Sympatric and allopatric differentiation delineates population structure in free-living terrestrial bacteria

Alexander B. Chase[1,2,5*], Philip Arevalo[3,6], Eoin L. Brodie[2,4], Martin F. Polz[3], Ulas Karaoz[2], and Jennifer B.H. Martiny[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA
[2]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[5]Present Address: Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, CA, USA
[6]Present Address: Department of Ecology and Evolutionary Biology, University of Chicago, Chicago, IL, USA

*Corresponding Author: AB Chase, abchase@ucsd.edu

ABSTRACT

In free-living bacteria and archaea, the equivalent of the biological species concept does not exist, creating several barriers to the study of the processes contributing to microbial diversification. These barriers are particularly high in soil, where high bacterial diversity inhibits the study of closely-related genotypes and therefore the factors structuring microbial populations. We isolated strains within a single *Curtobacterium* ecotype from surface soil (leaf litter) across a regional climate gradient and investigated the phylogenetic structure, recombination, and flexible gene content of this genomic diversity to infer patterns of gene flow. Our results indicate that microbial populations are delineated by gene flow discontinuities and exhibit evidence for population-specific adaptation. We conclude that the genetic structure within this bacterium is due to both adaptation within localized microenvironments (isolation-by-environment) as well as dispersal limitation between geographic locations (isolation-by-distance).

KEYWORDS: Curtobacterium, population structure, gene flow, microbial ecology, ecotype

INTRODUCTION

In eukaryotes, populations are typically defined as groups of interbreeding individuals within a species residing in the same geographic area (1). Geographically-distinct (i.e., allopatric) populations are also often genetically distinct because of reduced gene flow, or the exchange of genetic variation, between populations of the same species. However, in microorganisms, the equivalent of the biological species concept does not exist, creating several barriers to the study of the fine-scale genetic structure of microbial populations and thus, the processes contributing to microbial diversification (2–4).

The first of these barriers is that the genetic resolution delineating a microbial population is unclear. In eukaryotes, populations are, by definition, genetic units belonging to the same species, but defining a prokaryotic species remains challenging (5). Nonetheless, there is evidence for geographically-distinct, genetically-diverged groups of bacteria and archaea. Several studies have shown that the genetic similarity of closely-related microbial individuals are negatively correlated with geographic distance across continental and global scales (6–9). This pattern is consistent with isolation-by-distance, whereby dispersal limitation contributes to reproductive isolation over geographic distances (10). Further, in some cases, these geographically-localized genetic clades appear to be adapted to local environmental conditions, as individuals within these clades can differ in their temperature (11), nutrient (12), or habitat preference (13). However, the degree of divergence between genetic clades in such studies is usually quite high (<90% genome-wide average nucleotide identity), indicating they may not represent intra-species relationships (14). These genetic units would seem to be much broader than populations, or groups of individuals with the potential for contemporary interactions and exchange of genetic material (15). Therefore, a focus on much more closely-related microorganisms is needed to investigate the processes responsible for initial diversification.

A second, related obstacle is recovering genetically-similar individuals of the same species, however defined. Population genetic studies of eukaryotes typically characterize the genetic diversity among many individuals from a variety of geographic locations. For microbes, this sampling design requires reliable isolation of closely-related strains (but see (16)), which can be difficult in highly diverse microbial communities such as soil. Finally, even if a sample of closely-related individuals can be collected, a third barrier is quantifying the exchange of genetic variation (i.e., gene flow) between individuals. For prokaryotes, the exchange of genetic material is mediated through genetic recombination, whether homologous recombination or horizontal transfer of entirely new genes. However, the asexual nature of prokaryotes makes it a challenge to quantify this process, particularly among closely-related individuals. The more closely-related two genomes are, the more difficult it is to distinguish between differences caused by vertical inheritance and recombination (17).

In aquatic (18) and host-associated (19) systems, many of these obstacles have been addressed. In these environments, geographic proximity does not appear to be the most important factor in structuring microbial populations as typically observed in plants and animals. Indeed, an increasing number of studies find several, distinct genetic clades co-occurring in the same geographic location (20–23). For instance, the thermophilic archaeon, *Sulfolobus*, exhibited strong barriers to recombination between sympatric clades within a hotspring (24). Such evidence suggests that the genetic structure of microbial populations is influenced less by divergence among geographically-distinct (allopatric) groups, and more by ecological differentiation (isolation-by-environment (25)) among co-occurring (sympatric) groups (26). Thus, we might need to abandon the idea of defining microbial populations *a priori* based on geography (as done for larger organisms) and, instead, focus first on the emerging genetic structure among closely-related individuals (27).

Soils are highly heterogeneous systems where differences in microhabitats can contribute to environmental variation over many spatial scales (28,29). For this reason, one might expect that allopatric differentiation might be more evident for soil bacteria than that observed in aquatic or host-associated environments. Indeed, some soil fungi exhibit strong population structure at regional spatial scales (30,31). Therefore, we asked whether population structure in a free-living soil bacterium was consistent with patterns of allopatric or sympatric speciation. To do so, we investigated the abundant leaf litter taxon, *Curtobacterium* (32), which is relatively easy to culture from the leaf litter layer of soil. Previously, we demonstrated that *Curtobacterium* encompasses multiple ecotypes, or fine-scale genetic clades that correspond to ecologically relevant phenotypes (33). Here, we concentrated on the genetic diversity within a single ecotype, *Curtobacterium* Subclade IB/C, a unit that might be considered equivalent to a species designation (33). Specifically, we examined 26 strains (with identical full-length 16S rRNA regions and >97% mean genome-wide average amino acid identity) from a regional climate gradient, along with two closely-related strains isolated across continental distances. We hypothesized that soil bacteria would exhibit a pattern intermediate to that of aquatic free-living bacteria and archaea and soil fungi. In particular, we expected that sympatric populations of soil bacteria may exist within a particular geographic location, while also exhibiting a pattern consistent with allopatric differentiation among locations. Such a pattern would indicate that the genetic structure within this bacterium is due to both adaptation within localized microenvironments (isolation-by-environment) as well as dispersal limitation between geographic locations (isolation-by-distance).

RESULTS

Evolutionary History within a *Curtobacterium* ecotype. We identified 26 strains from a *Curtobacterium* ecotype, subclade IB/C, that share ecologically-relevant genotypic and phenotypic characteristics. These traits include the ability to degrade polymeric carbohydrates (i.e., cellulose and xylan), the degree of biofilm formation, and temperature preference for both growth and carbon degradation (33). These strains were previously isolated from leaf litter, the top layer of soil, at four geographic locations from a regional climate gradient in southern California (Supplementary Table 1). All analyzed strains have identical full-length 16S rRNA regions and share high sequence identity with ≥94.6% genome average nucleotide identity (ANI) and ≥95.3% genome average amino acid identity (AAI), congruent with previous observations for defining discrete sequence clusters within natural microbial communities (34). We also included two additional strains from subclade IB/C that were isolated from leaf litter in Boston, MA to provide varying geographic scales ($ANI_{[MEAN\ SIMILARITY]}$ = 94.9%; $AAI_{[MEAN\ SIMILARITY]}$ = 95.8%).

To examine whether genetically-similar strains within the IB/C subclade clustered by geographic location, we reconstructed the phylogenetic relationship among the strains using the core genome (Fig. 1A). The core genome phylogeny revealed highly structured genetic lineages; however, clusters contained strains isolated from a variety of geographic locations. While one strain from Boston, MA formed the outgroup, the other Boston strain was highly similar to a grassland strain from Loma Ridge, CA. At the regional scale within the climate gradient, most of the grassland strains clustered together, while strains from the scrubland and Salton Sea leaf litter communities were dispersed throughout the tree.

Phylogenetic analyses alone cannot delineate population structure as it is necessary to account for both vertical descent and contributions from shared ancestral gene pools. Therefore, we supplemented the phylogenetic analysis by computing ancestry coefficients for each strain across the core genome using a STRUCTURE-like (35) analysis (Fig. 1B). The most probable number of ancestral gene pools (K=4) contributing to the proportion of an individual genome (see Materials and Methods) demonstrated high congruence with the phylogenetic analysis. For example, an outgroup strain originating from Boston, MA exhibited little evidence for mixing with most of the climate gradient strains in CA across continental scales (Fig. 1B). Within the regional climate gradient, we detected three ancestral gene pools that may represent finer population structure across ecologically-similar strains in ecotype IB/C.

Gene Flow Delineates Bacterial Populations. Although STRUCTURE-like analyses can provide insights into the genetic structure among divergent lineages, populations (defined as groups with the potential to exchange genetic material) must be resolved by examining patterns of gene flow. However, in asexual organisms, measurements of homologous recombination can be overestimated when individuals are closely related as distinguishing between recombination and point mutations is difficult (17). Further, other forms of horizontal gene transfer can be ecologically relevant as well (36). To address these limitations, we employed a novel method, PopCOGenT, that attempts to detect all recent recombination events between pairs of strains (27).

To distinguish between vertical descent and homologous recombination in structuring populations, we used PopCOGenT to estimate the degree of recombination among the genomes. This analysis revealed three recombining populations that are evident as highly isolated clusters in the network (Fig. 2). One of the populations (population 2) was restricted to a single location (in the grassland site). The other two populations included strains from multiple sites along the climate gradient; for example, population 3 contained strains isolated from the grassland, scrubland, and Salton Sea leaf litter communities, which are geographically separated by 177 km (Supplementary Table 2).

This approach enabled the identification of recombining populations that would otherwise be masked with traditional phylogenetic analyses. For example, two strains (MMLR14002/014) isolated from the grassland site five years prior share no recent recombination events (Fig. 2) despite sharing a high degree of phylogenetic relatedness and a common ancestral gene pool to strains within population 3 (Fig. 1). Additionally, the analysis revealed that the highly similar strains isolated across the continent from one another (from Boston, MA and a CA grassland; Fig. 1) were not connected by recent recombination events. Indeed, this conservative approach to estimate recombination events reduced most strains within the IB/C subclade to singleton nodes, suggesting that no recent recombination events connect these individuals to the three identified populations (Fig. 2), and that these strains are probably representatives of other, unsampled populations.

To confirm the effect of homologous recombination on the genetic diversity within subclade IB/C, we employed ClonalFrameML (37). Specifically, we concentrated on the r/m ratio at which nucleotides are substituted from either recombination or point mutations. Throughout the evolution of the IB/C subclade, recombination rates were generally low (r/m = 0.94), indicating barriers to gene flow and the occurrence of mutation accumulation within the subclade. However, when we assessed the rates of recombination within each population assignment, we found that homologous recombination rates to be high in populations 1 and 3 (r/m = 3.34 and 2.75, respectively) while population 2 (r/m = 1.62) had intermediate recombination rates (Supplementary Table 2). The observed r/m values are especially notable as terrestrial free-living bacteria have previously been shown to have low r/m values (r/m <1) (38).

Population Differentiation of the Flexible Genome. Based on the recombination networks, we expected that individuals within the same population would also share more flexible genes (genes not present in all strains) than individuals between different populations. The similarity between flexible gene content among strains was highly congruent with the population assignments (Fig. 3); strains within a population (ANOSIM; R = 0.88, p = 0.001) shared more flexible genes than expected by chance. We also observed that flexible gene content differed significantly by site (ANOSIM; R = 0.81, p < 0.01), suggesting that processes within and across locations are structuring the differences in the flexible genome within the subclade IB/C.

The flexible genome also provides insights into the traits that distinguish populations. For example, flexible genes only present in all individuals within a particular population may have swept through the population by positive selection (26). We searched for population-specific genes shared among all members and discovered that many were highly localized to a limited number of genomic regions. Specifically, 16 of 48 population-specific genes in population 1 were highly localized in the genome, while 4 of 6 population-specific genes in population 3 were localized (Fig. 4A). Additionally, these population-specific genes had reduced nucleotide diversity when compared to whole-genome measurements (Supplementary Figure S1), which can be indicative of relatively recent selective sweeps. These putative sweep regions may have arrived prior to population diversification and subsequently co-diversified, but, nonetheless, represent genomic regions harboring population-specific flexible genes. We did not detect any localization of population-specific genes in population 2, perhaps due to its lower rate of homologous recombination (Supplementary Table 2).

The flanking genomic regions surrounding the population-specific genes exhibited high genomic conservatism across all members in the population as well, suggesting these genomic regions may be hotspots for genetic exchange within the populations (Fig. 4A). While we did not detect phage or integrative and conjugative elements (ICEs), we did identify other mobile genetic elements such as insertion sequences and clustered regularly interspaced short palindromic repeats (CRISPRs). Further, the regions were littered with pseudogenic exons, indicating the interruption of functional proteins due to recombining genomic segments. The genomic regions also contained rare (<25% of all members within Subclade IB/C) or strain-specific genes. In contrast to these variable regions, the flanking genes were highly conserved (shared by >85% of all members within Subclade IB/C) in nearly identical genetic architectures. Many of the conserved flanking core genes supported a strict monophyletic division of the population (Fig. 4B), suggesting integration of population-specific genes is mediated by homologous recombination of the conserved flanking homologous gene regions (4).

Most of the population-specific genes within the variable regions annotate as hypothetical proteins with some transcriptional regulators; however, other genes may be involved in differential use of environmental resources. For example, the regions contained a high number of metal uptake and transport proteins, along with glycoside hydrolase (GH) enzymes and glycosyltransferases that contribute to the breakdown of carbohydrates commonly found in leaf litter. To that end, we also observed a difference in the full genomic potential to degrade various carbohydrates in leaf litter between populations (ANOVA; $p < 0.01$). However, other predicted genomic traits (i.e. minimum generation time and optimal growth temperature) were indistinguishable between populations, most likely due to the calculation incorporating full genome-wide codon usage biases (Supplementary Figure S2).

DISCUSSION

Our results suggest that both allopatric and sympatric processes are responsible for structuring populations of free-living soil bacteria across a regional climate gradient. This genetic resolution was possible by isolating a variety of *Curtobacterium* strains from the same habitat (leaf litter) across geographic locations (33). Within the most abundant ecotype, Subclade IB/C, we quantified gene flow among closely-related, co-occurring lineages to identify distinct genetic populations of *Curtobacterium* across geographic distances. An analysis of the flexible genome confirmed that these populations are structured by gene flow discontinuities and provided additional evidence for population-specific adaptation. Finally, the distributional patterns of the populations suggest that both isolation-by-distance and isolation-by-environment contribute to *Curtobacterium* population structure. Thus, both dispersal limitation and local environmental adaptation contribute to the divergence among closely-related soil bacteria as observed in macroorganisms (39).

Previously, studies of two soil bacteria, *Streptomyces* and *Bradyrhizobium*, found continental-scale patterns consistent with allopatric diversification over distantly-related strains (<90% ANI) (6,7,13). Further, clonal sympatric strains of the social bacterium *Myxococcus* were found to have barriers to recombination over cm distances in soil (40). By isolating strains within a single *Curtobacterium* ecological cluster at varying geographic scales, we could characterize the processes driving recent population divergence between both co-occurring strains and across regional spatial scales. As a comparison, we included two strains within this ecotype that were isolated from Boston, MA and found no recent recombination events connecting strains across continental scales (Fig. 2). Notably, along the regional climate gradient, we found that closely-related strains isolated from similar leaf litter communities were constrained in their geographic extent (mean geographic range of populations = 62.4 ± 100 km), suggesting that observed gene flow patterns is consistent with allopatric differentiation. However, we also observed multiple, genetically-distinct populations overlapping at three of the sites. Two of these populations were comprised of individuals from spatially distinct sites that remained connected by gene flow, suggesting isolation-by-distance is reduced at regional spatial scales. These results are contrary to previous work in fungal populations conducted at similar spatial scales; where fungal populations were highly structured by geography insomuch that genomic differences strongly reflected local site adaptations, a pattern consistent with strictly allopatric differentiation (30,31).

The presence of sympatric *Curtobacterium* populations can indicate the presence of an isolating mechanism to maintain the cohesiveness of co-occurring genetic lineages (1). Indeed, the rate of homologous recombination between bacteria can decrease exponentially with increasing sequence divergence (41). Alternatively, the presence of sympatric populations could signify that spatial barriers between the populations existed in the past but have since been removed without sufficient time for genetic homogenization. The flexible genome of *Curtobacterium* provides two lines of evidence for the former and, specifically, that the identified populations have remained genetically isolated due to ecological differentiation, as others have observed in bacterial populations (42). First, *Curtobacterium* populations shared more flexible genes within populations than between, suggesting that the populations represent cohesive, ecologically differentiated clusters (Fig. 3). Flexible genes are thought to contribute to differences in niche exploitation (43) and can contribute to small fitness differences among microhabitats (15). For example, in the marine bacterium *Vibrio*, sympatric populations encoded habitat-specific genes (44) between free-living and particle-associated populations (45). At a similar microscale, *Curtobacterium* populations may differentiate between leaf litter microhabitats caused by variability in resources such as metals and carbohydrate availability. Accordingly, we observed differences in carbohydrate degradation potential and observed population-specific genomic islands encoding genes related to physiological features.

The second line of evidence that sympatric populations are being maintained by ecological differences is that all individuals within populations shared highly conserved genomic backbones containing population-specific genes (Fig. 4). The population-specific genomic backbones consisted of both core genes exhibiting a strict monophyletic division and population-specific flexible genes indicating recent selective sweeps within a population. These patterns have been previously identified in marine bacterial populations of *Vibrio* (44) and *Prochlorococcus* (16) and the archaeon *Sulfolobus* (24), where population-specific genomic regions were linked to small fitness differences and niche exploitation contributing to the coexistence of sympatric populations. Similarly, increased homologous recombination among strains of *Curtobacterium* populations could enable the rapid exchange of niche-adaptive genes for differential microhabitat specialization on leaf litter. This observation is consistent with isolation-by-environment where gene exchange rates among similar environments is higher than within geographic locations (25). Thus, the populations along the regional climate gradient seem to represent genetically-isolated lineages that are ecologically diverged by their partitioning microhabitats (within a location).

CONCLUSIONS

A major gap in our understanding of microbial diversity is the mechanisms contributing to the origin and maintenance of microbial diversification. Collectively, our results suggest a model for the recent microevolution of a soil bacterium. Similar to soil fungal populations and macroorganisms, free-living soil bacterial populations are geographically restricted. At the same time, distinct *Curtobacterium* populations may have also diverged to specialize on different leaf litter microhabitats, causing a reduction in gene flow between populations. Thus, overlapping populations are maintained within the same location, while also being connected via dispersal to individuals in other locations. Our results demonstrate that soil bacterial populations, similarly to those in other environments, are delineated by barriers to recombination where the proliferation of advantageous genes can spread in a population-specific manner (23,24,41,44).

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 (32), including a robust genomic dataset consisting of 26 strains from a climate gradient in southern California (33). We included two additional strains within the same ecotype from outside Boston, MA to provide varying spatial scales. Protein-coding regions and gene annotations were derived from the NCBI prokaryotic genome annotation pipeline (46). Genomes were further screened for the presence of mobile genetic elements by identifying integrating and conjugative elements (ICEs) with the ICEberg database (47), prophage sequences using PhiSpy (48), insertion sequences (IS) with ISfinder (49), and CRISPR with CRISPRCasFinder (50).

Evolutionary History of the Core Genome. We aligned all genomes using progressiveMauve (51) 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 (52) 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 (37). 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 (53). 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 (35) in the R software environment (54).

Gene Flow and Recombination Networks. To differentiate between vertical transmission and recent recombination, we identified recent transfer events across all pairs of genomes using PopCOGenT (https://github.com/philarevalo/PopCOGenT) (27). Briefly, we used a null model of sequence divergence to calculate the expected length distribution of identical genomic regions between strain pairs. Recently exchanged genes would enrich this distribution by introducing identical genomic regions that are longer and more frequent than expected. The extent of this enrichment is our measurement of recent transfer. Strains that were too closely related (<0.035% ANI divergence) to accurately assess recombination transfers were collapsed into clonal complexes. Finally, strains that were connected to any other strain in the recombination network were considered to be a part of the same recombining population. To confirm the importance of recombination events in structuring populations, we inferred the relative effects of recombination and mutations rates of the core genome (see above) within each population using ClonalFrameML.

Population Genetic Analyses. To perform within population genetic analyses, we identified all orthologous protein-coding genes (orthologs) shared across all strains. Orthologs were initially predicted using ROARY (55) with a minimum sequence identity of 90% to ensure all possible orthologs were included across populations (Supplementary Figure 3A). The resulting 2193 orthologs shared across all strains were individually aligned with ClustalO v1.2.3 (56) 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 (52) 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 3B). 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 nucleotide diversity within populations using the PopGenome package (57) in R, as outlined in (58).

Predicted orthologs that were not shared across all strains represent the flexible genome (Supplementary Figure 3A). 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 populations 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 and were not found in any member outside of the population. To reduce this list even further, we identified population-specific orthologs that were localized in genomic regions (<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 (59)[equation 1].

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

ENC = effective number of codons given %GC (60)

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 (61) using HMMer (62). 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 (32).

DECLARATIONS

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

The authors declare that they have no competing interests.

AVAILABILITY of DATA and MATERIALS

Relevant data and code used can be found at https://github.com/alex-b-chase/curto-popgen. Biosample identification numbers are available in Supplementary Table S1. Additional datasets used and/or analyzed during the current study are available from the corresponding author on request.

FUNDING

This work was supported by an US Department of Education Graduate Assistance of National Need (GAANN) fellowship and the US Department of Energy, Office of Science Graduate Student Research (SCGSR) fellowship to ABC, and the US Department of Energy, Office of Science, Office of Biological and Environmental Research (DE-SC0016410) to JBHM.

AUTHOR CONTRIBUTIONS

ABC, ELB, UK, and JBHM designed the research project; ABC, PA, and UK analyzed data; and ABC, MFP, and JBHM wrote the manuscript with input from all authors.

ACKNOWLEDGMENTS

The authors thank Claudia Weihe, Chamee Moua, and Michaeline Albright for their assistance in strain isolation. We also 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.

REFERENCES

1. Mayr E. What evolution is. Science Masters Series; 2001.
2. Chase AB, Martiny JBH. The importance of resolving biogeographic patterns of microbial microdiversity. Microbiol Aust. 2018;39(1):5–8.
3. Shapiro BJ, Leducq J-B, Mallet J. What is speciation? PLoS Genet. 2016;12(3):e1005860.
4. Rocha EPC. Neutral theory, microbial practice: challenges in bacterial population genetics. Mol Biol Evol. 2018;35(6):1338–47.
5. Ward DM, Cohan FM, Bhaya D, Heidelberg JF, Kühl M, Grossman A. Genomics, environmental genomics and the issue of microbial species. Heredity (Edinb). 2008;100(2):207.
6. Andam CP, Doroghazi JR, Campbell AN, Kelly PJ, Choudoir MJ, Buckley DH. A latitudinal diversity gradient in terrestrial bacteria of the genus Streptomyces. MBio. 2016;7(2):e02200-15.
7. Choudoir MJ, Doroghazi JR, Buckley DH. Latitude delineates patterns of biogeography in terrestrial Streptomyces. Environ Microbiol. 2016;18(12):4931–45.
8. Whitaker RJ, Grogan DW, Taylor JW. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science. 2003;301(5635):976–8.
9. Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D, et al. Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. Environ Microbiol. 2008;10(1):147–61.
10. Wright S. Isolation by distance. Genetics. 1943;28(2):114.
11. Choudoir MJ, Buckley DH. Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of Streptomyces sister-taxa. ISME J. 2018;1.
12. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science (80-). 2006;311(5768):1737–40.
13. VanInsberghe D, Maas KR, Cardenas E, Strachan CR, Hallam SJ, Mohn WW. Non-symbiotic Bradyrhizobium ecotypes dominate North American forest soils. ISME J. 2015;9(11):2435.
14. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High-throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. bioRxiv. 2017;225342.
15. Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol [Internet]. 2014;12(4):263–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24590245
16. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344(6182):416–20.
17. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. PLoS Comput Biol. 2015;11(5):e1004095.
18. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al. Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 2015;32(6):1396–410.
19. Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. Convergence of Campylobacter species: implications for bacterial evolution. Science. 2008;320(5873):237–9.
20. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320(5879):1081–5.
21. Cohan FM. Bacterial species and speciation. Syst Biol. 2001;50(4):513–24.
22. Chase AB, Karaoz U, Brodie EL, Gomez-Lunar Z, Martiny AC, Martiny JBH. Microdiversity of an Abundant Terrestrial Bacterium Encompasses Extensive Variation in Ecologically Relevant Traits. MBio. 2017;8(6):e01809-17.
23. Whitaker RJ, Grogan DW, Taylor JW. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol Biol Evol. 2005;22(12):2354–61.
24. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, et al. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. 2012;10(2):e1001265.
25. Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014;23(23):5649–62.
26. Polz MF, Alm EJ, Hanage WP. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 2013;29(3):170–5.
27. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A reverse ecology approach based on a biological

definition of microbial populations. Cell. 2019;

28. Ranjard L, Richaume A. Quantitative and qualitative microscale distribution of bacteria in soil. Res Microbiol. 2001;152(8):707–16.

29. Nannipieri P, Ascher J, Ceccherini M, Landi L, Pietramellara G, Renella G. Microbial diversity and soil functions. Eur J Soil Sci. 2003;54(4):655–70.

30. Amend A, Garbelotto M, Fang Z, Keeley S. Isolation by landscape in populations of a prized edible mushroom Tricholoma matsutake. Conserv Genet. 2010;11(3):795–802.

31. Branco S, Gladieux P, Ellison CE, Kuo A, LaButti K, Lipzen A, et al. Genetic isolation between two recently diverged populations of a symbiotic fungus. Mol Ecol. 2015;24(11):2747–58.

32. Chase AB, Arevalo P, Polz MF, Berlemont R, Martiny JBH. Evidence for ecological flexibility in the cosmopolitan genus Curtobacterium. Front Microbiol [Internet]. 2016;7(November):1874. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2016.01874/full

33. Chase AB, Gomez-Lunar Z, Lopez AE, Li J, Allison SD, Martiny AC, et al. Emergence of soil bacterial ecotypes along a climate gradient. Environ Microbiol. 2018;20(11):4112–26.

34. Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. Microbe. 2014;9(3):111–8.

35. Frichot E, François O. LEA: an R package for landscape and ecological association studies. Methods Ecol Evol. 2015;6(8):925–9.

36. van Elsas JD, Bailey MJ. The ecology of transfer of mobile genetic elements. FEMS Microbiol Ecol. 2002;42(2):187–97.

37. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11(2):e1004041.

38. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. Isme J [Internet]. 2008 Oct 2;3:199. Available from: https://doi.org/10.1038/ismej.2008.93

39. Sexton JP, Hangartner SB, Hoffmann AA. Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution (N Y). 2014;68(1):1–15.

40. Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ, et al. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. ISME J. 2016;

41. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science (80-). 2007;315(5811):476–80.

42. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22(5):235–47.

43. Rodriguez-Valera F, Ussery DW. Is the pan-genome also a pan-selectome? F1000Research. 2012;1.

44. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, et al. Population genomics of early events in the ecological differentiation of bacteria. Science (80-). 2012;336(6077):48–51.

45. Yawata Y, Cordero OX, Menolascina F, Hehemann J-H, Polz MF, Stocker R. Competition–dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc Natl Acad Sci. 2014;111(15):5622–7.

46. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614–24.

47. Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, et al. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 2011;40(D1):D621–6.

48. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res. 2012;40(16):e126–e126.

49. Siguier P, Pérochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(suppl_1):D32–6.

50. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 2018;

51. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.

52. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies.

Bioinformatics. 2014;30(9):1312–3.

53. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014;genetics-113.

54. Pinheiro J, Bates D, DebRoy S, Sarkar D. R Development Core Team. 2010. nlme: linear and nonlinear mixed effects models. R package version 3.1-97. R Found Stat Comput Vienna. 2011;

55. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.

56. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol [Internet]. 2011 Oct 11 [cited 2016 Nov 13];7(1):539. Available from: http://msb.embopress.org/content/7/1/539.abstract

57. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 2014;31(7):1929–36.

58. Lemieux JE, Tran AD, Freimark L, Schaffner SF, Goethert H, Andersen KG, et al. A global map of genetic diversity in Babesia microti reveals strong population structure and identifies variants associated with clinical relapse. Nat Microbiol. 2016;1(7):16079.

59. Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta) genomics. PLoS Genet. 2010;6(1):e1000808.

60. Subramanian S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics. 2008;178(4):2429–32.

61. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85.

62. Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 2011;39(SUPPL. 2):29–37.

FIGURE LEGENDS

FIGURE 1. (A) Phylogeny of the *Curtobacterium* ecotype, subclade IB/C, from a core genome alignment. (B) Ancestral population structure estimated from admixture analysis. 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 2. Recombination network across all pairwise strains. Thicker edges represent increased recombination between strains. Nodes are colored by population designation and node size indicates number of clonal clusters (strains too closely-related to differentiate recombination). D = Desert, Sc = Scrubland, G/MMLR = Grassland, SS = Salton Sea, MCBA = Boston

FIGURE 3. Flexible gene content similarity between strains. Tree is derived from a consensus neighbor-joining analysis showing only nodes with ≥750 support. Strains are colored by population assignments identified from the recombination network (Fig. 2).

FIGURE 4. Highly structured genomic backbones across strains. (A) Population-specific genomic backbones within all individuals in populations 1 and 3. Population-specific genes (colored in blue) are consistently flanked by highly conserved regions (in white). Putative mobile elements are also designated in boxes along the chromosome. (B) Phylogenies of a subset of conserved genes (white arrows in panel A) flanking the population-specific regions colored by the strains in each respective population.

SUPPORTING INFORMATION

Supplementary Figure 1. Boxplots show nucleotide diversity (π) across all population-specific genes (present in all members within the population). Each point is a population-specific gene and is colored whether the gene is localized in the genomic region displayed in Fig. 4. The dashed line shows the genome-wide average (π_{MEAN}) of each strain in the population across all core genes within subclade IB/C.

Supplementary Figure 2. Distributions of predicted genomic traits in strains belonging to populations. 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 3. 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.

Supplementary Table 1. Genomic and geographic characteristics of strains.

Supplementary Table 2. Ratio of nucleotide substitutions from recombination to point mutations (r/m).