries and, in total, analyzed 91 metagenomic libraries. The raw data is deposited on the metagenomics analysis server (MG-RAST) (Meyer *et al.*, 2008) under the project ID mgp17355.

Bacterial Community Analysis

To characterize the bacterial litter community, we built upon our previous pipeline (Chase *et al.*, 2017) using phylogenetic inference to characterize conserved single-copy marker genes (Wu *et al.*, 2013) within the metagenomic data. To compensate for the lack of genomic representation of soil microbes, we downloaded 7,392 publicly available genomes that are designated as "representative" genomes by the PATRIC database (Wattam *et al.*, 2014) and included representative *Curtobacterium* genomes from the climate gradient (see above). We translated all genomes using Prodigal (Hyatt *et al.*, 2010) and searched for the presence of 21 single-copy marker genes using HMMER v3.1b2 (Finn *et al.*, 2011) with an E value of 1×10^{-10}. Each protein was individually aligned with ClustalO v.1.2.0 (Sievers *et al.*, 2011) and used to create a 12,271 amino acid concatenated alignment for phylogenetic analysis using FastTree2 (Price *et al.*, 2010). The reference tree was manually curated for the misplacement of genomes based on assigned nomenclature, and a genome was removed if it did not group within its assigned family designation. This highly curated tree served as a reference tree to guide construction of each individual marker gene tree using RAxML v.8.0.0 (Stamatakis, 2014) under the PROTGAMMAWAG model for 100 replicates. The remaining 5,433 genomes were used to construct BLASTp (Altschul *et al.*, 1997) databases, HMMER profiles, and pplacer (Matsen *et al.*, 2010) reference packages (all databases available at https://github.com/alex-b-chase/elevation-community).

Metagenomic libraries were quality trimmed using BBMap (Bushnell, 2016) and filtered to remove eukaryotic DNA. Specifically, we mapped all reads to a reference genome using BWA (Li and Durbin, 2009) from both an abundant grass (*Lolium perenne*; Accession:

MEHO01000000) and fungus (*Pyrenophora* teres; Accession: NZ_AEEY00000000) found at the grassland site. All filtered reads were then merged using BBMap (Bushnell, 2016) to form paired-end reads. If a read could not be merged with its counterpart, we included only the forward read in further analyses. Reads were then translated using Prodigal (Hyatt *et al.*, 2010) with the metagenomic flag and searched against the reference marker gene databases, as previously described (Chase *et al.*, 2017). Briefly, we imposed a primary filter against the reference BLASTp database with an E value of 1×10^{-5} and a secondary filter against the reference HMMER profiles with an E value ranging from 1×10^{-10} to 1×10^{-25} depending on the individual marker gene. Passed reads were aligned using ClustalO v.1.2.0 (Sievers *et al.*, 2011) to the corresponding reference package and placed onto the reference phylogenies using pplacer v.1.1.alpha17 (Matsen *et al.*, 2010). Relative abundances were calculated by generating single branch abundance matrices and normalizing to the total number of marker genes present in each library.

Curtobacterium Ecotype Abundances

The above analyses provided an estimate of the total abundance of *Curtobacterium* and other taxa in the metagenomic libraries. However, to investigate the distribution of diversity within *Curtobacterium*, we first characterized *Curtobacterium* orthologous protein groups (orthologs) from the *Curtobacterium* genomes isolated from leaf litter. Publicly available *Frigoribacterium* genomes (N=5) were also included to serve as outgroups. Orthologs were identified using Roary (Page *et al.*, 2015) with coding regions predicted by Prokka (Seemann, 2014). Due to the diversity of these genomes, we decreased the percentage sequence identity to 50% to encompass all possible orthologs. The resulting 1075 orthologs were used to create a

core-genome tree, using RAxML v8.0.0 (Stamatakis, 2014) under the PROTGAMMAWAG model for 100 replicates, that was nearly identical to the reference tree derived from the genomic MLSA analysis. We built individual ortholog trees, using identical model parameters, with the core-genome tree as the guiding reference tree, to generate a *Curtobacterium* reference database (reference database can be found here: https://github.com/alex-b-chase/elevation-curto). We then removed orthologs that lacked a robust phylogenetic signal yielding a final set of 830 orthologs. We parsed the filtered metagenomic reads for the presence of each ortholog with a BLASTp E value of 1×10^{-20} and a secondary filter against the reference HMMER profiles with an E value of 1×10^{-40}. Each filtered metagenomic read was then placed onto the corresponding ortholog tree with pplacer v.1.1.alpha17 (Matsen *et al.*, 2010) and classified to each clade and subclade. Clade and subclade relative abundances were normalized by the total abundance of *Curtobacterium* calculated from the community analyses above. For the remainder of the subclade compositional analyses, subclades were treated as the proportion to all *Curtobacterium*, not the entire community, to limit compositional biases.

Statistical Analyses

Ecotype Identification – Linking Traits to Phylogeny

To tease apart the relative importance of isolation source (where the *Curtobacterium* strain was isolated from along the climate gradient) and phylogenetic relatedness (subclade designation) for each of our physiological assays, we implemented a statistical model with site of isolation and subclade designation as dependent variables with temperature as a covariate. To start, we examined various regression models to test the best model fit using Bayesian information criterion (BIC) to confirm that temperature covaried with the other variables across

all assays. For each assay, we then determined whether our regression models should be either linear or polynomial by comparing both BIC and residual values for each model. We constructed a linear regression model for biofilm formation and polynomial regression models for carbon degradation and growth rate. Finally, we used an analysis of covariance (ANCOVA) to test the effects of our main fixed factors, site and subclade, while controlling for the effects of the covariate, temperature. Within each ANCOVA design, we implemented a backward selection process (Mac Nally, 2002) to eliminate spurious relationships (Harrell, 2015) for each assay.

To further examine the physiological differences between *Curtobacterium* subclades, we constructed a nonmetric multidimensional scales (NMDS) ordination plot of each strain using the physiological measurements. Specifically, we included biofilm formation (at 6 temperatures), cellulose degradation (6 temps.), xylan degradation (6 temps.), A_{max} (4 temps.), μ_{max} (4 temps.), and lag phase (4 temps.). All variables were normalized by subtracting the mean from each measurement and dividing by the standard deviation. Before performing the NMDS analysis, we generated Spearman's correlation coefficients (ρ^2) for each physiological assay and clustered variables into groups when $\rho^2 > 0.6$. We kept one representative trait for each Spearman-defined cluster and generated a Euclidean similarity matrix across strains. Next, we fitted each physiological variable onto the ordination plot and calculated the significance of each variable over 9,999 permutations. Finally, we removed nonsignificant variables to reduce spurious relationships (Harrell, 2015) and reran all analyses. We report only the ordination plot generated from the remaining significant variables for each strain. The significance of strain groupings was assessed using an analysis of similarities (ANOSIM) for subclade or clade

designation and site of isolation for 9,999 permutations. All analyses were performed in the R software environment.

Ecotype Distributions Along the Climate Gradient

To test the effects of site on the distribution of *Curtobacterium* subclade composition, we used a permutational multivariate analysis of variance (PERMANOVA) (Clarke, 1993). The statistical model included the site along the climate gradient and season (wet or dry) as fixed effects. We generated a Bray-Curtis similarity matrix to run a type III partial sum of squares for 9,999 permutations of residuals under a reduced PERMANOVA model. The Bray-Curtis matrix was also used to generate principal coordinates analysis (PCO) ordination plot. To assess the effects of the abiotic environment (surface soil day- and night-time temperature and total precipitation) and leaf litter chemistry (i.e. cellulose and hemicellulose) on subclade composition, we applied a distance based linear model (distLM). Again, the Bray-Curtis matrix for subclade composition was analyzed using a step-wise forward procedure with adjusted R^2 as the model selection criterion. All multivariate statistical analyses were conducted using PRIMER6 with the PERMANOVA+ function (Primer-E Ltd., Ivybridge, UK).

ACKNOWLEDGEMENTS

We would like to thank Claudia Weihe for all her work organizing the climate gradient project. We would also like to thank Chamee Moua for her assistance in isolating strains, Brandon Gaut for his advisement on data analysis, Alyssa Kent and Andrew Oliver for computational assistance, and Gavin Lear and the Martiny lab for their helpful comments on earlier revisions. We also thank Michaeline Albright, Nameer Baker, Sydney Glassman, Mike Goulden, and Kathleen Treseder for their assistance in data collection. And finally, we would

like to thank the High-Performance Computing Cluster at UCI, specifically Harry Mangalam and Joseph Farran, for their continued help and support for computational resources. This work was supported by an US Department of Education Graduate Assistance of National Need (GAANN) fellowship to ABC, a UC MEXUS-CONACYT post-doctoral research fellowship to ZGL, a National Institutes of Health Maximizing Access to Research Careers (MARC) grant (GM-69337) to AEL, a National Science Foundation grant (DEB-1457160), and US Department of Energy Office of Science Biological and Environmental Research awards (DE-PS02-09ER09-25 and DE-SC0016410).

Figures

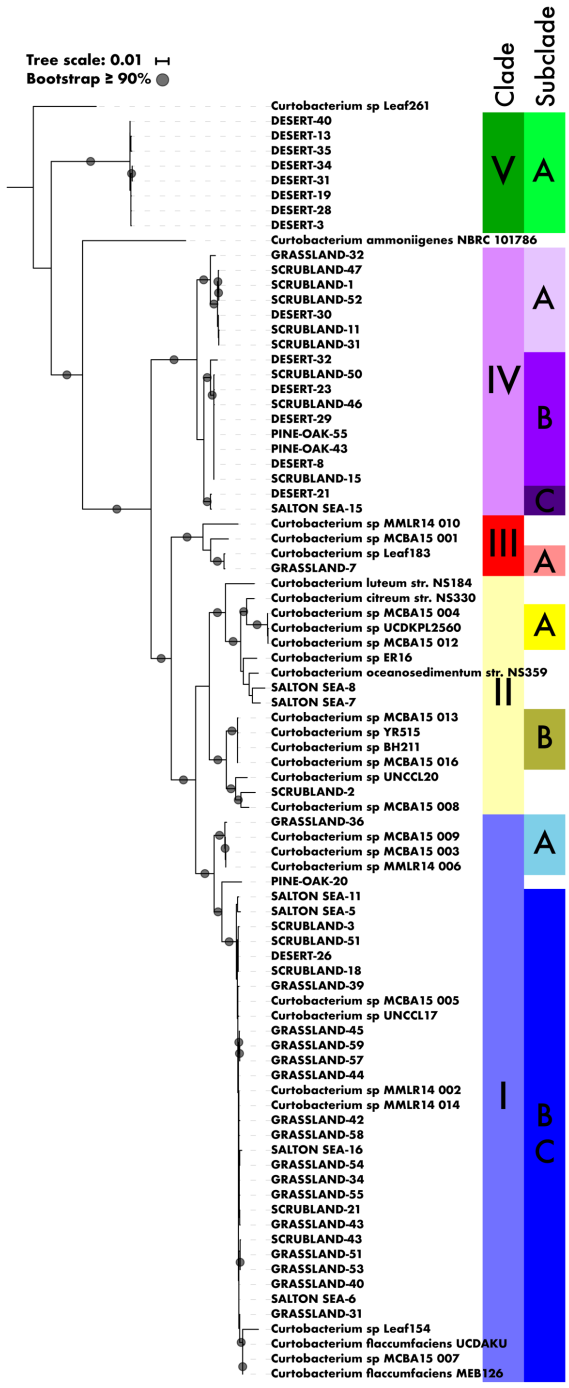


Figure 4.1 Phylogeny of *Curtobacterium* clades and subclades constructed from a multilocus phylogenetic analysis of 21 single-copy marker genes. Subclade designations are assigned if all genomes within subclade share ≥ 90% average amino acid identity. Strains with assigned nomenclature beginning with "Curtobacterium" are publicly available genomes, while the other labels designate the site of isolation along the climate gradient.

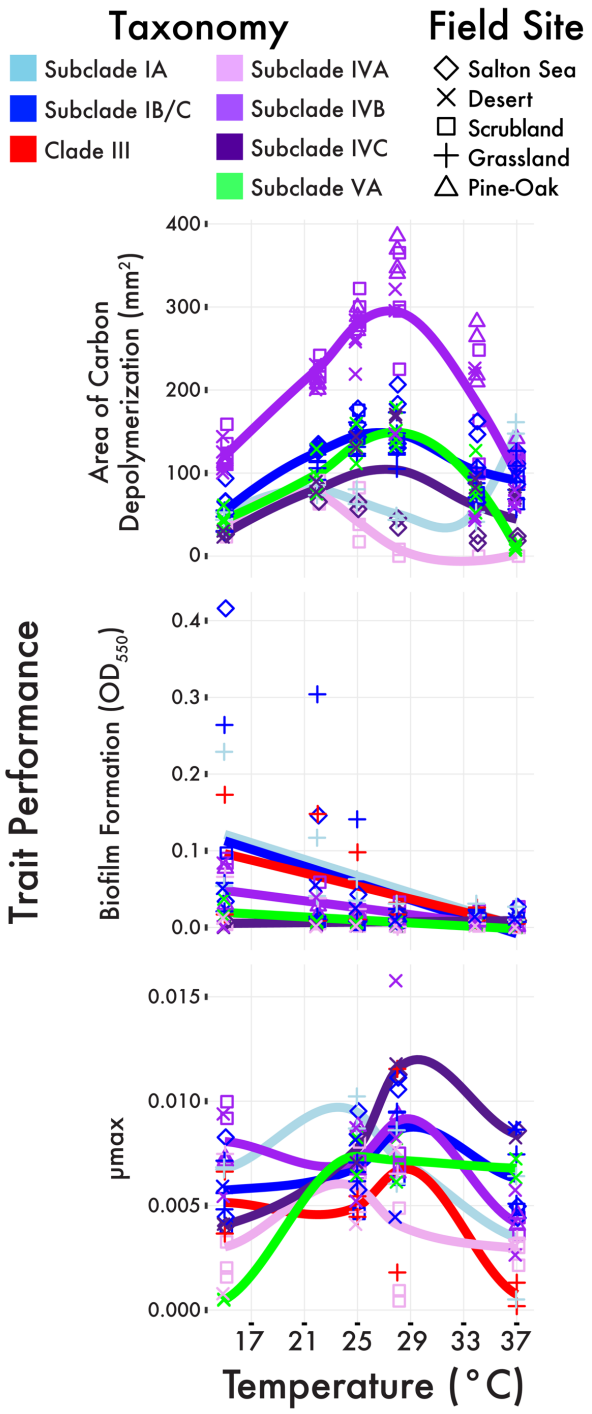


Figure 4.2 Physiological response curves plotting functional traits (carbon depolymerization, biofilm formation, and maximum growth rate, μ_{max}) versus temperature. The colors distinguish clade or subclade designation. The symbols represent the isolation site of each strain. Smoothed averages (lines) were calculated from locally weighted smoothing (LOESS) using either a polynomial (carbon and μ_{max}) or linear regression (biofilm).

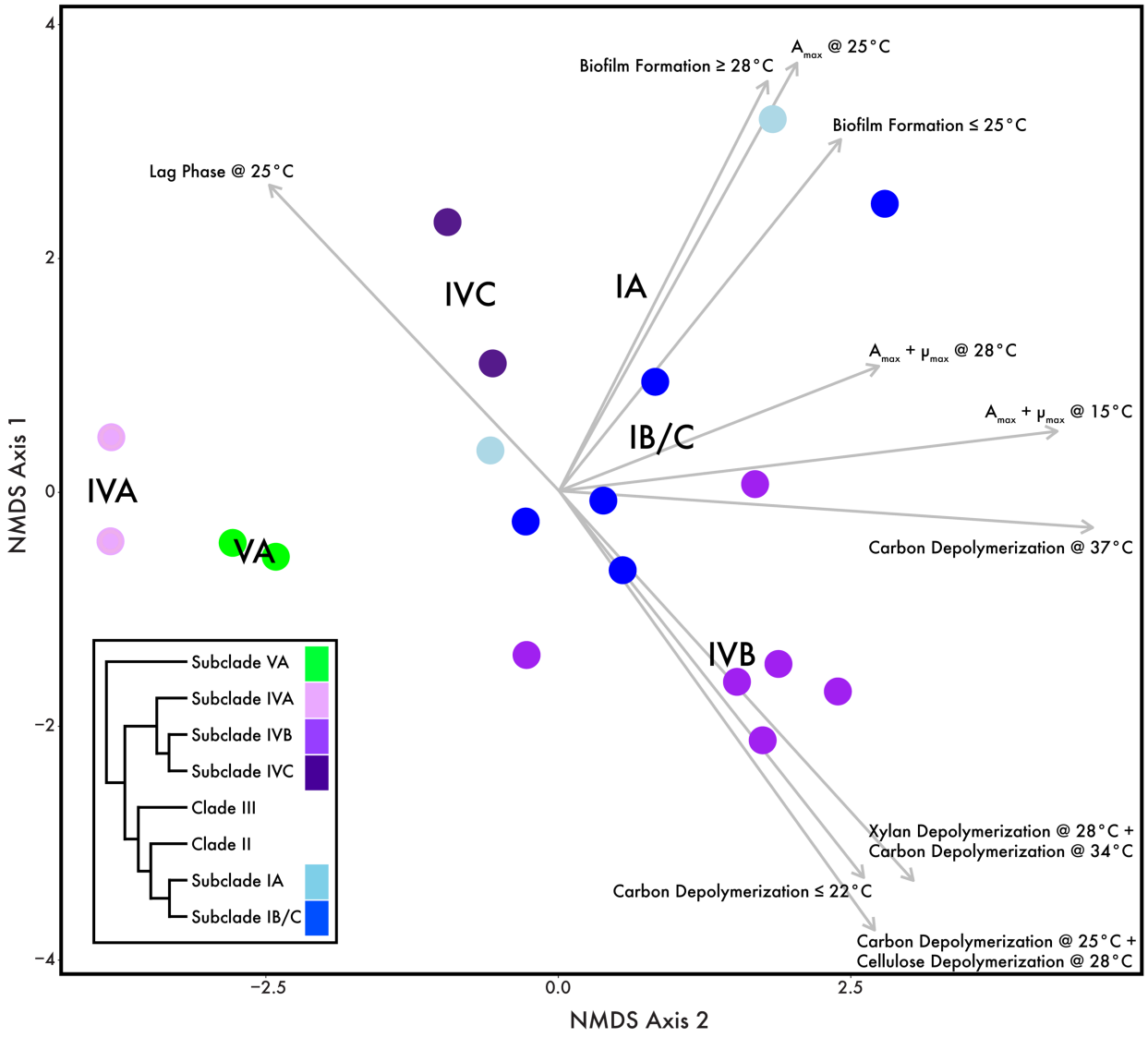


Figure 4.3 Non-metric multidimensional scaling (NMDS) plot (Euclidean distance) depicting physiological variables correlated with variation in *Curtobacterium* isolates. Variables assigned as "carbon" are collapsed to include both cellulose and xylan degradation. Each point represents an individual strain colored by subclade. Insert is a cladogram of *Curtobacterium* clades and subclades.

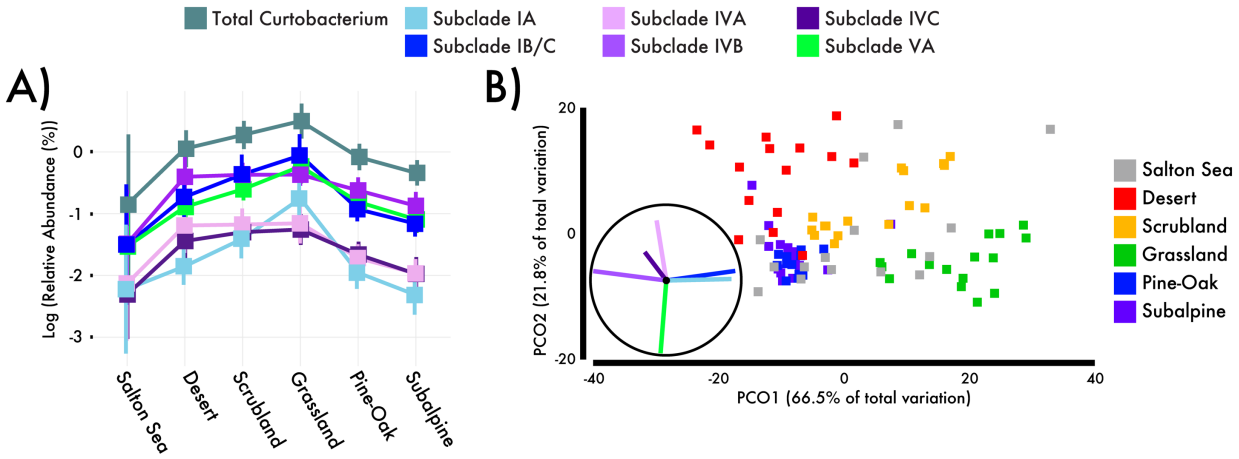


Figure 4.4 The composition of *Curtobacterium* ecotypes along the climate gradient. **(A)** Log_{10} of the mean relative abundances of *Curtobacterium* and its subclades (± 1 SD) with respect to the entire bacterial community by site. **(B)** Principal coordinates (PCO) ordination plot depicting the ecotype composition in each metagenomic sample, colored by site. Spearman correlation vectors illustrate the contribution of each subclade to compositional differences along the two PCO axes.

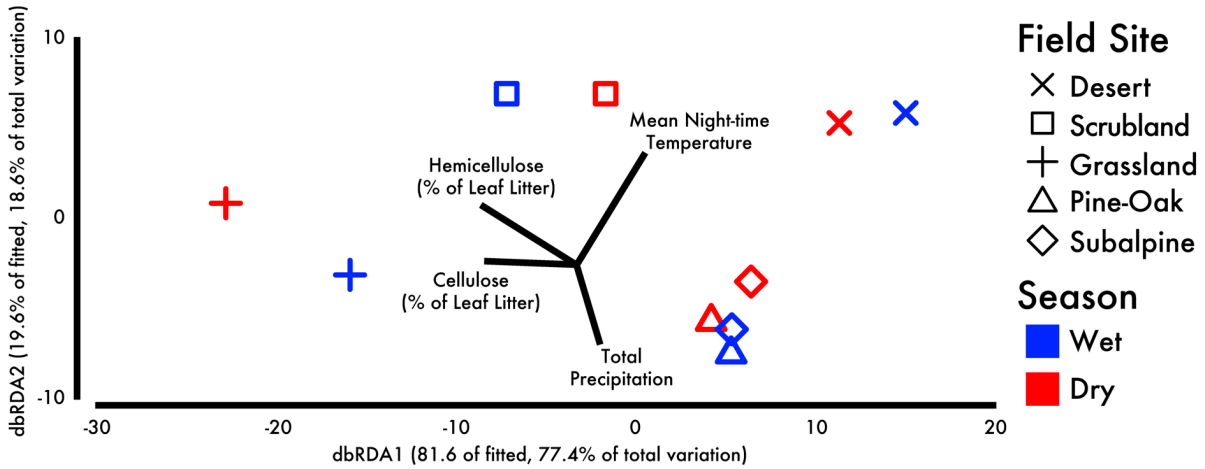


Figure 4.5 Results of a distance-based redundancy analysis (dbRDA) showing ecotype composition of a sample by site (symbols) and season (colors). Ecotype compositions were averaged across seasons with abiotic environmental variables averaged over the entire month of sampling. Vectors represent the direction and strength of correlations with environmental variables.

Supporting Information

This chapter contains supporting information that can be found online at https://doi.org/10.1111/1462-2920.14405.

CHAPTER 5

Gene flow delineates population structure in a terrestrial bacterium

INTRODUCTION

Evolutionary biologists study the genetic variation within and among populations to investigate the processes contributing to diversification. In eukaryotes, populations are typically defined as groups of interbreeding individuals within a species residing in the same geographic area (Mayr, 2001). Geographically-distinct populations are also often genetically distinct because of reduced gene flow, or the exchange of genetic variation, between populations. In microorganisms, specifically free-living bacteria and archaea, the equivalent of the biological species concept does not exist, which has created several barriers to the study of the fine-scale genetic structure within and between microbial populations (Shapiro *et al.*, 2016; Chase and Martiny, 2018).

One obstacle is that the genetic resolution delineating a microbial population is unclear. In eukaryotes, populations are, by definition, genetic units belonging to the same species. Of course, the definition of a prokaryotic species has its own unresolved challenges (Ward *et al.*, 2008). Nonetheless, there is evidence for geographically-distinct, genetically-diverged groups of bacteria and archaea. For instance, several studies have shown that the genetic similarity of closely-related microbial individuals are negatively correlated with geographic distance across continental and global scales (Whitaker *et al.*, 2003; Johnson, Zinser, Coe, Mcnulty, *et al.*, 2006; Andam *et al.*, 2016; Choudoir *et al.*, 2016). This pattern is consistent with isolation-by-distance, whereby dispersal limitation contributes to reproductive isolation over geographic distances (Wright, 1943). Further, in some cases, these geographically-localized genetic clades appear to

be adapted to local environmental conditions, as individuals within these clades differ in their temperature (Choudoir and Buckley, 2018) or habitat preferences (VanInsberghe *et al.*, 2015). However, the degree of divergence between genetic clades in such studies is quite high (<90% whole-genome ANI), indicating they may not account for intra-species relationships (Jain *et al.*, 2017). Therefore, these genetic units would seem to be much broader than the idea of a eukaryotic population, defined as a group of individuals with the potential for ecological interaction and exchange of genetic material (Cordero and Polz, 2014).

A second, related problem to studying microbial populations is assessing the genetic diversity of many microbial individuals of the same species, however defined. Indeed, population genetics in eukaryotes typically characterize the genetic diversity among many individuals from a variety of geographic locations. This sampling design for microbes would require reliable isolation of closely-related strains (but see (Kashtan *et al.*, 2014)), which can be difficult in highly diverse microbial communities (e.g. soil). Studies that do sample many microbial individuals from the same geographic location have found several co-occurring, but distinct genetic clades (Cohan, 2001; Whitaker *et al.*, 2005; Hunt *et al.*, 2008; Chase *et al.*, 2017). For instance, the thermophilic archaeon, *Sulfolobus*, demonstrated lower levels of recombination between sympatric clades within a hotspring (Cadillo-Quiroz *et al.*, 2012). Likewise, the marine bacterium, *Vibrio*, showed gradual separation of gene pools with increased habitat specificity between free-living and particle-associated microenvironments (Shapiro *et al.*, 2012; Yawata *et al.*, 2014). Such evidence suggests that the genetic structure of microbial populations is not only a function of geographic distance, but that ecological differentiation at microscales within a location may also be important (Polz *et al.*, 2013). Thus,

to investigate recent diversification among microorganisms, we might need to abandon the idea of defining populations *a priori* based on geography and, instead, focus first on the emerging genetic structure among closely-related individuals.

This possibility highlights a third barrier to investigating microbial populations: quantifying the exchange of genetic variation (i.e. gene flow). For prokaryotes, the exchange of genetic material is mediated through genetic recombination, whether homologous recombination or the transfer of entirely new genes. However, the asexual nature of prokaryotes makes it a challenge to quantify recombination, particularly among closely-related individuals. For instance, the more closely-related two genomes are, the harder it is to distinguish between differences caused by vertical inheritance and recombination. As such, many microbial studies infer gene flow from genetic divergence of sampled isolates, not necessarily considering whether isolates are from the same environment and/or represent interacting genotypes (Cordero and Polz, 2014). One approach to identify emerging genetic structure delineating populations is to examine gene-exchange networks between co-occurring strains to identify key ecological associations and barriers to recombination (Arevalo *et al.*, 2018).

To try and overcome the aforementioned obstacles, we focused on the abundant surface soil taxon, *Curtobacterium* (Chase *et al.*, 2016). Previously, we demonstrated that *Curtobacterium* encompasses multiple ecotypes, or fine-scale genetic clades that correspond to ecologically relevant phenotypes (Chase *et al.*, 2018). Thus, here we concentrate on the genetic diversity within a single ecotype, a unit that might be considered somewhat parallel to a species (Cohan, 2001). *Curtobacterium* is also relatively easy to culture from the leaf litter layer of soil;

therefore, we could isolate and sequence many individuals from a variety of geographic locations. Finally, we used three separate approaches to estimate the degree of genetic exchange among individuals: (1) a traditional admixture analysis to identify patterns of genetic similarity, (2) a network analysis to estimate the degree to which homologous recombination contributed to genetic structure, and (3) an analysis of the flexible genome to identify unique genes shared among individuals. We hypothesized that we could identify distinct populations (groups of individuals recombining more with one another than among groups) within the ecotype. Further, we hypothesized that sympatric populations may exist within geographic locations, while also varying in their geographic distribution. Such a pattern would indicate that the genetic structure within this bacterium is due to both adaptation among localized microenvironments as well as dispersal limitation between geographic locations.

RESULTS

Phylogenetic and Admixture Analysis

We identified 28 strains from the *Curtobacterium* ecotype, subclade IB/C (Chase *et al.*, 2018). These strains were previously isolated from leaf litter, the top layer of soil, at five geographic locations, including four from a regional climate gradient in southern California and from Boston, MA (Supplementary Table 1). All analyzed strains from the climate gradient have identical full-length 16S rRNA regions and share high sequence identity with ≥94.5% genome average nucleotide identity and ≥90% genome average amino acid identity, congruent with sequence similarity thresholds for defining sequence-discrete populations (Rodriguez-R and Konstantinidis, 2014). We reconstructed the phylogenetic history among all strains using the core genome (Fig. 1), revealing finer-scale genetic structure within the ecotype.

We then inferred population structure by computing ancestry coefficients for each strain (Fig. 1). The proportion of an individual genome originating from multiple ancestry gene pools (K=4) was congruent with the phylogenetic analysis. As expected, an outgroup strain originating from Boston, MA formed its own population (Population 1) and exhibited little evidence for recent mixing with the other strains (Fig. 1). After removing 'admixed' individuals (q-value < 0.75), we identified three potential populations across the climate gradient, Populations 2 (N = 4 individuals), 3 (N = 10), and 4 (N = 6). One population appeared to be restricted to one site (i.e. Population 3 was only found in the grassland location); however, the other two populations included strains from multiple sites along the climate gradient. For example, Population 4 contained strains from the grassland, scrubland, and Salton Sea litter communities.

Recombination Networks

While the admixture analysis identified overall patterns of genetic structure, it assumes that populations diverge and differentiate (via genetic drift) followed by a mixture phase. Therefore, we estimated the effect of recombination in structuring the core genome phylogeny using ClonalFrameML. The relative effect of recombination was similar to that of mutations (ratio of 0.94), signifying that the strains were not clonal, and that recombination accounted for as much nucleotide diversity as point mutations. Consequently, we next estimated the degree of homologous recombination among the individual genomes (defined as ≥99% nucleotide similarity for ≥500 bp) to create a gene-exchange network. The total pairwise recombination events between a pair of strains revealed a structured network of genome clusters that were congruent with the populations identified by the admixture analysis (Fig. 2). As we might

expect, total recombination events within populations were more frequent than between populations (Welsh two-sample t-test, $p < 0.0001$). The analysis also suggested even finer population subclusters (Supplementary Figure S1A vs. S1B); for example, it divided population 4 into two distinct subclusters, separating two strains (MMLR14002/014) that were isolated from the grassland site five years before the rest of the strains (Supplementary Figure S1B).

Estimations of homologous recombination can often be overestimated as the ability to distinguish between homologous recombination and vertically transmitted regions of the genome is reduced when genomic signatures are similar (Ravenhall *et al.*, 2015). To address this limitation, we employed a novel method that attempts to detect only recent recombination events between pairs of strains. Using this approach, we were able to mirror the results from the previous approach with the one exception of splitting Population 4 into finer subclusters (Supplementary Figure S1C). Further, this approach reduced all strains belonging to 'admixed' population groups (Admixed A-D) to individual subcluster nodes, suggesting that no recent recombination events connect 'admixed' individuals to the main populations observed along the climate gradient.

Both methods to characterize recombination networks displayed similar results and were also congruent with the admixture analysis. In addition, the frequency of recombination between strains was strongly related to phylogenetic distance (Mantel Test, $p < 0.01$), especially when we considered whether two strains were from the same population assignment (Fig. 3A). At the same time, total recombination was negatively correlated with geographic distance between strains (Fig. 3B; RELATE Test, $\rho = 0.58$, $p = 0.01$). Specifically, the average pairwise distance between every sampled member of the same population was constrained to

19.2 $\pm$ 49.4 km (Supplemental Figure S2A); while the population subclusters, were constrained to 29.5 $\pm$ 57.4 km (Supplemental Figure S2B). By combining the results from the admixture analysis with our gene-exchange network, we identified populations and finer population subclusters that were geographically constrained.

Composition of the Flexible Genome

Based on the recombination networks, we expected individuals within population assignments to exchange more genes with individuals of the same population than between. Therefore, we looked at flexible gene (genes not present in all strains) content similarity across all strains. Indeed, flexible gene content across strains generally recapitulated our population and subcluster designations (Fig. 4) such that strains within a population (analysis of similarities (ANOSIM); R = 0.83, p = 0.001) and within subclusters (ANOSIM; R = 0.79, p = 0.001) shared more similar flexible genes than expected by chance. Together, all three analyses (admixture, recombination, flexible gene content) indicate that the distinct genetic lineages observed across the climate gradient represent discrete microbial populations. We also observed a significant nonlinear relationship between flexible gene content with recombination (Supplementary Figure S3A) and phylogenetic distance (Supplementary Figure S3B), such that individuals within the same population consistently clustered together (Mantel Test, both p < 0.01).

Genomic Signatures of Ecological Association

Population Statistics of the Core Genome

After identifying distinct populations, we characterized the genetic variation within populations. Specifically, we compared genome-wide population genetic summary statistics for

all population assignments excluding 'admixed' individuals. Populations 3 and 4 had relatively high nucleotide diversity across the core genome (π = 10.6 $\pm$ 8.0 and 11.3 $\pm$ 9.1, respectively) indicating large, intermixing populations. Conversely, Population 2 had relatively low nucleotide diversity (π = 4.1 $\pm$ 4.7), suggesting a small, isolated population (Supplementary Figure S4A). However, when we evaluate other population metrics (i.e. Tajima's D) we observed different patterns. For example, Population 3 demonstrated evidence for recent population expansion (Tajima's D = -0.16); whereas, Population 4 appears to have undergone population contraction (Tajima's D = +0.18; Supplementary Figure S4B). These biological interpretations are further supported by the recombination networks, particularly if we only consider recent recombination (Supplementary Figure S1C). For example, the recent population expansion suggests a proliferation of rare variants, which is evidenced by the breakdown of Population 3 into four distinct population subclusters. On the other hand, Population 4 contained two strains isolated from five years prior (MMLR14002/014), enabling detection of the "ancestral" population shifting to higher frequency alleles and undergoing balancing selection.

Population Differentiation of the Flexible Genome

The flexible genome can provide insights into ecological genetics where traits related to fitness could proliferate through all members of a population via recombination. Therefore, we searched for flexible genes that were only present in all individuals within a population. To link flexible gene content to recombination, we searched for population-specific genes that were also localized on the genome. We concentrated on population subclusters as these assignments provided the best evidence for recombination delineating independent gene flow units (Supplementary Figure S1C). We observed two population subclusters, Subclusters 2 and 4.2,

with nearly identical genetic architectures shared among every member of the population in regions containing population-specific genes (Fig. 5). These regions did not contain phage or integrative and conjugative elements (ICEs) but did contain other mobile elements such as insertion sequences and clustered regularly interspaced short palindromic repeats (CRISPRs). Further, these genomic regions were littered with pseudogenic exons, suggesting interruption of functional proteins due to recombining genomic segments.

These population-specific genes were consistently flanked by highly conserved genes (present in >85% of all *Curtobacterium* strains) in nearly identical orientation, suggesting a mechanism related to increased homologous recombination. For example, Population Subcluster 4.2 contained a total of 6 population-specific genes, with 4 of those being highly localized (Fig. 5A) and were always flanked by conserved genomic regions. We observed similar patterns in Population Subcluster 2 as well, which contained high localization of the 16/48 population-specific genes (Fig. 5B). We did not detect any localization of population-specific genes in Subclusters 3.1 and 3.2, most likely due to its recent population expansion (Supplementary Figure S4B). In both subclusters where we did observe high gene localization, conserved flanking regions exhibited highly consistent phylogenetic relationships within population subclusters (Fig. 5C), indicating these genes are not an artifact of measurement but represent evidence for increased homologous recombination within subclusters. Further, conserved genes that did not provide robust phylogenetic signals typically flanked variable regions with genes shared across population boundaries or singleton genes (genes 3 and 4 in Fig. 5C).

Together, our results indicate that population subclusters represent cohesive genetic clusters sharing highly conserved genomic backbones with population-specific gene cassettes. The most compelling evidence for this observation is the population-specific genes contained in these backbones (Fig. 5) have nearly swept to fixation in the populations. For instance, population-specific genes from both Subcluster 4.2 (N=4) and Subcluster 2 (N=16) had low nucleotide diversity (π_{MEAN} = 2.2 and 3.9, respectively) and high nucleotide percent identity (94.1% and 99.3%, respectively) across all individuals within a population subcluster. The presence of selective sweeps within individuals of the same population subcluster indicates strong positive selection of beneficial genes.

Due to these population-specific genes becoming nearly fixed within population subclusters, these genomic backbones are most likely attributing to differential ecological associations. The population-specific genes, themselves, did not provide any indication of ecological differentiation (all annotate as hypothetical protein), but flanking regions and singleton genes suggest these genomic backbones are related to small fitness differences in environmental resources. For example, we observed high occurrences of metal uptake and transport proteins, along with glycoside hydrolase (GH) enzymes for the breakdown of carbohydrates in leaf litter. Likewise, when we assayed each strain for genomic traits within population subclusters for indications of ecological differentiation, we observed some trait differences between population subclusters in minimum generation time, optimal temperature, and GH composition (Supplementary Figure S5). Collectively, our analyses indicate that the correlations between core and flexible genes, specifically population-specific gene cassettes, suggests these subclusters are in the early stages of ecological differentiation.

DISCUSSION

Our study highlights the growing understanding that the origin and maintenance of microbial populations is governed by the same ecological and evolutionary processes shaping macroorganisms. Utilizing population genetics and gene exchange networks, we observed distinct genetic lineages that recombined more with individuals of the same population than between, suggesting limited gene flow between populations. Together, our results indicate biogeographic barriers (via dispersal limitation) as well as genetic isolation (via ecological differentiation) contribute to population structure in microbes as one would typically observe in geographically-distinct eukaryotic populations. More broadly, our study highlights the importance of both dispersal limitation and local adaptation in governing the processes contributing to the divergence among closely-related bacteria.

Previously, only two instances in soil bacteria, *Streptomyces* and *Bradyrhizobium*, have shown that dispersal limitation at continental scales and distant evolutionary events (e.g., free-living v symbiont) are consistent with allopatric speciation (VanInsberghe *et al.*, 2015; Andam *et al.*, 2016; Choudoir *et al.*, 2016). None, to our knowledge have investigated the ecological and evolutionary processes driving initial population divergence between interacting genotypes in soil systems. Here, using strains within a cohesive ecological unit (akin to a eukaryotic species (Cohan, 2001)), we demonstrated population differentiation within this *Curtobacterium* ecotype (Fig. 1). However, teasing apart the relative contributions of biogeographic barriers and ecological differentiation was complex. Based on previous work in soil fungal populations conducted at similar regional spatial scales, we expected that populations might be structured geographically and genomic differences would reflect local site adaptation along the gradient

(Amend *et al.*, 2010; Branco *et al.*, 2015). While we observed a geographic signal constraining population intermixing (Fig. 3B), we also observed the co-occurrence of multiple, genetically distinct populations within and across sites.

One possible explanation for the appearance of multiple *Curtobacterium* populations is that populations were once separated by a biogeographic barrier and have yet to genetically homogenize across spatial scales between ecologically-similar populations. Moreover, the measured geographic distance may not be indicative of contemporary population dynamics as we cannot account for the frequency or abundance of populations at each site. However, once we incorporated gene flow (i.e. the exchange of genetic variation) in our analysis, we observed distinct and independent gene flow units, suggesting that barriers to recombination were highly constrained by both population boundaries and geography. Our results suggest that the observed populations were not explicitly delineated by geography. Rather, sympatric microbial populations must be subjected to some isolating mechanism that protects the integrity of cohesive genotypes (Mayr, 2001). Indeed, our results suggest that populations are genetically isolated from one another most likely due to ecological differentiation, as others have noted in marine microbial populations (Shapiro and Polz, 2014).

Evidence for ecological differentiation between populations is evidenced in the flexible genome, which can provide insights into potential habitat-specific associations. While the flexible genome can enable sharing of genes for habitat-specific adaptation between distantly-related organisms (Tettelin *et al.*, 2008), our results indicate that the putative exchange of adaptive genes are, for the most part, restricted by population boundaries (Fig. 4A). These flexible genes are thought to contribute to differences in niche exploitation (Rodriguez-Valera

and Ussery, 2012) and could contribute to small fitness differences between populations that enable ecological differentiation even among sympatric populations (Cordero and Polz, 2014). For example, in the marine bacterium *Vibrio*, sympatric populations encoded habitat-specific genes (Shapiro *et al.*, 2012) to ecologically differentiate at a microscale between free-living and particle-associated populations (Yawata *et al.*, 2014). Along similar spatial scales, *Curtobacterium* populations may differentiate between microhabitats on leaf litter due to variation in environmental resources, such as metals and carbohydrate availability. To that end, we also observed differences in growth strategies and carbohydrate degradation potential (Supplementary Figure S5) that may contribute to ecological differentiation among populations.

Differences in ecological associations between populations is also leading to decreased gene flow and a separation of gene pools, a mechanism which has previously been shown to lead to the early stages of speciation in archaea (Cadillo-Quiroz *et al.*, 2012). Moreover, the presence of highly conserved genomic backbones shared across members within a population suggest that the mechanism reinforcing differentiation is primarily homologous recombination (Fig. 5). Within these population-specific genomic backbones, we observed gene-specific selective sweeps localized in large genomic islands that proliferate in a population-specific manner. Indeed, gene-specific sweeps have been identified in marine populations in *Vibrio* (Shapiro *et al.*, 2012) and *Prochlorococcus* (Kashtan *et al.*, 2014), where sweeps were linked to small fitness differences contributing to the coexistence of sympatric populations. Similarly, increased homologous recombination within strains of *Curtobacterium* populations could enable the rapid exchange of niche-adaptive genes for differential microhabitat specialization on leaf litter.

Collectively, our results suggest a model for differentiating populations within a Curtobacterium ecotype through a combination of sympatric and allopatric processes. Populations along the regional climate gradient represent genetically-isolated lineages that are ecologically differentiating by partitioning of microhabitat resources. As in *Prochlorococcus* (Kashtan *et al.*, 2014), the exchange of flexible genes and the homologous recombination of population-specific genomic backbones (Fig. 5) may contribute to new a dimension of niche differentiation. At the same time, these populations also are able to disperse and intermix across sites along the regional gradient with other subpopulations that specialize on similar microhabitats. The presence of geographically-distinct strains within a population sharing nearly identical genomic backbones suggests that the acquisition of a beneficial flexible gene cassette can proliferate in a population-specific manner across geographic distances. Our results, therefore, suggest that microbial populations may differ in the degree of granularity in microhabitat preference while at the same time being connected via dispersal.

A major gap in our understanding of microbial diversity are the mechanisms contributing to the origin and maintenance of microbial speciation. Most of our estimates related to evolutionary adaptation are restricted to laboratory measurements or concentrated on pathogenic strains (Ingle *et al.*, 2016; Lemieux *et al.*, 2016). And while insights into community-wide approaches using metagenomics can reveal evidence for genome recombination (Denef and Banfield, 2012), these approaches are limited in the ability to detect the proliferation of population-specific genes (Bendall *et al.*, 2016), which are indicative of the early stages of ecological differentiation (Polz *et al.*, 2013; Takeuchi *et al.*, 2015). Our results, and others, demonstrate that free-living bacterial populations are delineated by barriers to

recombination that enable the proliferation of advantageous genes in a population-specific manner (Whitaker *et al.*, 2005; Fraser *et al.*, 2007; Cadillo-Quiroz *et al.*, 2012; Shapiro *et al.*, 2012). Finally, by sampling across a regional climate gradient, we can identify the sympatric and allopatric mechanisms contributing to population divergence.

MATERIALS AND METHODS

Field Sites and *Curtobacterium* Strains

We downloaded 28 *Curtobacterium* genomes (Supplementary Table 1) from the National Center for Biotechnology Information (NCBI) [https://www.ncbi.nlm.nih.gov/] database that were previously isolated from leaf litter (Chase *et al.*, 2016), including a robust genomic dataset consisting of 24 strains from a climate gradient in southern California (Chase *et al.*, 2018). We included two additional strains within the same ecotype from outside Boston, MA to provide varying spatial scales for population comparisons. Protein-coding regions and gene annotations were derived from the NCBI prokaryotic genome annotation pipeline (Tatusova *et al.*, 2016). Genomes were screened for the presence of mobile elements by identifying integrating and conjugative elements (ICEs) with the ICEberg database (Bi *et al.*, 2011), prophage sequences using PhiSpy (Akhter *et al.*, 2012), insertion sequences (IS) with ISfinder (Siguier *et al.*, 2006), and CRISPR with CRISPRCasFinder (Couvin *et al.*, 2018).

Core Genome Population Structure

We aligned all genomes using progressiveMauve (Darling *et al.*, 2010) to identify locally collinear blocks (LCBs) of genomic data. We identified 49,610 LCBs >1500 bp found across all 28 genomes that represented 1.28 Mbp of the core genome. This core genome alignment was used to perform a maximum likelihood bootstrap analysis using RAxML v8.2.10 (Stamatakis,

2014) under the general time reversal model with a gamma distribution for 100 replicates. Using the core genome, we performed an initial analysis to infer the relative effects of recombination and mutation rates using ClonalFrameML v1.11 (Didelot and Wilson, 2015). Specifically, we attempted to reconstruct phylogenetic relationships by detecting regions of recombination across the phylogeny to provide an initial estimate for clonal genealogy.

Due to the weak clonal structure among strains, we sought to infer population structure from multilocus genotype data. First, we converted the core genome sequence data to a genotype matrix reflecting the distance between polymorphic sites of all individuals (https://github.com/xavierdidelot). We then used this genotype matrix to compute ancestry coefficients to delineate genetic clusters. Specifically, we employed sparse non-negative matrix factorization algorithms to estimate the cross-entropy parameter (Frichot *et al.*, 2014). Based on the cross-entropy criterion which best fit the statistical model, we designated the number of ancestral populations to K=4 to estimate individual admixture coefficients using the LEA package (Frichot and François, 2015) in the R software environment (Pinheiro *et al.*, 2011). Finally, individual membership values to each population (*q*-value) were used to classify isolates to Populations 1-4 (*q*-value > 0.75) and 'admixed' groups (*q*-value < 0.75). Admixed groups were further divided by phylogenetic clusters at a 0.075 phylogenetic distance.

Gene Flow Subclusters and Recombination

We employed two separate analyses to estimate genetic exchange by 1) investigating all potential recombination events and 2) accounting for only recent recombination between pairs of strains.

As a proxy for total recombination, we compared all protein coding genes that shared sequence similarity ≥99% of at least 500 bp, as defined in (Bonham *et al.*, 2017). Identified neighboring candidate genes were further grouped into recombination blocks if they were separated by ≤5000 bp. Due to the high genomic similarity among strains, we needed to verify each designated recombinant block in a pairwise manner. Therefore, we built a local BLAST database containing all protein-coding regions of all strains and searched all identified candidate recombining genes with blastn v2.6.0+ (Camacho *et al.*, 2009). Consequently, we only defined recombination events if the query and subject gene matched genome block designation, were ≥99% sequence similarity, and ≥90% query coverage to avoid spurious hits. Finally, pairwise recombining genes were used to construct a gene exchange network analysis and were normalized by the total number of genes found within a recombination block. To assign population subclusters using only recombination, we aggregated the pairwise analyses to reflect the total recombination events between each pairs of strains. Each strain was then normalized by subtracting the mean recombination events from all other strains and dividing by the standard deviation. Finally, we computed a Euclidean distance matrix and conducted a hierarchical clustering analysis to delineate subclusters. To test for correlations between total recombination with phylogenetic and geographic distance, we used a Mantel and RELATE test.

To differentiate between vertical transmission and recent homologous recombination, we identified identical genomic regions (100%) between each pair of strains. To correct for the influence of vertical inheritance, we created a null model of expected mutational divergence and looked for enriched genomic regions that deviated from the null model. To do so, we implemented a decay rate constant to identify large genomic regions that remained identical

across strains despite the null expectation that larger regions will be less identical over time due to divergence and mutation accumulation. We then extrapolated a length measurement bias to generate an average genome wide measurement to create gene networks between strains. This process allows us to focus on recent recombination transfer events that should correspond to population units.

Population Genetic Analyses

To perform population genetic summary analyses, we identified all orthologous protein-coding genes (orthologs) shared across all strains. Orthologs were initially predicted using ROARY (Page *et al.*, 2015) with a minimum sequence identity of 90% to ensure all possible orthologs were included across populations (Supplementary Figure 6A). The resulting 2193 orthologs shared across all strains were individually aligned with ClustalO v1.2.3 (Sievers *et al.*, 2011) and used to create a 2.14 Mbp concatenated nucleotide alignment. Note, the size of this alignment differs from the core genome alignment since genes do not necessarily need to be located on LCBs. To verify the effects of using a gene x gene approach on the core genome, we reconstructed the phylogenetic relationship of the concatenated alignment of all orthologous protein-coding genes, using RAxML v8.2.10 (Stamatakis, 2014) under the general time reversal model with a gamma distribution for 100 replicates, and compared to phylogeny derived from the Mauve core genome alignment (Supplementary Figure 6B). Next, all individual ortholog alignments were screened for complete codon reading frames (i.e. multiple of 3 bp) and the resulting 2137 genes were individually used to calculate population genetic summary statistics. Specifically, we calculated nucleotide diversity and neutrality statistics within populations using the PopGenome package (Pfeifer *et al.*, 2014) in R, as outlined in (Lemieux *et al.*, 2016).

Predicted orthologs that were not shared across all strains represent the flexible genome (Supplementary Figure 6A). Using all identified orthologs, we computed a Jaccard distance between pairs of strains to estimate shared gene content. The distance matrix was used to generate a neighbor-joining tree based on 1000 re-samplings and to create a heatmap showing gene content similarity across strains. We tested the significance of gene content using an analysis of similarities (ANOSIM) for population or subcluster designations and site of isolation for 9999 permutations. In addition, we looked for orthologs that were unique to our populations. Specifically, we identified orthologs that were encoded by every member within a population subcluster (N ≥ 3 individuals) and were not found in any member outside of the subcluster designation. To reduce this list even further, we identified population-specific orthologs that were localized in genomic space (<10 kbp separation).

Analysis of Genomic Traits

We analyzed all genomic sequences for specific ecological traits that may contribute to population divergence. We concentrated on genomic traits related to growth strategies and substrate (i.e. carbohydrate) utilization that may be advantageous on leaf litter.

To infer growth strategies, we estimated minimum generation times (MGT) and optimal growth temperature (OGT). We predicted MGT by comparing codon-usage biases between highly expressed ribosomal proteins and all other encoded genes following a linear regression model (Vieira-Silva and Rocha, 2010)[equation 1].

[1] $$\Delta ENC = \frac{ENC_{all} - ENC_{ribosomal\ proteins}}{ENC_{all}}$$

where ENC is the effective number of codons given %GC (Subramanian, 2008)

We analyzed each strain for the genomic potential to degrade various carbohydrates by searching the predicted coding-regions against the Pfam-A v30.0 database (Finn *et al.*, 2016) using HMMer (Finn *et al.*, 2011). Identified protein families were reduced to only known protein families that encode for glycoside hydrolase (GH) and carbohydrate binding module (CBM) proteins as described in (Chase *et al.*, 2016).

ACKNOWLEDGEMENTS

The authors would like to thank Claudia Weihe, Chamee Moua, and Michaeline Albright for their assistance in sample collection and strain isolation. We would also like to thank Brandon Gaut for invaluable insight into data analysis and interpretation, Sarai Finks, Kendra Walters, Cynthia Rodriguez, and the rest of the Martiny Lab for helpful comments, and Xavier Didelot and Kevin Bonham for software assistance. This work was supported by an US Department of Education Graduate Assistance of National Need (GAANN) fellowship and US Department of Energy Office of Science Graduate Student Research (SCGSR) fellowship to ABC.

Figures and Tables

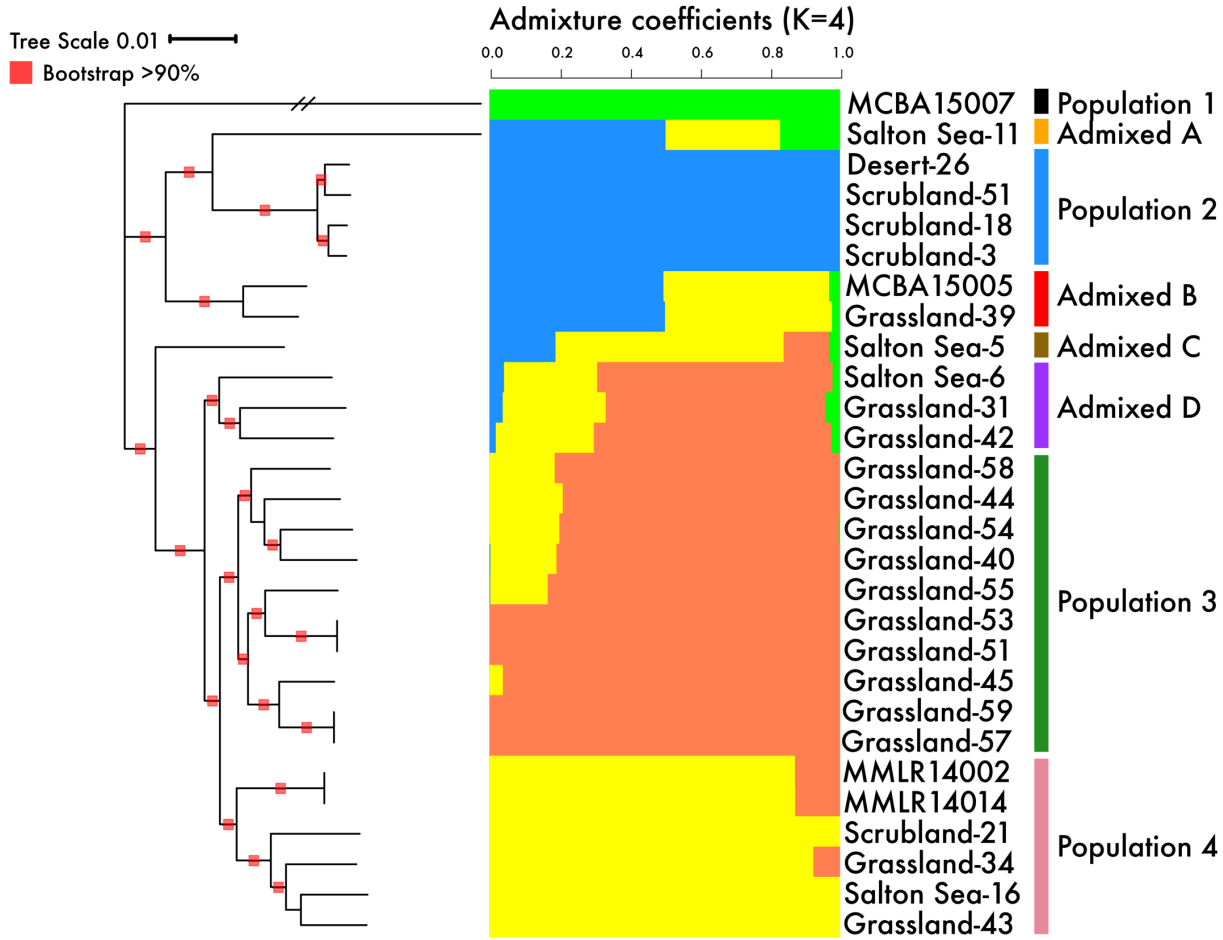


Figure 5.1 Phylogeny of the *Curtobacterium* ecotype, subclade IB/C, and its underlying populations constructed from a core genome alignment. Bar plots reflect the proportion of an individual genome that originate from estimated ancestral gene pools (K = 4). Genome names designate the site of isolation along the climate gradient except for MCBA = Boston and MMLR = Grassland isolate from 2010.

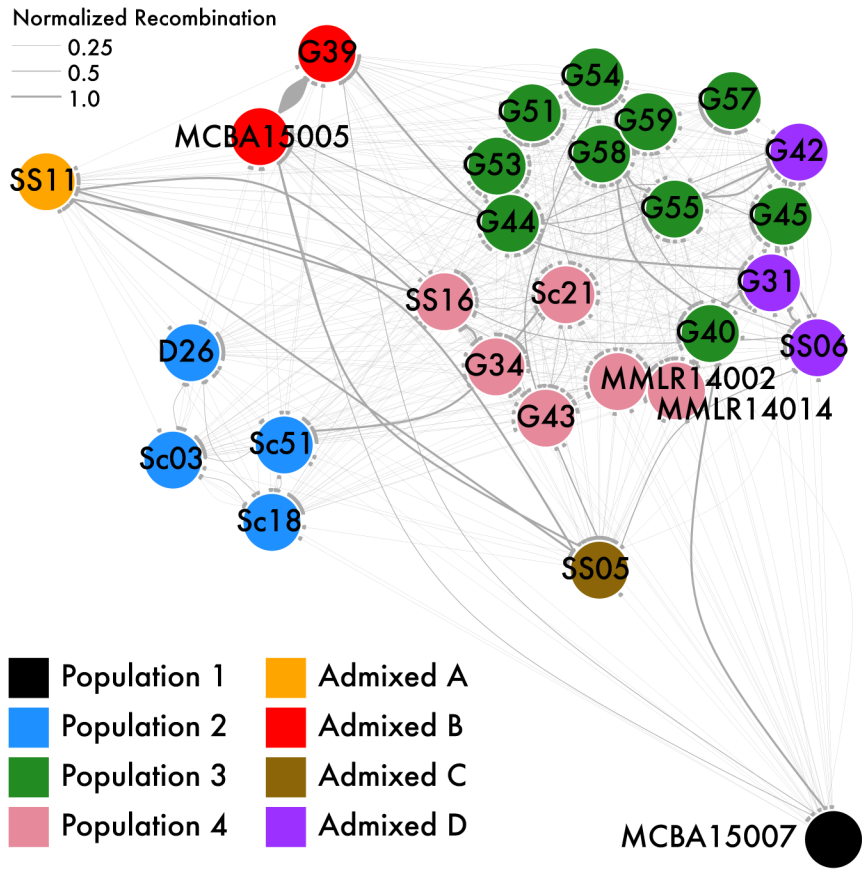


Figure 5.2 Gene-exchange network across all pairwise strains with nodes colored by population assignments from admixture analysis (Fig. 1) and edge widths normalized by the percent of genes within identified recombinant blocks.
D = Desert, Sc = Scrubland, G/MMLR = Grassland, SS = Salton Sea, MCBA = Boston.

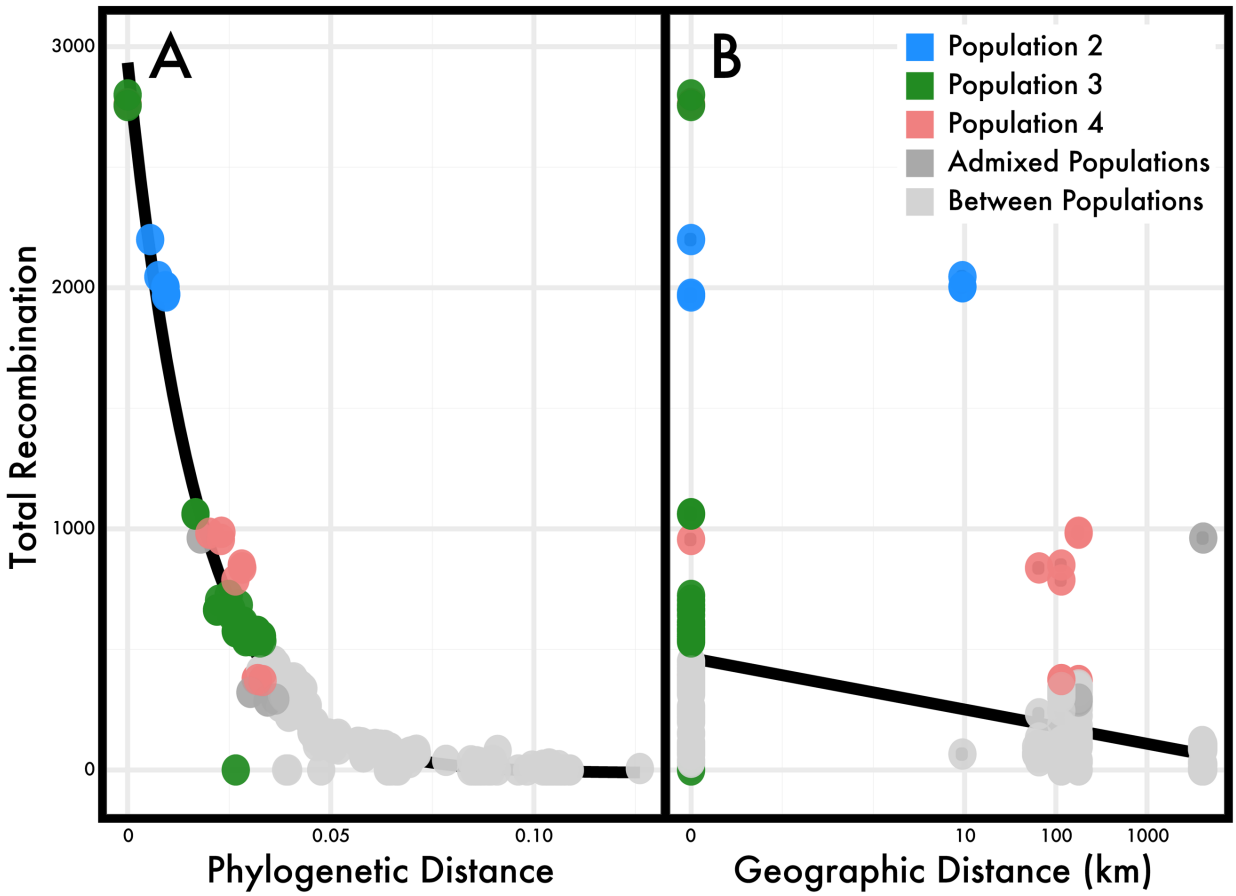


Figure 5.3 Total recombination by **(A)** phylogenetic and **(B)** geographic distance. Each point represents a pairwise genome comparison and colored by population assignment. Total regression line is colored in black.

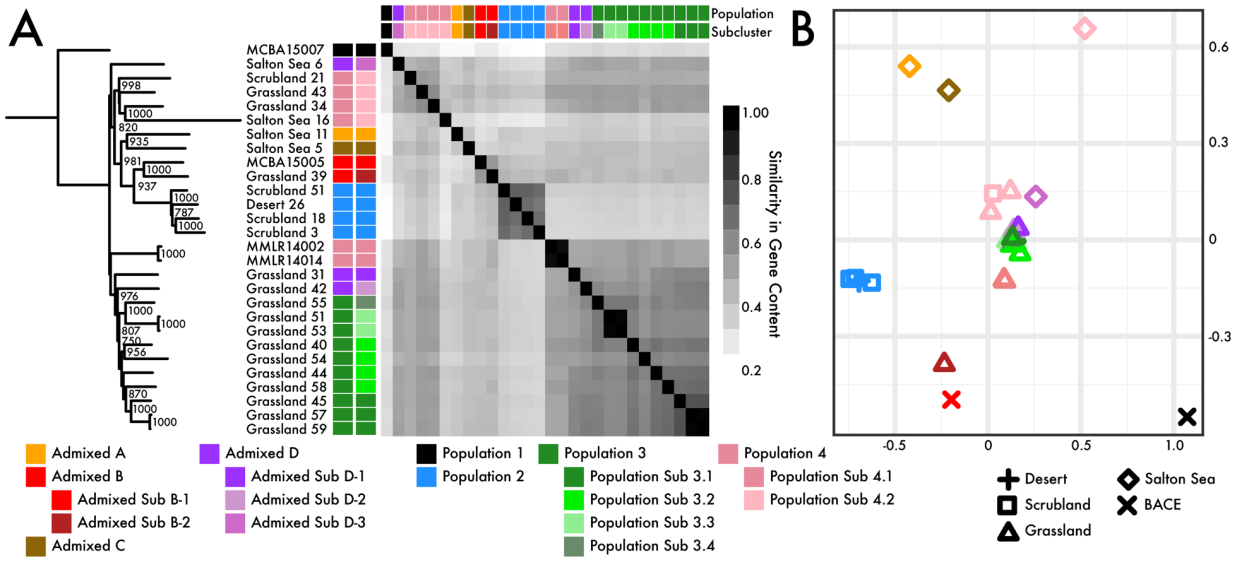


Figure 5.4 Flexible gene content similarity between strains. **(A)** Tree is derived from a consensus neighbor-joining analysis showing only nodes with ≥750 support. Strains are colored by both population and population subcluster assignments. **(B)** Multidimensional scaling (MDS) plot depicting total gene composition for *Curtobacterium* strains. Each point represents an individual strain and is colored based on the assigned population subcluster with the symbol denoting the site of isolation.

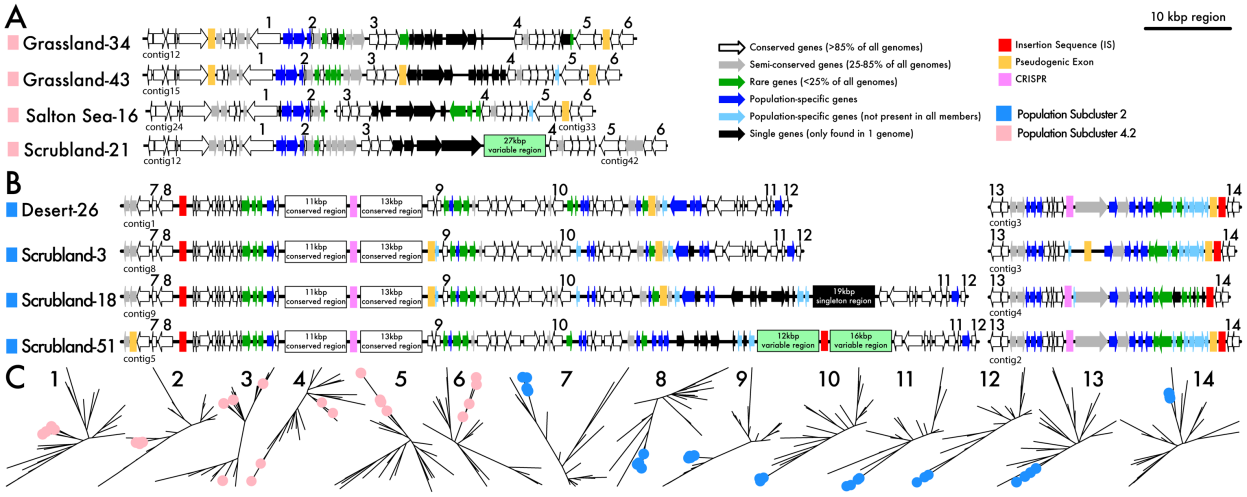
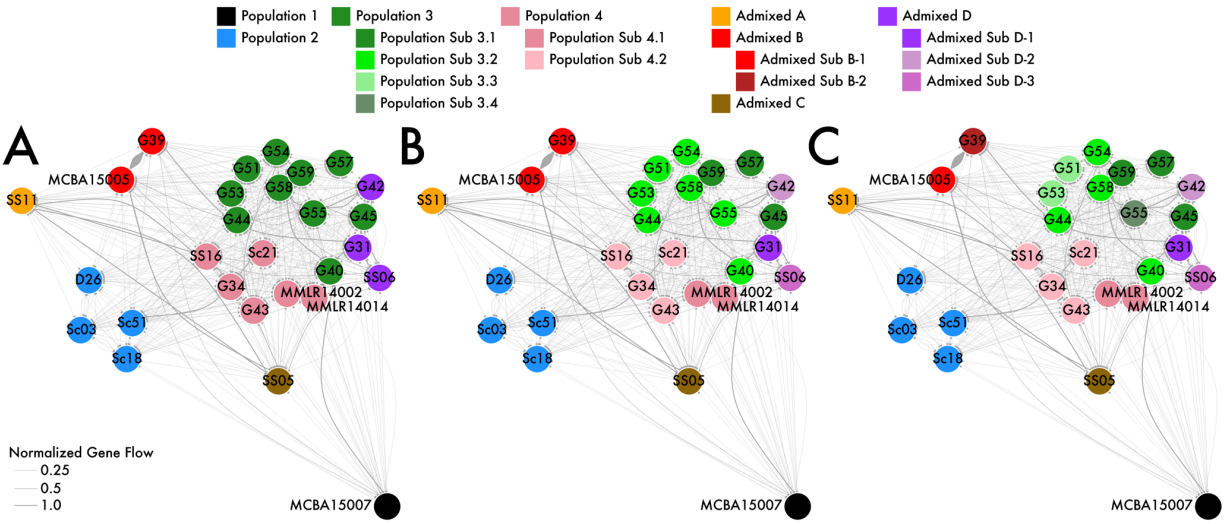


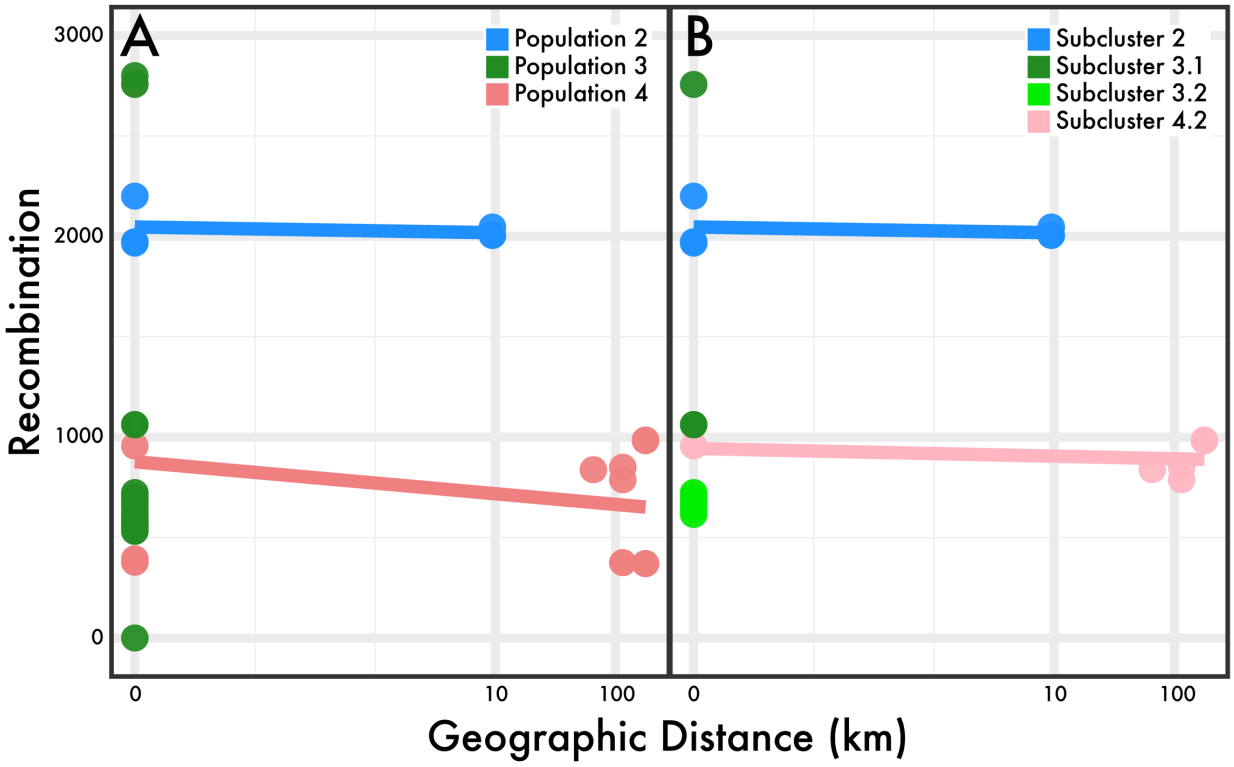
Figure 5.5 Highly structured genomic backbones across strains. Population-specific genomic backbones within **(A)** Population Subcluster 4.2 and **(B)** Population Subcluster 2. Population-specific genes (colored in blue) are flanked by highly conserved regions (in white). Putative mobile elements are also designated in boxes along the chromosome. **(C)** Phylogenies of a subset of conserved genes flanking the population-specific regions colored by each respective population subcluster.

Supplementary Table 5.1 Genomic and geographic characteristics of isolates.

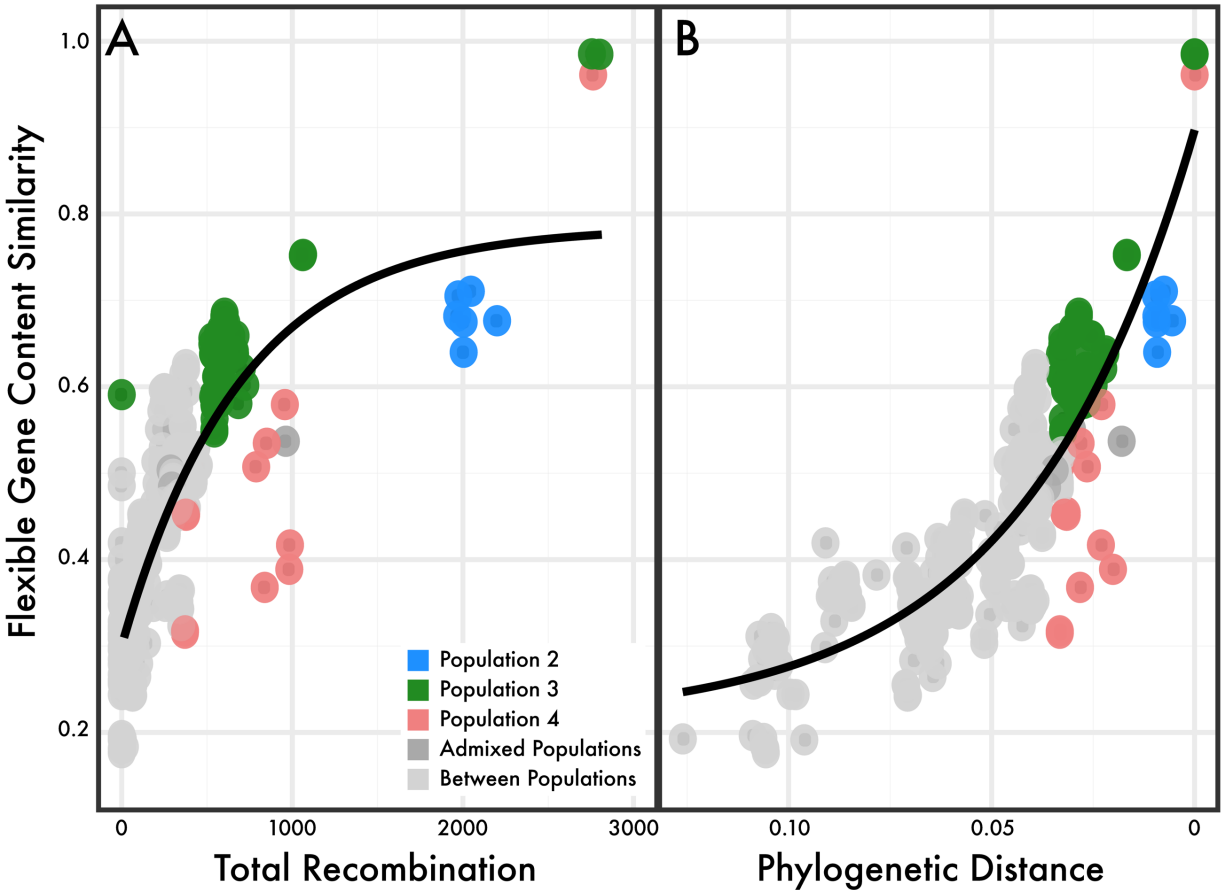
Genome Name	Location	Lat-Long	Isolation Year	Total Contigs	Genome Length (bp)	%GC	Reference
MCBA15005	Boston, MA, USA	42.38, -71.21	2015	91	3772244	69	AB Chase et al. Frontiers in Microbiology. 2016
MCBA15007	Boston, MA, USA	42.38, -71.21	2015	87	3768639	69	AB Chase et al. Frontiers in Microbiology. 2016
MMLR14002	Loma Ridge, CA, USA	33.74, -117.69	2011	137	3808678	70	AB Chase et al. Frontiers in Microbiology. 2016
MMLR14014	Loma Ridge, CA, USA	33.74, -117.69	2011	139	3822836	69	AB Chase et al. Frontiers in Microbiology. 2016
DESERT-26	Deep Canyon, CA, USA	33.65, -116.37	2016	40	3645388	70	AB Chase et al. Environmental Microbiology. 2018
GRASS-31	Loma Ridge, CA, USA	33.74, -117.69	2016	51	3777956	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-34	Loma Ridge, CA, USA	33.74, -117.69	2016	35	3820963	70	AB Chase et al. Environmental Microbiology. 2018
GRASS-39	Loma Ridge, CA, USA	33.74, -117.69	2016	37	3819069	70	AB Chase et al. Environmental Microbiology. 2018
GRASS-40	Loma Ridge, CA, USA	33.74, -117.69	2016	39	3761440	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-42	Loma Ridge, CA, USA	33.74, -117.69	2016	37	3838097	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-43	Loma Ridge, CA, USA	33.74, -117.69	2016	30	3695116	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-44	Loma Ridge, CA, USA	33.74, -117.69	2016	47	3789494	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-45	Loma Ridge, CA, USA	33.74, -117.69	2016	57	3764851	70	AB Chase et al. Environmental Microbiology. 2018
GRASS-51	Loma Ridge, CA, USA	33.74, -117.69	2016	41	3785722	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-53	Loma Ridge, CA, USA	33.74, -117.69	2016	42	3786104	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-54	Loma Ridge, CA, USA	33.74, -117.69	2016	52	3927711	70	AB Chase et al. Environmental Microbiology. 2018
GRASS-55	Loma Ridge, CA, USA	33.74, -117.69	2016	34	3752819	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-57	Loma Ridge, CA, USA	33.74, -117.69	2016	47	3728355	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-58	Loma Ridge, CA, USA	33.74, -117.69	2016	42	3800007	71	AB Chase et al. Environmental Microbiology. 2018
GRASS-59	Loma Ridge, CA, USA	33.74, -117.69	2016	25	3729778	71	AB Chase et al. Environmental Microbiology. 2018
SALT-11	Salton Sea, CA, USA	33.33, -115.84	2016	63	3632003	70	AB Chase et al. Environmental Microbiology. 2018
SALT-16	Salton Sea, CA, USA	33.33, -115.84	2016	58	4357272	70	AB Chase et al. Environmental Microbiology. 2018
SALT-5	Salton Sea, CA, USA	33.33, -115.84	2016	32	3744522	70	AB Chase et al. Environmental Microbiology. 2018
SALT-6	Salton Sea, CA, USA	33.33, -115.84	2016	29	3739706	70	AB Chase et al. Environmental Microbiology. 2018
SCRUBLAND-18	Pinon Flats, CA, USA	33.61, -116.46	2016	30	3658022	70	AB Chase et al. Environmental Microbiology. 2018
SCRUBLAND-21	Pinon Flats, CA, USA	33.61, -116.46	2016	54	3833068	70	AB Chase et al. Environmental Microbiology. 2018
SCRUBLAND-3	Pinon Flats, CA, USA	33.61, -116.46	2016	34	3695248	70	AB Chase et al. Environmental Microbiology. 2018
SCRUBLAND-51	Pinon Flats, CA, USA	33.61, -116.46	2016	36	3558624	71	AB Chase et al. Environmental Microbiology. 2018



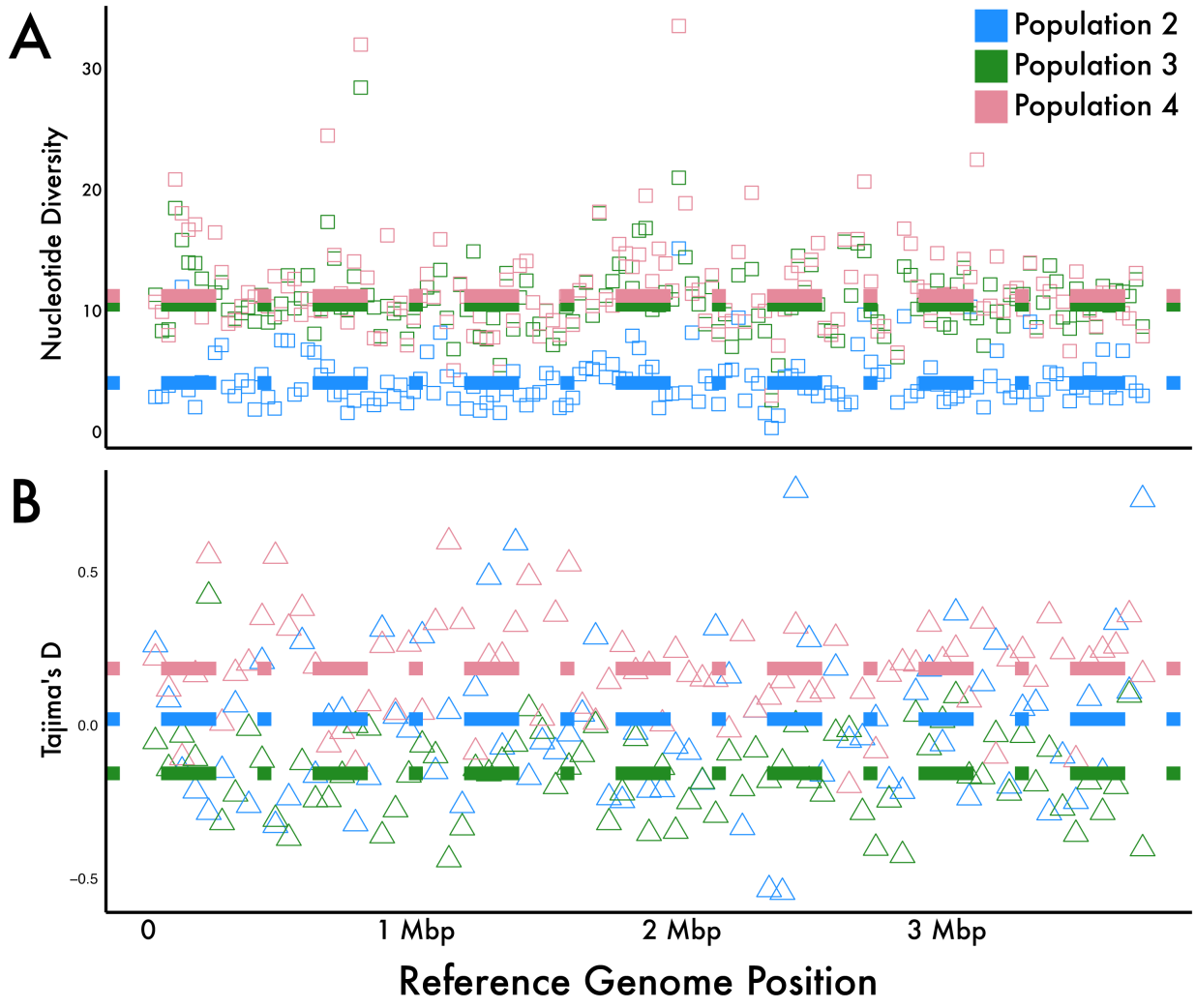
Supplementary Figure 5.1 Comparison of recombination network analyses and derived population subcluster designations. All panels show identical recombination networks but have individual strains (nodes) colored based on separate analyses. **(A)** Population assignments derived from admixture analysis (same as Figure 2). **(B)** Total recombination between strains separating populations subclusters. **(C)** Recent recombination delineating subclusters.
D = Desert, Sc = Scrubland, G/MMLR = Grassland, SS = Salton Sea, MCBA = Boston.



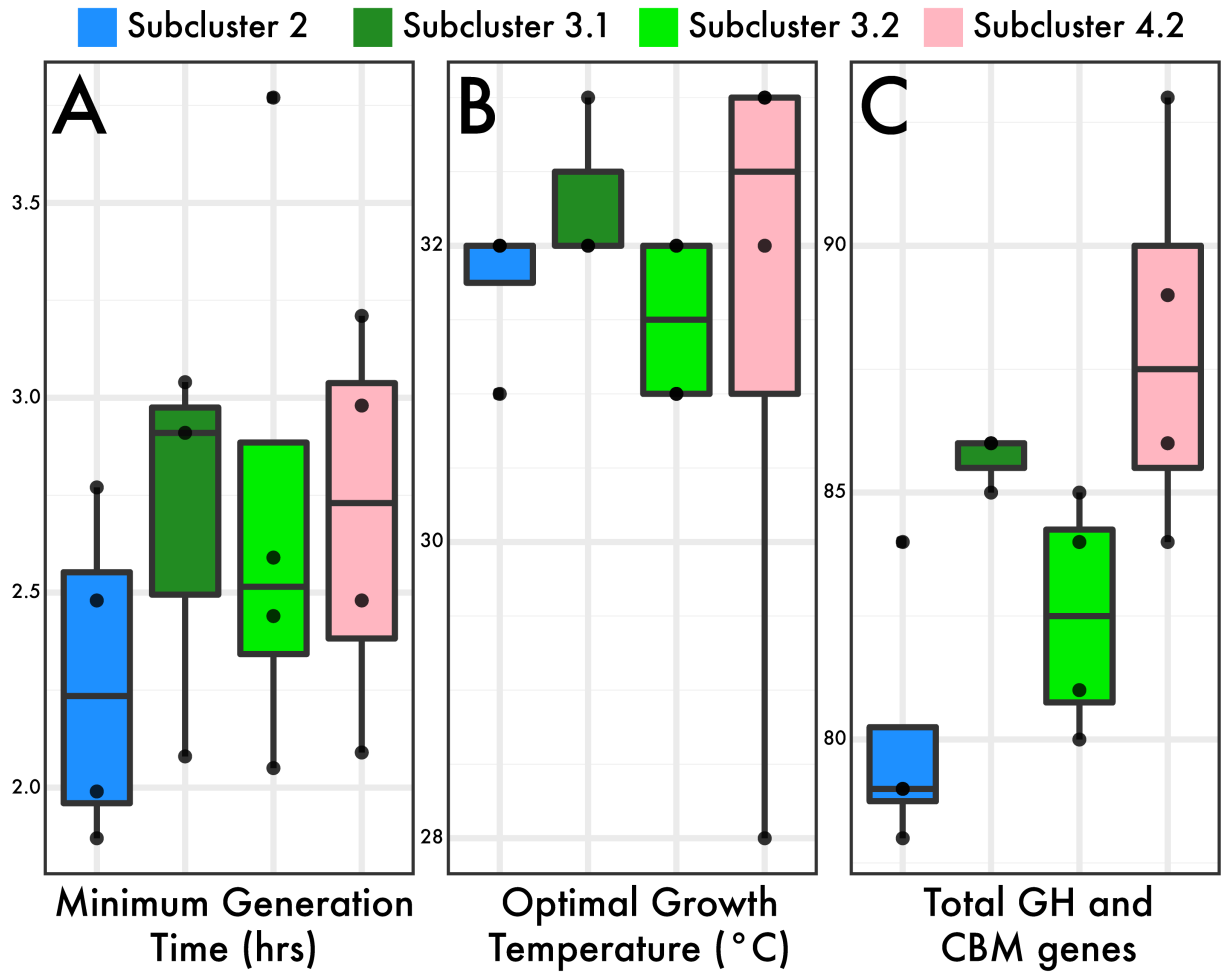
Supplementary Figure 5.2 Total recombination within and between populations in relation to geographic distance between strains. Each point represents a pairwise genome comparison and colored by whether the comparison is from **(A)** populations (N ≥ 4 strains) or **(B)** population subclusters (N ≥ 3 strains).



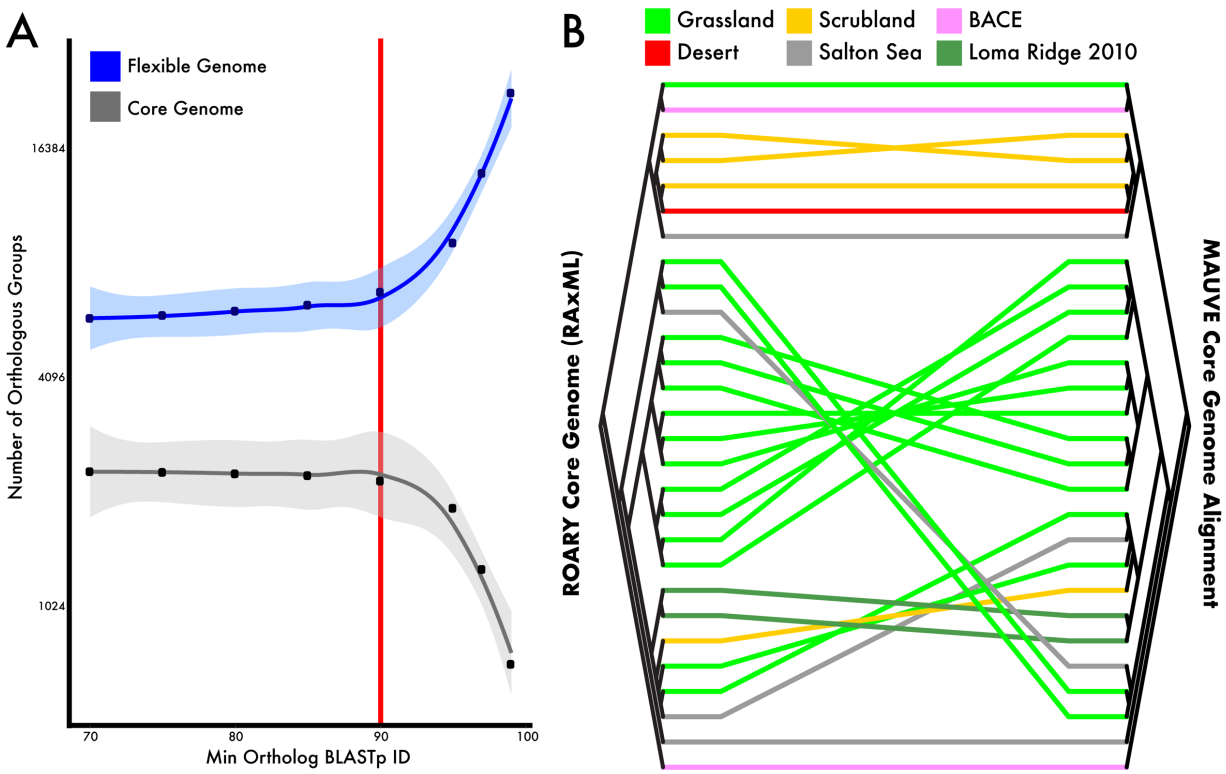
Supplementary Figure 5.3 Flexible gene content similarity in relation to **(A)** total recombination and **(B)** phylogenetic distance. Each point represents a pairwise genome comparison and colored by population assignment. Total regression line is colored in black.



Supplementary Figure 5.4 Genome-wide population genetic summary analyses. **(A)** Pairwise nucleotide diversity (π) in 25 kbp sliding windows by population. **(B)** Tajima's D statistic in 50 kbp sliding windows by population. Each point represents the average value in the sliding window and dashed line shows total genome average. Genomic position is to reference strain MMLR14002.



Supplementary Figure 5.5 Distributions of predicted genomic traits in strains belonging to population subclusters. Traits include: **(A)** minimum generation time (hrs), **(B)** optimal growth temperature (°C), and **(C)** total abundance of glycoside hydrolase (GH) and carbohydrate binding module (CBM) proteins.



Supplementary Figure 5.6 Breakdown of orthologous protein groups derived from all strains. **(A)** Number of identified orthologous protein groups in both the core and flexible genome based on initial clustering of proteins. **(B)** Cladogram comparison of core genes (N=2193 orthologous proteins) and core genome alignment (defined as locally collinear blocks). Terminal branches are colored by geographic location with lines connecting identical strains in each respective cladogram.

References

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., and Polz, M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551.

Adair, E.C., Parton, W.J., Del Grosso, S.J., Silver, W.L., Harmon, M.E., Hall, S.A., et al. (2008) Simple three-pool model accurately describes patterns of long-term litter decomposition in diverse climates. *Glob. Chang. Biol.* **14**: 2636–2660.

Agarkova, I.V., Lambrecht, P.A., Vidaver, A.K., and Harveson, R.M. (2012) Genetic Diversity among Curtobacterium flaccumfaciens pv. flaccumfaciens populations in the American High Plains. *Can. J. Microbiol.* **58**: 788–801.

Aizawa, T., Ve, N.B., Kimoto, K., Iwabuchi, N., Sumida, H., Hasegawa, I., et al. (2007) Curtobacterium ammoniigenes sp. nov., an ammonia-producing bacterium isolated from plants inhabiting acidic swamps in actual acid sulfate soil areas of Vietnam. *Int. J. Syst. Evol. Microbiol.* **57**: 1447–1452.

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.* **40**: e126–e126.

Allison, S.D. (2012) A trait-based approach for modelling microbial litter decomposition. *Ecol. Lett.* **15**: 1058–1070.

Allison, S.D., Lu, Y., Weihe, C., Goulden, M.L., Martiny, A.C., Treseder, K.K., and Martiny, J.B.H. (2013) Microbial abundance and composition influence litter decomposition response to environmental change. *Ecology* **94**: 714–725.

Alster, C.J., German, D.P., Lu, Y., and Allison, S.D. (2013) Microbial enzymatic responses to drought and to nitrogen addition in a southern California grassland. *Soil Biol. Biochem.* **64**: 68–79.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amend, A., Garbelotto, M., Fang, Z., and Keeley, S. (2010) Isolation by landscape in populations of a prized edible mushroom Tricholoma matsutake. *Conserv. Genet.* **11**: 795–802.

Amend, A.S., Martiny, A.C., Allison, S.D., Berlemont, R., Goulden, M.L., Lu, Y., et al. (2016) Microbial response to simulated global change is phylogenetically conserved and linked with functional potential. *ISME J.* **10**: 109–118.

Andam, C.P., Doroghazi, J.R., Campbell, A.N., Kelly, P.J., Choudoir, M.J., and Buckley, D.H. (2016) A latitudinal diversity gradient in terrestrial bacteria of the genus Streptomyces. *MBio* **7**: e02200-15.

Araújo, W.L., Maccheroni Jr, W., Aguilar-Vildoso, C.I., Barroso, P.A. V, Saridakis, H.O., and Azevedo, J.L. (2001) Variability and interactions between endophytic bacteria and fungi isolated from leaf tissues of citrus rootstocks. *Can. J. Microbiol.* **47**: 229–236.

Arevalo, P., VanInsberghe, D., and Polz, M.F. (2018) A Reverse Ecology Framework for Bacteria and Archaea.

Avise, J.C. (2000) Phylogeography: the history and formation of species Harvard university press.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008) The RAST

Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75.
Baker, N.R. and Allison, S.D. (2017) Extracellular enzyme kinetics and thermodynamics along a climate gradient in southern California. *Soil Biol. Biochem.* **114**: 82–92.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**: 455–477.
Becraft, E.D., Cohan, F.M., Kühl, M., Jensen, S.I., and Ward, D.M. (2011) Fine-scale distribution patterns of Synechococcus ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* **77**: 7689–97.
Behrendt, U., Ulrich, A., Schumann, P., Naumann, D., and Suzuki, K.-I. (2002) Diversity of grass-associated Microbacteriaceae isolated from the phyllosphere and litter layer after mulching the sward; polyphasic characterization of Subtercola pratensis sp. nov., Curtobacterium herbarum sp. nov. and Plantibacter flavus gen. nov., sp. *Int. J. Syst. Evol. Microbiol.* **52**: 1441–1454.
Behrendt, U., Ulrich, A., Schumann, P., Naumann, D., and Suzuki, K.I. (2002) Diversity of grass-associated Microbacteriaceae isolated from the phyllosphere and litter layer after mulching the sward; polyphasic characterization of Subtercola pratensis sp. nov., Curtobacterium herbarum sp. nov. and Plantibacter flavus gen. nov., sp. *Int. J. Syst. Evol. Microbiol.* **52**: 1441–1454.
Bendall, M.L., Stevens, S.L.R., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., et al. (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**: 1589.
Benhamou, N., Gagné, S., Le Quéré, D., and Dehbi, L. (2000) Bacterial-mediated induced resistance in cucumber: beneficial effect of the endophytic bacterium Serratia plymuthica on the protection against infection by Pythium ultimum. *Phytopathology* **90**: 45–56.
Bennett, A.F., Lenski, R.E., and Mittler, J.E. (1992) Evolutionary adaptation to temperature. I. Fitness responses of Escherichia coli to changes in its thermal environment. *Evolution (N. Y).* 16–30.
Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2008) GenBank. *Nucleic Acids Res.* **36**: D25.
Berlemont, R., Allison, S.D., Weihe, C., Lu, Y., Brodie, E.L., Martiny, J.B.H., and Martiny, A.C. (2014) Cellulolytic potential under environmental changes in microbial communities from grassland litter. *Front. Microbiol.* **5**: 1–10.
Berlemont, R. and Martiny, A.C. (2015) Genomic potential for polysaccharide deconstruction in bacteria. *Appl. Environ. Microbiol.* **81**: 1513–1519.
Berlemont, R. and Martiny, A.C. (2013) Phylogenetic distribution of potential cellulases in bacteria. *Appl. Environ. Microbiol.* **79**: 1545–1554.
Bi, D., Xu, Z., Harrison, E.M., Tai, C., Wei, Y., He, X., et al. (2011) ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.* **40**: D621–D626.
Bonham, K.S., Wolfe, B.E., and Dutton, R.J. (2017) Extensive horizontal gene transfer in cheese-associated bacteria. *Elife* **6**: e22144.
Branco, S., Gladieux, P., Ellison, C.E., Kuo, A., LaButti, K., Lipzen, A., et al. (2015) Genetic isolation between two recently diverged populations of a symbiotic fungus. *Mol. Ecol.* **24**:

2747–2758.
Brown, M. V, Lauro, F.M., DeMaere, M.Z., Muir, L., Wilkins, D., Thomas, T., et al. (2012) Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* **8**: 595.
Bulgari, D., Casati, P., Brusetti, L., Quaglino, F., Brasca, M., Daffonchio, D., and Bianco, P.A. (2009) Endophytic bacterial diversity in grapevine (Vitis vinifera L.) leaves described by 16S rRNA gene sequence analysis and length heterogeneity-PCR. *J. Microbiol.* **47**: 393–401.
Bulgari, D., Casati, P., Crepaldi, P., Daffonchio, D., Quaglino, F., Brusetti, L., and Bianco, P.A. (2011) Endophytic bacterial community is restructured in grapevine yellows-diseased and recovered Vitis vinifera L. plants. *Appl. Environ. Microbiol.* AEM-00051.
Bushnell, B. (2016) BBMap short read aligner. *Univ. California, Berkeley, California. URL http//sourceforge. net/projects/bbmap*.
Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., et al. (2012) Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol.* **10**: e1001265.
Cadotte, M.W., Arnillas, C.A., Livingstone, S.W., and Yasui, S.-L.E. (2015) Predicting communities from functional traits. *Trends Ecol. Evol.* **30**: 510–511.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.
Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**: 335.
Chase, A.B., Arevalo, P., Polz, M.F., Berlemont, R., and Martiny, J.B.H. (2016) Evidence for ecological flexibility in the cosmopolitan genus Curtobacterium. *Front. Microbiol.* **7**: 1874.
Chase, A.B., Gomez-Lunar, Z., Lopez, A.E., Li, J., Allison, S.D., Martiny, A.C., and Martiny, J.B.H. (2018) Emergence of soil bacterial ecotypes along a climate gradient. *Environ. Microbiol.* **20**: 4112–4126.
Chase, A.B., Karaoz, U., Brodie, E.L., Gomez-Lunar, Z., Martiny, A.C., and Martiny, J.B.H. (2017) Microdiversity of an Abundant Terrestrial Bacterium Encompasses Extensive Variation in Ecologically Relevant Traits. *MBio* **8**: e01809-17.
Chase, A.B. and Martiny, J.B.H. (2018) The importance of resolving biogeographic patterns of microbial microdiversity. *Microbiol. Aust.* **39**: 5–8.
Chase, J.M. and Leibold, M.A. (2003) Ecological niches: linking classical and contemporary approaches University of Chicago Press, Chicago.
Cho, J.-C. and Tiedje, J.M. (2000) Biogeography and degree of endemicity of fluorescent Pseudomonas strains in soil. *Appl. Environ. Microbiol.* **66**: 5448–5456.
Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., et al. (2016) Strategies to improve reference databases for soil microbiomes. *ISME J.* **11**: 1–6.
Choudhary, D.K. and Johri, B.N. (2011) Ecological significance of microdiversity: coexistence among casing soil bacterial strains through allocation of nutritional resource. *Indian J. Microbiol.* **51**: 8–13.
Choudoir, M.J. and Buckley, D.H. (2018) Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of Streptomyces sister-taxa. *ISME J.* 1.
Choudoir, M.J., Doroghazi, J.R., and Buckley, D.H. (2016) Latitude delineates patterns of biogeography in terrestrial Streptomyces. *Environ. Microbiol.* **18**: 4931–4945.
Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure.

Austral Ecol. **18**: 117–143.
Cohan, F.M. (2001) Bacterial species and speciation. *Syst. Biol.* **50**: 513–524.
Cole, J.R., Konstantinidis, K., Farris, R.J., and Tiedje, J.M. (2010) Microbial diversity and phylogeny. *Env. Mol Microbiol* **17**: 1339–1346.
Conner, R.L., Balasubramanian, P., Erickson, R.S., Huang, H.C., and Mündel, H.-H. (2008) Bacterial wilt resistance in kidney beans. *Can. J. Plant Sci.* **88**: 1109–1113.
Connor, N., Sikorski, J., Rooney, A.P., Kopac, S., Koeppel, A.F., Burger, A., et al. (2010) Ecology of speciation in the genus Bacillus. *Appl. Environ. Microbiol.* **76**: 1349–1358.
Consortium, T.H.M.P. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–214.
Cordero, O.X. and Polz, M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**: 263–273.
Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., et al. (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.*
Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., Bik, H.M., and Eisen, J.A. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.
Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**: e11147.
Delmont, T.O. and Eren, A.M. (2018) Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **6**: e4320.
Denef, V.J. and Banfield, J.F. (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science (80-.).* **336**: 462–466.
DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**: 5069–5072.
Diaz, S., Cabido, M., and Casanoves, F. (1998) Plant functional traits and environmental filters at a regional scale. *J. Veg. Sci.* **9**: 113–122.
Didelot, X. and Wilson, D.J. (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**: e1004041.
Dıaz, S. and Cabido, M. (2001) Vive la difference: plant functional diversity matters to ecosystem processes. *Trends Ecol. Evol.* **16**: 646–655.
Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., et al. (2011) Geneious, version 5.4. *Geneious, Auckland, New Zeal.*
Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C., and Fitter, A.H. (2009) Relative roles of niche and neutral processes in structuring a soil microbial community. *Isme J.* **4**: 337.
Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.
Edgar, R.C. (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **1**: 5.
Edwards, K.F., Litchman, E., and Klausmeier, C.A. (2013) Functional traits explain phytoplankton community structure and seasonal dynamics in a marine ecosystem. *Ecol. Lett.* **16**: 56–63.
Edwards, U., Rogall, T., Blöcker, H., Emde, M., and Böttger, E.C. (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for

16S ribosomal RNA. *Nucleic Acids Res.* **17**: 7843–7853.
Elbeltagy, A., Nishioka, K., Suzuki, H., Sato, T., Sato, Y.-I., Morisaki, H., et al. (2000) Isolation and characterization of endophytic bacteria from wild and traditionally cultivated rice varieties. *Soil Sci. plant Nutr.* **46**: 617–629.
Enquist, B.J., Norberg, J., Bonser, S.P., Violle, C., Webb, C.T., Henderson, A., et al. (2015) Scaling from traits to ecosystems: developing a general trait driver theory via integrating trait-based and metabolic scaling theories. In, *Advances in Ecological Research*. Elsevier, pp. 249–318.
Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**: 1575–1584.
Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**: 968.
Eren, A.M., Sogin, M.L., Morrison, H.G., Vineis, J.H., Fisher, J.C., Newton, R.J., and McLellan, S.L. (2015) A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* **9**: 90–100.
Evtushenko, L. and Takeuchi, M. (2006) The prokaryotes: Springer, New York.
Federhen, S. (2012) The NCBI Taxonomy. *Nucleic Acids Res.* **40**: D136--D143.
Fierer, N. and Jackson, R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 626–631.
Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**: 29–37.
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., et al. (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44**: D279–D285.
Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., et al. (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proc. Natl. Acad. Sci.* **110**: 9824–9829.
Fontes, C.M.G.A. and Gilbert, H.J. (2010) Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu. Rev. Biochem.* **79**: 655–681.
Fraser, C., Hanage, W.P., and Spratt, B.G. (2007) Recombination and the nature of bacterial speciation. *Science (80-.).* **315**: 476–480.
Frese, S.A., Benson, A.K., Tannock, G.W., Loach, D.M., Kim, J., Zhang, M., et al. (2011) The evolution of host specialization in the vertebrate gut symbiont Lactobacillus reuteri. *PLoS Genet* **7**: e1001314.
Frichot, E. and François, O. (2015) LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**: 925–929.
Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* genetics-113.
Garcia-Martinez, J. and Rodriguez-Valera, F. (2000) Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Mol. Ecol.* **9**: 935–948.
Gilbert, J.A., Jansson, J.K., and Knight, R. (2014) The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**: 69.
Giovannoni, S.J., Rappé, M.S., Vergin, K.L., and Adair, N.L. (1996) 16S rRNA genes reveal

stratified open ocean bacterioplankton populations related to the green non-sulfur bacteria. *Proc. Natl. Acad. Sci.* **93**: 7979–7984.

Gonzalez, A., Vázquez-Baeza, Y., Pettengill, J.B., Ottesen, A., McDonald, D., and Knight, R. (2016) Avoiding pandemic fears in the subway and conquering the Platypus. *mSystems* **1**: e00050-16.

Goodfellow, M. and Williams, S.T. (1983) Ecology of actinomycetes. *Annu. Rev. Microbiol.* **37**: 189–216.

Green, J.L., Bohannan, B.J.M., and Whitaker, R.J. (2008) Microbial biogeography: from taxonomy to traits. *Science (80-.).* **320**: 1039–1043.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., and Martiny, J.B.H. (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* 1–10.

Harrell, F. (2015) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis Springer.

Hartel, P.G. and Alexander, M. (1986) Role of extracellular polysaccharide production and clays in the desiccation tolerance of cowpea Bradyrhizobia. *Soil Sci. Soc. Am. J.* **50**: 1193–1198.

Harveson, R.M., Schwartz, H.F., Vidaver, A.K., Lambrecht, P.A., and Otto, K.L. (2006) New outbreaks of bacterial wilt of dry bean in Nebraska observed from field infections. *Plant Dis.* **90**: 681.

Hedges, F. (1926) Bacterial wilt of Beans (Bacterium flaccum-faciens Hedges), including comparisons with Bacterium phaseoli. *Phytopathology* **16**:.

Hehemann, J.-H., Arevalo, P., Datta, M.S., Yu, X., Corzett, C.H., Henschel, A., et al. (2016) Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nat. Commun.* **7**:.

Hsieh, T.F., Huang, H.C., Mündel, H., Conner, R.L., Erickson, R.S., and Balasubramanian, P.M. (2005) Resistance of common bean (Phaseolus vulgaris) to bacterial wilt caused by Curtobacterium flaccumfaciens pv. flaccumfaciens. *J. Phytopathol.* **153**: 245–249.

Huang, H.C., Erickson, R.S., Hsieh, T.F., Conner, R.L., and Balasubramanian, P.M. (2009) Resurgence of bacterial wilt of common bean in North America. **300**: 290–300.

Hubbell, S.P. (2001) The Unified Neutral Theory of Biodiversity and Biogeography Princeton University Press.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., and Polz, M.F. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science (80-.).* **320**: 1081–1085.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Ingle, D.J., Tauschek, M., Edwards, D.J., Hocking, D.M., Pickard, D.J., Azzopardi, K.I., et al. (2016) Evolution of atypical enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants. *Nat. Microbiol.* **1**: 15010.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2017) High-throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *bioRxiv* 225342.

Jaspers, E. and Overmann, J. (2004) Ecological significance of microdiversity: identical 16S rRNA

gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl. Environmantal Microbiol.* **70**: 4831–4839.
Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**: 431.
Johnson, N.C., Graham, J., and Smith, F.A. (1997) Functioning of mycorrhizal associations along the mutualism–parasitism continuum. *New Phytol.* **135**: 575–585.
Johnson, Z.I., Zinser, E.R., Coe, A., Mcnulty, N.P., Malcolm, E.S., Chisholm, S.W., et al. (2006) Partitioning among Prochlorococcus ecotypes along environmental gradients. *Science (80-.).* **311**: 1737–1740.
Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., and Chisholm, S.W. (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science (80-.).* **311**: 1737–1740.
Júnior Silva, T.A.F.S., Negrão, D.R., Itako, A.T., Soman, J.M., and Maringoni, A.C. (2012) Survival of Curtobacterium flaccumfaciens pv. flaccumfaciens in soil and bean crop debris. *J. Plant Pathol.* **94**: 331–337.
Kasana, R.C., Salwan, R., Dhar, H., Dutt, S., and Gulati, A. (2008) A rapid and easy method for the detection of Mmcrobial cellulases on agar plates using Gram's Iodine. *Curr. Microbiol.* **57**: 503–507.
Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., et al. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science (80-.).* **344**: 416–420.
Kent, A.G., Dupont, C.L., Yooseph, S., and Martiny, A.C. (2016) Global biogeography of Prochlorococcus genome diversity in the surface ocean. *ISME J.* **10**: 1856.
Kim, M.K., Kim, Y.-J., Kim, H.-B., Kim, S.-Y., Yi, T.-H., and Yang, D.-C. (2008) Curtobacterium ginsengisoli sp. nov., isolated from soil of a ginseng field. *Int. J. Syst. Evol. Microbiol.* **58**: 2393–2397.
Kogel, K.-H., Franken, P., and Hückelhoven, R. (2006) Endophyte or parasite–what decides? *Curr. Opin. Plant Biol.* **9**: 358–363.
Komagata, K., IIzuka, H., and Takahashi, M. (1965) Taxonomic Evaluation of Nitrate Respiration and Carbohydrate Fermentation in Aerobic Bacteria. *J. Gen. Appl. Microbiol.* **11**: 191–201.
Konstantinidis, K.T. and Tiedje, J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10**: 504–509.
Konstantinidis, K.T. and Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**: 6258–6264.
Korkama-Rajala, T., Müller, M.M., and Pennanen, T. (2008) Decomposition and fungi of needle litter from slow-and fast-growing Norway spruce (Picea abies) clones. *Microb. Ecol.* **56**: 76.
Lacava, P.T., Li, W., Arau´jo, W.L., Azevedo, J.L., and Hartung, J.S. (2007) The endophyte Curtobacterium flaccumfaciens reduces symptoms caused by Xylella fastidiosa in Catharanthus roseus. *J. Microbiol.* **45**: 388–393.
Lan, Y., Rosen, G., and Hershberg, R. (2016) Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* **4**: 18.
Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**: 357–359.

Larkin, A.A. and Martiny, A.C. (2017) Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environ. Microbiol. Rep.* **9**: 55–70.
Lavorel, S. and Garnier, É. (2002) Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Funct. Ecol.* **16**: 545–556.
Lednická, D., Mergaert, J., Cnockaert, M.C., and Swings, J. (2000) Isolation and identification of cellulolytic bacteria involved in the degradation of natural cellulosic fibres. *Syst. Appl. Microbiol.* **23**: 292–299.
Leff, J.W., Jones, S.E., Prober, S.M., Barberán, A., Borer, E.T., Firn, J.L., et al. (2015) Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl. Acad. Sci.* **112**: 10967–10972.
Lemieux, J.E., Tran, A.D., Freimark, L., Schaffner, S.F., Goethert, H., Andersen, K.G., et al. (2016) A global map of genetic diversity in Babesia microti reveals strong population structure and identifies variants associated with clinical relapse. *Nat. Microbiol.* **1**: 16079.
Lennon, J.T., Aanderud, Z.T., Lehmkuhl, B.K., and Schoolmaster, D.R. (2012) Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology* **93**: 1867–1879.
Lenormand, T. (2002) Gene flow and the limits to natural selection. *Trends Ecol. Evol.* **17**: 183–189.
Letunic, I. and Bork, P. (2006) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127–128.
Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2014) MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674–1676.
Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
Litchman, E., Edwards, K.F., and Klausmeier, C.A. (2015) Microbial resource utilization traits and trade-offs: implications for community structure, functioning, and biogeochemical impacts at present and in the future. *Front. Microbiol.* **6**: 254.
Litchman, E. and Klausmeier, C.A. (2008) Trait-based community ecology of phytoplankton. *Annu. Rev. Ecol. Evol. Syst.* **39**: 615–639.
Lomolino, M. V, Riddle, B.R., Brown, J.H., and Brown, J.H. (2006) Biogeography Sinauer Associates Sunderland, MA.
Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., et al. (2013) Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. *ISME J.* **7**: 184.
Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006) Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci.* **103**: 12552–12557.
Martiny, A.C., Tai, A.P.K., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009) Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environ. Microbiol.* **11**: 823–832.
Martiny, A.C., Treseder, K., and Pusch, G. (2013) Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7**: 830–8.
Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J. a, Green, J.L., et al. (2006) Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.*

4: 102–112.
Martiny, J.B.H., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**: aac9323.
Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: 538.
Matulich, K.L., Weihe, C., Allison, S.D., Amend, A.S., Berlemont, R., Goulden, M.L., et al. (2015) Temporal variation overshadows the response of leaf litter microbial communities to simulated global change. *ISME J.* **9**: 2477–2489.
Mayr, E. (2001) What evolution is Science Masters Series.
McGill, B.J., Enquist, B.J., Weiher, E., and Westoby, M. (2006) Rebuilding community ecology from functional traits. *Trends Ecol. Evol.* **21**: 178–185.
McGill, B.J., Maurer, B.A., and Weiser, M.D. (2006) Empirical evaluation of neutral theory. *Ecology* **87**: 1411–1423.
McLaren, M.R. and Callahan, B.J. (2018) In Nature, There Is Only Diversity. *MBio* **9**: e02149-17.
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., et al. (2008) The metagenomics RAST server -- a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.
Moore, L.R. and Chisholm, S.W. (1999) Photophysiology of the marine cyanobacterium Prochlorococcus: ecotypic differences among cultured isolates. *Limnol. Oceanogr.* **44**: 628–638.
Moore, L.R., Post, A.F., Rocap, G., and Chisholm, S.W. (2002) Utilization of different nitrogen sources by the marine cyanobacteria Prochlorococcus and Synechococcus. *Limnol. Oceanogr.* **47**: 989–996.
Moore, L.R., Rocap, G., and Chisholm, S.W. (1998) Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature* **393**: 464–467.
Morrison, S.S., Williams, T., Cain, A., Froelich, B., Taylor, C., Baker-Austin, C., et al. (2012) Pyrosequencing-based comparative genome analysis of Vibrio vulnificus environmental isolates. *PLoS One* **7**: e37553.
Mouginot, C., Kawamura, R., Matulich, K.L., Berlemont, R., Allison, S.D., Amend, A.S., and Martiny, A.C. (2014) Elemental stoichiometry of fungi and bacteria strains from grassland leaf litter. *Soil Biol. Biochem.* **76**: 278–285.
Mac Nally, R. (2002) Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodivers. Conserv.* **11**: 1397–1401.
Nannipieri, P., Ascher, J., Ceccherini, M., Landi, L., Pietramellara, G., and Renella, G. (2003) Microbial diversity and soil functions. *Eur. J. Soil Sci.* **54**: 655–670.
Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.
Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K.S. (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**: 1612–1625.
Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Bilinski, T.M., Stanish, L.F., et al. (2013) Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**:

342–356.
Newton, A.C., Fitt, B.D.L., Atkins, S.D., Walters, D.R., and Daniell, T.J. (2010) Pathogenesis, parasitism and mutualism in the trophic space of microbe–plant interactions. *Trends Microbiol.* **18**: 365–373.
O'Toole, G.A. (2011) Microtiter dish biofilm formation assay. *J. Vis. Exp. JoVE*.
Ochman, H., Elwyn, S., and Moran, N.A. (1999) Calibrating bacterial evolution. *Proc. Natl. Acad. Sci.* **96**: 12638–12643.
Ohya, H., Komai, Y., and Yamaguchi, M. (1986) Occurrence of Curtobacterium sp. possessing ω-cyclohexyl fatty acids in soil with zinc added. *Arch. Microbiol.* **145**: 9–12.
Osdaghi, E., Pakdaman Sardrood, B., Bavi, M., Akbari Oghaz, N., Kimiaei, S., and Hadian, S. (2015) First Report of *Curtobacterium flaccumfaciens* pv. *flaccumfaciens* Causing Cowpea Bacterial Wilt in Iran. *J. Phytopathol.* **163**: 653–656.
Osdaghi, E., Taghavi, S.M., Fazliarab, A., Elahifard, E., and Lamichhane, J.R. (2015) Characterization, geographic distribution and host range of Curtobacterium flaccumfaciens: an emerging bacterial pathogen in Iran. *Crop Prot.* **78**: 185–192.
Osono, T. (2006) Role of phyllosphere fungi of forest trees in the development of decomposer fungal communities and decomposition processes of leaf litter. *Can. J. Microbiol.* **52**: 701–716.
Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., et al. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**: 3691–3693.
Partensky, F., Hess, W.R., and Vaulot, D. (1999) Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**: 106–127.
Pereira, S.L. and Baker, A.J. (2006) A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol. Biol. Evol.* **23**: 1731–1740.
Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**: 1929–1936.
Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B., and Hallin, S. (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* **8**: 523–529.
Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2011) R Development Core Team. 2010. nlme: linear and nonlinear mixed effects models. R package version 3.1-97. *R Found. Stat. Comput. Vienna*.
Pold, G., Billings, A.F., Blanchard, J.L., Burkhardt, D.B., Frey, S.D., Melillo, J.M., et al. (2016) Long-term warming alters Ccrbohydrate degradation potential in temperate forest soils. **82**: 6518–6530.
Polz, M.F., Alm, E.J., and Hanage, W.P. (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29**: 170–175.
Polz, M.F., Hunt, D.E., Preheim, S.P., and Weinreich, D.M. (2006) Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philos. Trans. R. Soc. B Biol. Sci.* **361**: 2009 LP-2021.
Poretsky, R., Rodriguez-R, L.M., Luo, C., Tsementzi, D., and Konstantinidis, K.T. (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial

community dynamics. *PLoS One* **9**: e93827.
Potts, D.L., Suding, K.N., Winston, G.C., Rocha, A.V., and Goulden, M.L. (2012) Ecological effects of experimental drought and prescribed fire in a southern California coastal grassland. *J. Arid Environ.* **81**: 59–66.
Potts, M. (1994) Desiccation tolerance of prokaryotes. *Microbiol. Rev.* **58**: 755–805.
Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.
Pruesse, E., Peplies, J., and Glöckner, F.O. (2012) SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**: 7188–7196.
Raes, J., Letunic, I., Yamada, T., Jensen, L.J., and Bork, P. (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**: 473.
Ranjard, L. and Richaume, A. (2001) Quantitative and qualitative microscale distribution of bacteria in soil. *Res. Microbiol.* **152**: 707–716.
Raupach, G.S. and Kloepper, J.W. (2000) Biocontrol of cucumber diseases in the field by plant growth-promoting rhizobacteria with and without methyl bromide fumigation. *Plant Dis.* **84**: 1073–1075.
Raupach, G.S. and Kloepper, J.W. (1998) Mixtures of plant growth-promoting rhizobacteria enhance biological control of multiple cucumber pathogens. *Phytopathology* **88**: 1158–1164.
Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. (2015) Inferring horizontal gene transfer. *PLoS Comput. Biol.* **11**: e1004095.
Redman, R.S., Dunigan, D.D., and Rodriguez, R.J. (2001) Fungal symbiosis from mutualism to parasitism: who controls the outcome, host or invader? *New Phytol.* **151**: 705–716.
Rho, M., Tang, H., and Ye, Y. (2010) FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**: 1–12.
Richter, M. and Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* **106**: 19126–19131.
Roberson, E.B. and Firestone, M.K. (1992) Relationship between desiccation and exopolysaccharide production in a soil Pseudomonas sp. *Appl. Environ. Microbiol.* **58**: 1284–1291.
Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., et al. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042.
Rodriguez-R, L.M. and Konstantinidis, K.T. (2014) Bypassing cultivation to identify bacterial species. *Microbe* **9**: 111–118.
Rodriguez-R, L.M. and Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes PeerJ Preprints.
Rodriguez-Valera, F. and Ussery, D.W. (2012) Is the pan-genome also a pan-selectome? *F1000Research* **1**:.
Schimel, J., Balser, T.C., and Wallenstein, M. (2007) Microbial stress-response physiology and its

implications for ecosystem function. *Ecology* **88**: 1386–1394.
Schlatter, D.C. and Kinkel, L.L. (2014) Global biogeography of *Streptomyces* antibiotic inhibition, resistance, and resource use. *FEMS Microbiol. Ecol.* **88**: 386–397.
Schloss, P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* **6**: e1000844.
Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., and Hollister, E.B. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Env. Microbiol* **75**:.
Schloter, M., Lebuhn, M., Heulin, T., and Hartmann, A. (2000) Ecology and evolution of bacterial microdiversity. *FEMS Microbiol. Rev.* **24**: 647–660.
Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.
Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science (80-.).* **336**: 48–51.
Shapiro, B.J., Leducq, J.-B., and Mallet, J. (2016) What is speciation? *PLoS Genet.* **12**: e1005860.
Shapiro, B.J. and Polz, M.F. (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* **22**: 235–247.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**: 539.
Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**: D32–D36.
Simmons, S.L., DiBartolo, G., Denef, V.J., Goltsman, D.S.A., Thelen, M.P., and Banfield, J.F. (2008) Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6**: e177.
Soares, R.M., Fantinato, G.G.P., Darben, L.M., Marcelino-Guimarães, F.C., Seixas, C.D.S., and Carneiro, G.E. de S. (2013) First report of Curtobacterium flaccumfaciens pv. flaccumfaciens on soybean in Brazil. *Trop. Plant Pathol.* **38**: 452–454.
South, A. (2011) rworldmap: A New R package for Mapping Global Data. *R J.* **3**:.
Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
Sturz, A. V, Christie, B.R., Matheson, B.G., Arsenault, W.J., and Buchanan, N.A. (1999) Endophytic bacterial communities in the periderm of potato tubers and their potential to improve resistance to soil-borne plant pathogens. *Plant Pathol.* **48**: 360–369.
Sturz, A. V, Christie, B.R., Matheson, B.G., and Nowak, J. (1997) Biodiversity of endophytic bacteria which colonize red clover nodules, roots, stems and foliage and their influence on host growth. *Biol. Fertil. Soils* **25**: 13–19.
Subramanian, S. (2008) Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics* **178**: 2429–2432.
Taghavi, S., Garafola, C., Monchy, S., Newman, L., Hoffman, A., Weyens, N., et al. (2009) Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees. *Appl. Environ. Microbiol.* **75**: 748–757.

Takeuchi, N., Cordero, O.X., Koonin, E. V, and Kaneko, K. (2015) Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* **13**: 20.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**: 6614–6624.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**: 472–477.

Tharayil, N., Suseela, V., Triebwasser, D.J., Preston, C.M., Gerard, P.D., and Dukes, J.S. (2011) Changes in the structural composition and reactivity of Acer rubrum leaf litter tannins exposed to warming and altered precipitation: climatic stress-induced tannins are more reactive. *New Phytol.* **191**: 132–145.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science (80-.).* **307**: 1311–1313.

VanInsberghe, D., Maas, K.R., Cardenas, E., Strachan, C.R., Hallam, S.J., and Mohn, W.W. (2015) Non-symbiotic Bradyrhizobium ecotypes dominate North American forest soils. *ISME J.* **9**: 2435.

Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., and Pati, A. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**: 6761–6771.

Vidaver, A.K. (1982) The plant pathogenic corynebacteria. *Annu. Rev. Microbiol.* **36**: 495–517.

Vieira-Silva, S. and Rocha, E.P.C. (2010) The systemic imprint of growth and its uses in ecological (meta) genomics. *PLoS Genet.* **6**: e1000808.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**:.

Ward, D.M., Cohan, F.M., Bhaya, D., Heidelberg, J.F., Kühl, M., and Grossman, A. (2008) Genomics, environmental genomics and the issue of microbial species. *Heredity (Edinb).* **100**: 207.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., et al. (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**: 581–591.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science (80-.).* **301**: 976–978.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2005) Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol. Evol.* **22**: 2354–2361.

Whitman, W., Goodfellow, M., Kämpfer, P., Busse, H.-J., Trujillo, M., Ludwig, W., et al. eds. (2012) Bergey's Manual of Systematic Bacteriology: Volume 5: The Actinobacteria 2nd ed. Springer-Verlag, New York.

Wielbo, J., Marek-Kozaczuk, M., Kubik-Komar, A., and Skorupska, A. (2007) Increased metabolic potential of Rhizobium spp. is associated with bacterial competitiveness. *Can. J. Microbiol.* **53**: 957–967.

Wiens, J.J. and Donoghue, M.J. (2004) Historical biogeography, ecology and species richness.

Trends Ecol. Evol. **19**: 639–644.
Williams, T.C., Blackman, E.R., Morrison, S.S., Gibas, C.J., and Oliver, J.D. (2014) Transcriptome sequencing reveals the virulence and environmental genetic programs of Vibrio vulnificus exposed to host and estuarine conditions. *PLoS One* **9**: e114376.
Wilson, D.B. (2011) Microbial diversity of cellulose hydrolysis. *Curr. Opin. Microbiol.* **14**: 259–263.
Woebken, D., Lam, P., Kuypers, M.M.M., Naqvi, S., Kartal, B., Strous, M., et al. (2008) A microdiversity study of anammox bacteria reveals a novel Candidatus Scalindua phylotype in marine oxygen minimum zones. *Environ. Microbiol.* **10**: 3106–3119.
Wood, B.A. and Easdown, W.J. (1990) A new bacterial disease of mung bean and cowpea for Australia. *Australas. Plant Pathol.* **19**: 16–21.
Wright, S. (1943) Isolation by distance. *Genetics* **28**: 114.
Wu, D., Jospin, G., and Eisen, J.A. (2013) Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One* **8**: e77033.
Yao, H., Gao, Y., Nicol, G.W., Campbell, C.D., Prosser, J.I., Zhang, L., et al. (2011) Links between ammonia oxidizer community structure, abundance, and nitrification potential in acidic soils. *Appl. Environ. Microbiol.* **77**: 4618–4625.
Yawata, Y., Cordero, O.X., Menolascina, F., Hehemann, J.-H., Polz, M.F., and Stocker, R. (2014) Competition–dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proc. Natl. Acad. Sci.* **111**: 5622–5627.
Young, J.M., Saddler, G.S., Takikawa, Y., De Boer, S.H., Vauterin, L., Gardan, L., et al. (1996) Names of plant pathogenic bacteria 1864-1995. *Rev. Plant Pathol.* **75**: 721–763.