Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations

Sarahi L Garcia[a,b,*], Sarah L R Stevens[a,*], Benjamin Crary[c], Manuel Martinez-Garcia[d], Ramunas Stepanauskas[e], Tanja Woyke[f], Susannah G Tringe[f], Siv Andersson[g], Stefan Bertilsson[b], Rex R. Malmstrom[f], Katherine D McMahon[a,c]

a Department of Bacteriology, University of Wisconsin-Madison Madison, WI, USA
b Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden
c Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA
d Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain.
e Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA
f DOE Joint Genome Institute, Walnut Creek, CA, USA
g Department of Molecular Evolution, Uppsala University Uppsala, Sweden
* Equal contributors

Corresponding Author:
Katherine McMahon
1550 Linden Drive
Madison, WI 53711
trina.mcmahon@wisc.edu

Abstract

To understand the forces driving differentiation and diversification in wild bacterial populations, we must be able to delineate and track ecologically relevant units through space and time. Mapping metagenomic sequences to reference genomes derived from the same environment can reveal genetic heterogeneity within populations, and in some cases, be used to identify boundaries between genetically similar, but ecologically distinct, populations. Here we examine population structure within abundant and ubiquitous freshwater bacterial groups such as the acI Actinobacteria and LD12 Alphaproteobacteria (the freshwater sister clade to the marine SAR11) using 33 single cell genomes and a 5-year metagenomic time series. The single cell genomes grouped into 15 monophyletic clusters (termed “tribes”) that share at least 97.9% 16S rRNA identity. Distinct populations were identified within most tribes based on the patterns of metagenomic read recruitments to single-cell genomes representing these tribes. Genetically distinct populations within tribes of the acI actinobacterial lineage living in the same lake had different seasonal abundance patterns, suggesting these populations were also ecologically distinct. In contrast, sympatric LD12 populations were much less genetically differentiated and had similar temporal abundance patterns. This suggests that within one lake, some freshwater lineages harbor genetically discrete (but still closely related) and ecologically distinct populations, while other lineages are composed of less differentiated populations with overlapping niches. Our results point at an interplay of evolutionary and ecological forces acting on these communities that can be observed in real time.

Introduction

Bacteria represent a significant biomass component in almost all ecosystems and drive most biogeochemical cycles on Earth. Yet we know little about the population structure of bacteria in natural ecosystems and have yet to find and define the boundaries for ecological populations. Cohesive temporal dynamics and associations inferred from distribution patterns have been documented for many habitats and these observations are consistent with the notion of such populations as locally coexisting members of a species [1]. The most compelling cases are from collections of closely related isolates [1-3], but cultured species represent only a very small portion of the bacteria populating the Earth [4,5], and thus we still know little about the most abundant lineages. Therefore it is critical to study microorganisms in their natural environments [6], in order to test if and how their population structure differs from the established models based on isolates. The advent of culture-independent approaches, such as single-cell genomics and metagenomics, provides an opportunity for gaining new insights about genome-level diversity at the population level. These approaches sample entire communities directly in their environment, thereby bypassing the need to isolate and culture individual community members [7,8].

The delineation of ecologically differentiated lineages within complex microbial communities remains controversial because direct evidence for such differentiation is usually sparse [9]. Additionally, the appropriate level of phylogenetic resolution defining ecologically equivalent groups has not yet been established and likely varies across different groups [10]. Past explorations for defining such groups have used genome-wide average nucleotide identity (gANI) across shared regions of isolate genome sequences [11,12]. These studies have found that gANI greater than 94-96% unites past classical species definitions and separates known sequenced strains into consistent and distinct groups. Such genetically distinct populations have also been observed in microbial communities using metagenomics [7,13,14]. In

several large-scale metagenomic studies performed in aquatic ecosystems, the sampled microbial communities were found to contain collections of individuals sharing gANIs greater than 95%, as inferred by mapping metagenomic reads against reference genomes [13,15-19]. A closer inspection of coverage discontinuity further revealed that few reads typically map at 90-95% identity, enabling delineation of 'sequence-discrete' populations. That is, reads mapping with identities above the coverage discontinuity are defined as originating from a 'sequence discrete population' of genetically nearly identical cells that are distinct from other cells whose sequences map with identities below the coverage discontinuity. Metagenomic read recruitment can also be used to track spatial and temporal dynamics in the abundance and microscale diversity of genetically distinct populations[19]. For the remainder of the manuscript, we will use the terms 'population' and 'sequence-discrete population' interchangeably.

We used a combination of time-series metagenomics and single cell genomics to define genetic diversification within ubiquitous and abundant freshwater lineages such as acI and LD12. The term "tribe" was previously coined to delineate these groups using 16S rRNA gene sequences, where tribes are defined by monophyly and >97.9% within-clade 16S rRNA gene sequence identity [20,21]. We remind readers that a synoptic review of the diversity and phylogenetic relationships among recognized freshwater bacterial groups proposed a controlled hierarchical vocabulary within which "lineage" is roughly analogous to "family", "clade" roughly equates to "genus", and "tribe" to "species" [21]. We avoid the classical Linnaean taxonomy vocabulary because so many of these organisms cannot yet be obtained in axenic cultures and thus cannot be formally assigned to an Order, Family, Genus, or Species. Indeed, one main motivation for the present study is the challenge of delineating ecologically relevant taxonomic units given observed patterns of population structure within naturally assembling communities. This study includes thirty-three Single Amplified

Genomes (SAGs) representing fifteen phylogenetically coherent groups (i.e. freshwater "tribes").

The SAGs in this study originated from four lakes geographically isolated from one another and represent a rich source of reference genomes that can be used to recruit metagenomic reads in order to study population structure and dynamics through time in naturally assembled communities. In particular, the contrasting origin of these SAGs provide the opportunity to assess the differences in populations belonging to the same "tribe" while having evolved in different island-like habitats (i.e. lakes). Two of the lineages featured in the present study are the abundant and ubiquitous freshwater Actinobacteria acI and Alphaproteobactera alfV containing the freshwater SAR11 sister-clade, LD12. Members of these lineages are intriguing in their own right, as they represent groups of free-living ultramicrobacteria that dominate many freshwater ecosystems [22-28]. They differ markedly with respect to within-lineage diversity: LD12 is the sole tribe defined within the freshwater alfV lineage, while the acI lineage is comprised of 13 tribes [21]. The acI and LD12 have no axenic cultured representatives and share a large number of genomic and cellular traits. First, both lineages have genomes with GC content values lower than 40% and estimated sizes of about 1.5 Mb or less [29-31]. These genome characteristics are all the more striking since most cultivated species in the Alphaproteobacteria and Actinobacteria have GC-rich genomes up to 10 Mb in size. Second, both lineages have evolved by massive gene loss [30,32]. Third, the fraction of gained genes is only about 10% of the lost genes. Fourth, both groups of bacteria have small cell volumes[27,28]. However, acI and LD12 seem to employ different substrate niche specialization. While acI is thought to primarily use polyamines, oligopeptides and carbohydrates, LD12 specializes in carboxylic acids and lipids [29,33].

By combining genome information from twenty-one previously published [29,30] and twelve new SAGs from different freshwater lineages and an extensive five-year time series of

lake metagenomes (94 samples), we investigated the population structure of such ubiquitous freshwater bacteria for the first time. Our results confirm the existence of coherent sequence-discrete populations within these ubiquitous freshwater bacterial groups in natural communities and we could trace the abundance and gANI of these populations over monthly to seasonal time scales. Our work demonstrates the power of combining time-series metagenomics and single cell genomics for studying bacterial diversification and for describing ecologically meaningful population structure within the uncultured majority inhabiting natural ecosystems.

Results

The SAG collection represents multiple clades within cosmopolitan freshwater lineages

We analyzed 33 SAGs from four different freshwater lakes. Twenty-one of these SAGs were previously analyzed for their genomic features and phylogenetic relationships [29-31,34]. The 33 SAGs had total assembly sizes between 0.33 and 2.42 Mbp and were organized into 8 to 103 contigs with GC contents between 29.1% and 51.7% (**Table 1**). Estimated genome completeness, calculated using two different methods, ranged between 30% and 99%. Throughout the paper we will use mostly the shorter name version to facilitate reading, for example, M14 in place of AAA027-M14.

The 33 SAGs in the study represent fifteen different previously defined freshwater "tribes" that are each monophyletic and defined by >97.9% within-clade 16S rRNA gene sequence identity, measured across the nearly full-length 16S rRNA gene [20,21]. Freshwater microbial ecology researchers generally discuss and track these tribes as if they were coherent units that are ecologically distinct from one another. Ten tribes are represented by only one SAG each, while four tribes (LD12, acI-A1, acI-A7 and acI-B1) have more than one SAG representative in our dataset. To illustrate phylogenetic and taxonomic placement of the

LD12 and acI SAGs at finer scale resolution than previously achieved using partial 16S rRNA genes, we used the PhyloPhlAn pipeline [35] to generate a multi-gene tree (**Figure 1A and 1B**). The tree supported the 16S-based tribe designations but did not reveal a clear biogeographic pattern, in agreement with previous analyses, i.e. members of the same tribes were found in different lakes [30,32]. However, our SAG collection was not designed to explore biogeography and much deeper sampling of each population would be needed to address this question rigorously.

Genome-wide nucleotide identity is consistent with phylogeny

To further examine the genetic diversity within and among tribes, we determined the gANI using the set of four tribes that each contained more than one SAG representative. This general approach has been proposed as a way to compare genome pairs using a single metric that robustly reflects phylogenetic and taxonomic groupings obtained using other polyphasic methods [11,12]. We asked whether all genome pairs from the same tribe shared a consistent minimum gANI. Most SAGs shared gANI of at least 78% and alignment fractions greater than 40% with other members of the same tribe (**Figure 1C and Table S1**). All pairs from the same tribe that were also recovered from the same lake shared at least 84% gANI, but some pairs were much more similar (gANI approaching 99%). gANIs between pairs belonging to different tribes but still within the same lineage were markedly lower and typically below 74% (e.g. acI-A1 vs acI-B1) (**Figure 1C and Table S1**).

Although gANI is a useful univariate metric for comparing genome pairs, it masks the differences in sequence similarity of individual genes or genome regions that arise due to varying rates of divergence across loci. This variation can be visualized by plotting the frequency distribution of nucleotide identities calculated using a sliding window across the genome [11]. We asked whether different homologous genomic regions from two SAGs would have markedly different nucleotide identities even if they were from the same tribe. We used

the most complete SAGs from the acI-B1 and LD12 tribes as reference genomes and calculated nucleotide identity using a sliding window with other SAGs from the same respective tribe and visualized the results as a frequency distribution (**Figure 2**). The acI-B1 SAGs featuring the highest gANI (L06 and A23) were both from Lake Mendota and shared nucleotide identity consistently greater than 95% with a peak at 99-100%. The acI-B1 SAG P03 recovered from a lake in Germany had a frequency distribution with a peak more near 97% and a distinctly different shape. Other acI-B1 SAGs shared genomic regions with primarily 80-85% nucleotide identity. This was even true for J17, which was also collected from Lake Mendota and shared an average gANI of 79% with L06/A23 (**Table S1**), suggesting that cells belonging to the same tribe (acI-B1) and living in the same environment can have substantial genetic differences. The LD12 SAGs, which all belonged to the same tribe, also displayed three distinct patterns, with one peak near 85%, several near 91%, and two near 97%. Lake origin did not appear to explain these differences. That is, some LD12 cells from Lake Mendota were more similar to LD12 cells from Sparkling Lake than to other LD12 cells from Lake Mendota.

Diversity and structure of wild populations inferred using SAGs

The variety of patterns observed in **Figure 2** indicates substantial within-tribe variability even among cells recovered from the same lake. This made us wonder if tribes were composed of genetically and ecologically distinct populations coexisting in the same environment. SAGs can serve as relevant reference points to study the diversity of uncultured populations sampled using shotgun metagenomics by recruiting metagenomic reads and examining the extent of nucleotide identity for each aligned read [8]. The results can also be used to identify sequence-discrete populations whose boundaries are revealed by recruitment patterns and specifically the dramatic drop in coverage observed around 95% sequence

identity [7,18,19]. To examine the diversity and structure of wild freshwater bacterial populations, metagenomic reads from Lake Mendota, WI, USA, were mapped to the 33 SAGs, 19 of which were collected from this lake.

Each of the SAGs was first used to recruit reads from a single metagenomic dataset collected from Lake Mendota on 29 April 2009 (**Figure S1**). This time point was chosen because it was the sample collected closest to the date on which the single cells were collected (12 May 2009). Frequency distribution plots of the same data (**Figure 3**) revealed patterns that were similar to those obtained with SAG pairs (**Figure 2**). The five acI-SAGs from Lake Mendota (J17, L06, A23, M14 and I14) recruited more reads than the acI-SAGs from other lakes, with many reads recruiting at nucleotide identity greater than 97.5% (**Figure 3A**). All of the acI-SAGs also recruited many reads at 60 – 90% identity (**Figure 3A and D**), creating the characteristic bimodal distribution observed in previous work [7]. Based on these results, we hereafter consider reads sharing > 97.5% nucleotide identity as coming from the same, operationally defined *population* as the reference SAG. Thus, the acI lineage in Lake Mendota was composed of multiple sequence-discrete populations. Interestingly, the acI-B1 tribe in Lake Mendota, a subset of the acI lineage, appeared to be composed of at least two coexisting and genetically distinct populations, one represented by SAG J17 and the other by SAGs A23 and L06, consistent with the pairwise gANI observed using only the SAGs (**Figure 2**).

To determine if we recovered representative SAGs from all acI populations in Lake Mendota, we next performed recruitments competitively, allowing each read to only map to the SAG with the greatest % identity (**Figure S2**). As the patterns in **Figure 3** were generated by non-competitive mapping, some reads mapping with 100% similarity to one SAG might for example also have mapped with 60-90% similarity to SAGs from different sequence-discrete populations. After competitive mapping the resulting frequency distributions

changed and the fraction of reads mapping with 60-90% identity to each acI SAG dropped dramatically (**Figure S2**). However, a secondary peak around 80% identity still remained in most cases, and it is possible these reads originated from cells belonging to other acI populations lacking a representative SAG.

LD12 SAGs collected from Lake Mendota (C06, J10, L15, C07 and D10) also had a distinctive peak of recruited reads at >97.5% sequence identity (**Figure 3B**), although the overall shape of the recruitment patterns differed dramatically from those of the acI lineage. For example, LD12 SAGs had a secondary recruitment peak at ~92% identity whereas the acI SAGs had secondary peaks at ~75% with non-competitive mapping. This suggests the sequence-discrete populations within the LD12 tribe were more similar genetically than populations comprising the acI-B1 tribe. In fact, the populations were sufficiently similar that the hallmark coverage discontinuity below 97% similarity was not particularly pronounced (**Figure 3B**). Under competitive recruiting conditions, the LD12 recruitment distribution plots had remarkably different shapes (**Figure S2B and D**), as compared to the uncompetitive recruiting conditions (**Figure 3B**), and each SAG had only a single peak at >97.5% identity. This suggests the majority of LD12 cells in Lake Mendota belong to sequence-discrete populations represented by the SAGs in our collection.

All but one (I06) of the other freshwater SAGs in this study that were collected from Lake Mendota generated the distinctive read recruitment frequency peak above 97.5% identity (**Figure 3C**) that was observed for acI (**Figure 3A**). A negligible number of reads recruited to the SAGs collected from other lakes under the competitive recruiting conditions (data not shown). Since each of these SAGs represent just one tribe, it is not appropriate to infer any general conclusions for these populations or tribes, but we present them here to show the intriguing diversity of recruitment patterns. We finally underscore the need to more deeply sample individual population members using SAGs, to better capture and describe the

range of variation in population structure.

Are sequence-discrete populations within a tribe ecologically discrete too?

Results from a single metagenome sample suggested that individual tribes were composed of multiple genetically distinct populations that could be delineated and tracked using metagenomic read recruitment. Next we hypothesized that these populations might also be ecologically distinct and fill different realized niches. If so, we might expect these populations to display different temporal abundance patterns. We followed changes in population abundance through time by recruiting reads from a five-year metagenomic time-series applying a nucleotide identity cutoff of 97.5%. SAGs from the LD12 tribe recruited more reads than all of the acI SAGs summed together, on almost all sample dates (Figure 4A).

Using the relative number of reads recruited as a proxy for abundance, we found the J17 population, which belonged to the acI-B1 tribe, to be the most abundant acI population in almost every sample (**Figure 4B and 5A**). The abundance of the J17 population was poorly correlated over time with the other acI-B1 population represented by L06 and A23 (maximum Spearman rank correlation = 0.294), indicating each population had a different temporal abundance pattern. This suggests the two sequence-discrete populations comprising the acI-B1 tribe were also ecologically distinct. The different tribes of acI, which were more distantly related than the populations within the B1 tribe, also displayed different abundance patterns in Lake Mendota. For example, the acI-A1 I14 SAG population peaked in spring, but at markedly higher levels in 2009 and 2012 than in other years (**Figure 4B**). The acI-A6 I14 SAG population was consistently in low abundance compared to other tribes, but had small peaks in June and July.

In contrast to the acI-B1 tribe, the populations comprising the LD12 tribe had highly

similar abundance patterns. (**Figure 4C and S3**). The abundances of J10, L15, and C06 populations were strongly correlated (Spearman rank correlation = 0.997-0.999) and tended to peak both in Spring and Fall (Figure S3). The D10 population was the most abundant in the dataset but its abundance was not as strongly correlated to the other LD12 populations (Spearman rank correlation = 0.712-0.725). The C07 population was the least abundant but was also correlated to both the J10-L15-C06 populations and the D10 population (Spearman rank correlation = 0.861-0.873). Based on the similar temporal abundance patterns, the ecological differences among genetically discrete LD12 populations might be small, at least compared to the presumed major differences among acI-B1 populations that resulted in substantially different abundance patterns.

Does the genetic diversity of populations change over time?

We also examined the extent to which within-population diversity varied through time by quantifying changes in population-wide ANI, i.e. the average identity of all reads mapping with at least 97.5% identity (**Figure 5B**). For Mendota SAGs, the more abundant populations (such as LD12 and acI-B1 J17) generally had lower population-wide ANI variance through time compared to some less abundant populations (such as acSTL-A1-D23 and acI-A6-I14). For example, the SAG bacI-A1 G08 population had relatively high population-wide ANI in June 2009, around the time when the sample was collected for SAG library collection, but had markedly lower ANI on all other dates. One interesting exception to this observation was a significantly lower ANI for the relatively abundant acI-B1 L06-A23 population in 2012, as compared to 2007-2011 (Mann-Whitney U test p=1.4e-06). However, we note that for those populations recruiting a very low number of reads, it is possible that the sampling is not deep enough to reflect the true ANI value of the populations thus leading to a higher observed variance.

Discussion

Comparative genomics can reveal the diversity and structure of bacterial populations. This approach is particularly powerful when applied using single cells recovered from environmental samples (SAGs) and shotgun metagenomes from the same or similar ecosystems. Here we used a combination of 33 SAGs and 94 metagenomes collected over five years to ask the following questions: 1) How well does our SAG collection represent the diversity found in natural communities? 2) Do common freshwater bacterial groups have similar population structure? and 3) How stable is population abundance and diversity through time? We used the answers to these questions to gain insight into the population structure and ecology of the cosmopolitan and abundant freshwater bacteria, LD12 (Alphaproteobacteria) and acI (Actinobacteria).

Pairwise genome-wide ANI has been proposed as a useful metric for determining if two genomes belong to the same species [11,12]. This kind of analysis has been used to illustrate genetic differences between classically defined species in pure cultures. The analogous approach of recruiting metagenomic reads from wild populations has been used to gather evidence for the existence of sequence-discrete populations (which may function as cohesive species-like groups) [7,18]. We found that sequence-discrete populations could be delineated in the Lake Mendota metagenome using our 33 SAGs as references, as has previously been demonstrated in other lakes using genomes assembled from metagenomes [7,19]. We interpret the occurrence of these populations in the context of previously defined phylogenetically coherent and ostensibly ecologically distinct “tribes” composed of cells with >97.9% 16S rRNA identity [36]. We conclude that the canonical freshwater tribes can contain multiple sequence-discrete populations. The converse is, of course, not true: sequence-discrete populations can never represent multiple tribes.

Pair-wise gANI analysis of SAGs and metagenomic read recruitment indicated that cells belonging to the same tribe but inhabiting different lakes were usually genetically distinct. For example, SAGs collected from other lakes generally recruited very few reads from Lake Mendota at ANI >97.5% (**Figure 5**) while many recruited a substantial number of reads in the 89-92% range (**Figure 3**). However there were two prominent exceptions: LD12 N17 and L09, both of which are from Sparkling Lake. N17 and L09 share 97% gANI with Mendota SAG D10, which is substantially higher than the average (88%) and median (90%) within-tribe gANI (Table S1). These SAGs also recruited roughly the same number of reads with >97.5% identity as did the LD12 SAGs from Lake Mendota, though around 20% of the base pairs in the genomes did not recruit any reads (17% for L09 and 23% for N17). This implies that some gene content was present in the Sparkling Lake populations but missing in Lake Mendota. However, 10% of the base pairs in the D10 genome also did not recruit any reads, even though it was from Lake Mendota. This rare genome content could represent flexible or low frequency genes in the population, or contamination in the SAG preparation. We examined the phylogenetic distribution of low-coverage contigs and did not discern any evidence of contamination.

In Lake Mendota, acI cells are organized into genetically discrete populations, but the forces creating this organization remain a mystery. The consistent lack of coverage around 90-97% identity in recruitment plots indicates Lake Mendota lacks acI genotypes sharing this degree of sequence similarity with our SAGs, or at least that these putative genotypes were consistently at much lower abundances than their close relatives over the five years surveyed. This raises the questions “how do sequence discrete populations persist?” and “why don’t we detect a continuum of genotypes across the range of 90-97% identity?”. The P03 SAG from Stechlin Lake shares gANI of 96% with acI-B1 SAGs from Mendota, indicating that genotypes within this locally excluded sequence space do exist, at least as long as they are

from different environments. We infer the persistence of the coverage discontinuity between populations to be less a factor of dispersal limitation and more likely the result of competitive exclusion and barriers to recombination within Mendota populations. Additional SAG and metagenomic studies are necessary to determine if the same forces maintain the coverage discontinuities among sequence discrete populations observed in other phylogenetic groups and in different environments.

We know that both acI tribes and LD12 vary in abundance over seasonal and annual time-scales, based on previous work using 16S rRNA gene sequencing, quantitative PCR, and FISH [27,28,37,38]. Here we used our SAGs to track such populations at monthly intervals over five years. The results confirmed prior work that showed acI tribes and LD12 are among the most abundant non-cyanobacterial groups in Lake Mendota [21,39] but also revealed dynamics at unprecedentedly high phylogenetic resolution. Based on our extensive comparison of how SAGs recruited relative to one another, we are confident that our metagenomic recruitment filters allowed us to delineate discrete populations that would not be possible to resolve using more traditional and widely used methods (e.g. 16S rRNA gene sequencing or FISH). Specifically, metagenomic recruitments to LD12 SAGs revealed strikingly different patterns compared to the acI lineage, suggesting fundamental differences in evolutionary history and/or lifestyles among such abundant and ubiquitous freshwater bacteria. We discovered that LD12 populations were not as strongly genetically separated as acI populations; pair-wise gANIs between SAGs were higher and recruitment plots showed secondary peaks between 90-95% identity (**Figure 3B**), the same range where coverage of acI SAGs was at a minimum (**Figure 3A**). Under a competitive recruitment analysis, wherein each read is counted only once and attributed to the best match SAG, the secondary peaks disappear (**Figure S2**), indicating the LD12 SAGs represent highly similar, but still genetically discrete, populations. Temporal abundance patterns of these LD12 populations were strongly

correlated over five years, whereas acI populations showed much lower correlation within tribes. This suggests that the acI-B1 populations are ecologically distinct (i.e. occupying temporally discrete niches) while LD12 populations are less differentiated genetically, and possibly ecologically neutral, leading to co-occurrence and synchronization of temporal abundance patterns. LD12 is a particularly fascinating group because it is also a subclade of the broader SAR11 clade, with hypothesized ancient transition from marine to freshwater [40] followed by specialization through gene flux and mutation, with comparatively low recombination rates [30]. Over time, low recombination rates should lead to large genetic divergence among coexisting populations. Thus, we propose that LD12 populations are simply at earlier stages of differentiation as compared to acI populations while we cannot exclude that something fundamental about their lifestyle is "holding" the populations together genetically and ecologically. In any case, the lack of coherence among acI-B1 populations challenges our concept of tribes as ecologically coherent units and will force freshwater microbial ecologists to re-examine conventions for tracking these units through space and time.

Our observations stand in contrast to those reported for the marine species *Vibrio cyclitrophicus* [41]. Shapiro and colleagues examined strains inhabiting large-size (L) and small-size (S) particles and considered their particle-association to represent ecological differentiation. These L and S strains had an average across-group gANI of 99.0% (**Table S6**) and would not appear genetically distinct using the metagenomic read recruitment method applied in our study. That is, the *V. cyclitrophicus* L and S strains appear to be less genetically differentiated than LD12 but possibly more ecologically differentiated. Thus, our work provides further evidence for conceptual models of bacterial evolution in which different lineages can diversify in different ways, and no single mode will explain all extant diversity.

The metagenomic recruitments allowed us to also examine the extent to which

diversity varied within and among populations as well as how diversity changed over time. We calculated the population-wide ANI for reads that recruited only above 97.5% and found the resulting value was remarkably stable through time for most of the abundant populations (**Figure 5B**). This was particularly true for the LD12 populations. However, one striking contrast was the acI-B1 population represented by L06/A23, which had consistent population-wide ANI of 99.3% during 2008-2011 but 99.0% during 2012 (Mann-Whitney U test p=1.4e-06). This change suggests a substantial shift in the relative abundance of genotypes comprising the population. Similar shifts were observed previously in sequence-discrete populations inhabiting Trout Bog Lake, indicating this could be a common phenomenon among freshwater clades [19]. Unlike the genome-wide selective sweep observed in one *Chlorobium* population from Trout Bog Lake, the distribution of single nucleotide polymorphisms within the L06/A23 population before and after 2012 exhibited no clear pattern of gene- or genome-wide sweep (data not shown). It is difficult or impossible to separate genotypes within sequence-discrete populations using short-read shotgun sequencing, so further work using long-read technologies will be needed to link SNPs in populations to individual genomes. This kind of approach will likely be required to tease apart the paths leading to diversification within and among populations.

Methods

Single amplified genomes (SAGs)

Water samples (1-ml) were collected from the upper 0.5m to 1m of each of four lakes (Mendota, Sparkling, Damariscotta, Stechlin) and cryopreserved, as previously described [31,42]. These lakes were originally selected because they represent different freshwater trophic status (eutrophic, oligotrophic, mesoeutrophic, and oligotrophic, respectively) and geographic regions (Wisconsin and Maine, USA, and Germany). Bacterial single amplified genomes

(SAGs) were generated by fluorescence-activated cell sorting (FACS) and multiple displacement amplification (MDA), and identified by PCR-sequencing of their 16S rRNA genes at the Bigelow Laboratory Single Cell Genomics Center (SCGC; http://scgc.bigelow.org). Thirty-two SAGs from lakes Mendota, Sparkling and Damariscotta were selected for sequencing based on the previously sequenced 16S rRNA gene as well as the kinetics of the MDA reactions [42]. The one SAG from Lake Stechlin was selected from a separate library because its 16S rRNA gene was 100% identical to an acI-B1 SAG previously analyzed (AAA027-L06) [31]. In the present study we analyze 21 previously published and 12 new SAGs. All 33 SAGs were analyzed (**Table 1**) after genome sequencing, assembly, contamination removal and annotation as previously described (Ghylin et al. 2014). Estimation of completeness was done using CheckM [43] and the gene markers from a recent study examining a large collection of draft environmental genomes [44].

Tree construction, Average Amino acid and Average Nucleotide Identity (AAI, ANI)

A phylogenomic analysis was conducted using PhyloPhlAn [35]. ANI was calculated by using the method described in [11] with fragment size of 1000, minimum alignment length of 700 bp, percent identity of 70, and e-value of 0.001. AAI was calculated by averaging the identity of the reciprocal best hits from the BLASTP searches of the predicted proteins for each pair of genomes. 16S rRNA gene similarity for each pair was calculated using the overlapping region in an alignment created using a multiple alignment (default options) in Geneious Version R6 [45].

SAG-to-SAG recruitments

SAG pairs from the same tribe were used to examine the frequency distribution of nucleotide identities across homologous regions of the two genomes. In order to create a sliding window for comparison, the contigs of all SAGs were shredded into 301bp fragments

with 150 bp overlap. Two SAGs were selected as reference genomes: L06 as the most complete from the tribe acI-B1 and C06 as the most complete LD12. The contigs of each of the two selected SAGs were used as a reference for recruiting from the shredded SAGs using Blast 2.2.28 [46]. Ribosomal RNAs were masked from the SAGs prior to performing blast.

Five-year time series metagenome data: sampling, sequencing and recruitments

Samples were collected from Lake Mendota over the course of five years, as previously described [47,48]. Lake Mendota, Madison, Wisconsin, (N 43°06, W 89°24) is one of the most well-studied lakes in the world, and is a Long Term Ecological Research site affiliated with the Center for Limnology at the University of Wisconsin Madison [49]. It is dimictic and eutrophic with an average depth of 12.8 m, maximum depth of 25.3 m, and total surface area of 3938 ha. Depth integrated water samples were collected from 0 to 12 m of the epilimnion (upper mixed layer) at 94 different time points during ice-free periods from summer 2008 to summer 2012, and filtered onto 0.2 μm pore-size polyethersulfone filters (Supor, Pall) prior to storage at -80°C. DNA was later purified from these filters using the FastDNA kit (MP Biomedicals). DNA sequencing was performed at the Department of Energy Joint Genome Institute using standard protocols (Walnut Creek, CA, USA). DNA from the 94 samples was used to generate libraries that were sequenced on the Illumina HiSeq 2000 platform. Paired-end sequences of 2 X 150bp were generated for all libraries. Adapter sequences, low quality reads (i.e. ≥80% of bases had quality scores <20), and reads dominated by short repeats of ≥3 bp were removed. The remaining high quality reads were merged with the Fast Length Adjustment of Short Reads [50] with a mismatch value of ≤0.25 and a minimum of ten overlapping bases from paired sequences, resulting in merged read lengths of 150 to 290 bp (**Table S3**). Metagenomes were pooled by month to reduce the time-series data to 30 observations and increase coverage.

All contigs from each of the 33 SAGs were used as a reference to recruit reads from the Mendota metagenomes using blastn. Metagenome reads that recruited to the SAGs were filtered and only alignments 200bp or longer were considered. An additional filter requiring an alignment percent identity of at least 97.5% was applied when analyzing the metagenome time series. Ribosomal RNAs were masked from the SAGs prior to performing the recruitments.

Statistics, Visualization, Reproducible Methods

Datasets were analyzed and results were visualized using custom scripts written in R [51] and python. Pipeline and scripts for analysis can be found at https://github.com/sstevens2/blast2ani.

References

1 Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in microbiology*, (2014).

2 Luo, C. W. *et al.* Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7200-7205, (2011).

3 Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *Bmc Biol* **3**, (2005).

4 Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143-169, (1995).

5 Kaeberlein, T., Lewis, K. & Epstein, S. S. Isolating "Uncultivable" Microorganisms in Pure Culture in a Simulated Natural Environment. *Science* **296**, 1127-1129, (2002).

6 Little, A. E. F., Robinson, C. J., Peterson, S. B., Raffa, K. E. & Handelsman, J. Rules of Engagement: Interspecies Interactions that Regulate Microbial Communities. *Annual Review of Microbiology* **62**, 375-401, (2008).

7 Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environmental microbiology* **14**, 347-355, (2012).

8 Stepanauskas, R. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**, 613-620, (2012).

9 Hunt, D. E. *et al.* Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081-1085, (2008).

10 Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**, 133-146, (2015).

11 Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2567-2572, (2005).

12 Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Research* **43**, 6761-6771, (2015).
13 Oh, S. *et al.* Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Applied and Environmental Microbiology* **77**, 6000-6011, (2011).
14 Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344**, 416-420, (2014).
15 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43, (2004).
16 Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304**, 66-74, (2004).
17 Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77, (2007).
18 Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**, 1052-1065, (2008).
19 Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*, (2016).
20 Newton, R. J., Jones, S. E., Helmus, M. R. & McMahon, K. D. Phylogenetic Ecology of the Freshwater Actinobacteria acI Lineage. *Appl. Environ. Microbiol.* **73**, 7169-7176, (2007).
21 Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14-49, (2011).
22 Rösel, S., Allgaier, M. & Grossart, H.-P. Long-Term Characterization of Free-Living and Particle-Associated Bacterial Communities in Lake Tiefwaren Reveals Distinct Seasonal Patterns. *Microbial Ecology* **64**, 571-583, (2012).
23 Salcher, M. M., Pernthaler, J. & Posch, T. Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limno. Oceanography* **55**, 846-856; 846, (2010).
24 Warnecke, F., Sommaruga, R., Sekar, R., Hofer, J. S. & Pernthaler, J. Abundances, Identity, and Growth State of Actinobacteria in Mountain Lakes of Different UV Transparency. *Appl. Environ. Microbiol.* **71**, 5551-5559, (2005).
25 Zwart, G., Crump, B. C., Agterveld, M. P. K.-v., Hagen, F. & Han, S.-K. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology* **28**, 141-155, (2002).
26 Glöckner, F. O. *et al.* Comparative 16S rRNA Analysis of Lake Bacterioplankton Reveals Globally Distributed Phylogenetic Clusters Including an Abundant Group of Actinobacteria. *Applied and Environmental Microbiology* **66**, 5053-5065, (2000).
27 Heinrich, F., Eiler, A. & Bertilsson, S. Seasonality and environmental control of freshwater SAR11 (LD12) in a temperate lake (Lake Erken, Sweden). *Aquatic Microbial Ecology* **70**, 33-44, (2013).
28 Salcher, M. M., Pernthaler, J. & Posch, T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *The ISME journal* **5**, 1242-1252, (2011).
29 Ghylin, T. W. *et al.* Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *ISME J* **8**, 2503-2516, (2014).
30 Zaremba-Niedzwiedzka, K. *et al.* Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biology* **14**, (2013).
31 Garcia, S. L. *et al.* Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J* **7**, 137-147, (2013).
32 Zhao, W., S.L., G., Ohlsson, R., McMahon, K. D. & Anderson, S. Single cell and metagenome data reveals novel gene acquisition and diversification of freshwater Actinobacteria along the salinity gradient of the Baltic Sea., (2016).
33 Salcher, M. M., Posch, T. & Pernthaler, J. In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake. *ISME J* **7**, 896-907, (2013).
34 Eiler, A. *et al.* Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *ISME J*, (2016).
35 Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications* **4**, 2304, (2013).
36 Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**, 784-798, (2010).

37 Eiler, A., Heinrich, F. & Bertilsson, S. Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J* **6**, 330-342, (2012).

38 Allgaier, M. & Grossart, H.-P. Diversity and Seasonal Dynamics of Actinobacteria Populations in Four Lakes in Northeastern Germany. *Appl. Environ. Microbiol.* **72**, 3489-3497, (2006).

39 Newton, R. J. & McMahon, K. D. Seasonal differences in bacterial community composition following nutrient additions in a eutrophic lake. *Environmental Microbiology* **13**, 887-899, (2011).

40 Logares, R., Brate, J., Heinrich, F., Shalchian-Tabrizi, K. & Bertilsson, S. Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Mol Biol Evol* **27**, 347-357, (2010).

41 Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51, (2012).

42 Martinez-Garcia, M. *et al.* High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**, 113-123, (2011).

43 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, (2015).

44 Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437, (2013).

45 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, (2012).

46 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, (2009).

47 Shade, A. *et al.* Interannual Dynamics and Phenology of Bacterial Communities in a Eutrophic Lake. *Limnology and Oceanography* **52**, 487-494, (2007).

48 Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L. & McMahon, K. D. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *Isme J* **7**, 680-684, (2013).

49 Carpenter, S. R. *et al. The ongoing experiment: Restoration of lake mendota and its watershed". Long term dynamics of lakes in the landscape.* . (Oxford Press, 2006).

50 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963, (2011).

51 R Core Team. *R: A language and environment for statistical computing.*, <http://www.R-project.org/.> (2014).

Acknowledgements

We thank Dr. Todd Miller and Sara Yeo for collecting the original water samples used to retrieve single cells from Lake Mendota and Sparkling Lake. We thank the Joint Genome Institute for supporting this work through the Community Science Program, performing the bioinformatics, and providing technical support. We thank Moritz Buck for informatics and statistical support. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. KDM acknowledges funding from the United States National Science Foundation (NSF) Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-0822700), an INSPIRE award (DEB-1344254), and the Swedish Wenner-Gren Foundation. RS acknowledges funding from NSF (DEB-0841933, EF-0633142 and OCE-821374). SB acknowledges funding from the Swedish Research Council. Sarahi Garcia thanks and acknowledges the JSMC for funding. MMG acknowledges funding from Ministry of Economy and Competitiveness (CGL2013-405064-R and SAF2013-49267-EXP)

Author contribution

SLG, SLRS, RM, SB and KDM conceived the research. RM, MMG, TW and SGT conducted experiments and generated the data. SLG, SLRS and KDM analyzed the data. SLG, SLRS and BC prepared the figures. SLG, SLRS, RM and KDM wrote the manuscript. All authors participated in revision of the manuscript.

Additional Information

The raw shotgun metagenome reads are publicly available in the JGI portal and the assembly is available in IMG/MER under the ER submission ID XXXXX. The access

number for each SAG and metagenome can be found in **Table 1** and **Table S3**.

Conflict of interest Statement: The authors declare no conflict of interest.

Figure and table legends

Figure 1. A. Phylogenetic tree of acI SAGs based on conserved single copy genes selected by PhyloPhlAn. Amino acid sequences from 400 genes were aligned. B. Phylogenetic tree of LD12 SAGs based on conserved single copy genes selected by PhyloPhlAn, representing 400 genes. Prior work provided evidence for finer-scale groups within the LD12 tribe including the A group (N17, L09, and D10), the B group (C06, J10, and L15) and the C group (P20 and M09) [30]. C. Genome-wide nucleotide identity (gANI) versus 16S rRNA gene identity for pairs of SAGs. Alignment fractions for homologous genomic regions and 16S genes are given in Table S1. Shapes indicate the lake the tribe is from, if same, otherwise different lake is indicated. Colors indicate the tribe a pair is from, if same, otherwise different tribe is indicated.

Figure 2. Nucleotide identity density plots for SAG versus SAG genome-wide comparison using a sliding window. Results are shown for two reference SAGs representing the most complete genomes from the most thoroughly sampled tribes. All SAG pairs were from the same tribe. Nucleotide identity was calculated with blastn using 301 bp fragments that overlapped by 150 bp. A. acI-B1 SAGs and other selected acI SAGs vs L06. Note that the purple line (D18) is hidden underneath the orange (I18) and red (J17) lines. B. selected LD12 SAGs vs C06. Note the dark blue line (L09) is hidden under the light green (N17) line.

Figure 3. Mapping metagenomic reads from Lake Mendota to SAGs. The x-axis represents nucleotide identity of the recruited reads. The metagenome sample was collected from Lake Mendota on 29 April 2009. Reads were only counted if they aligned over a minimum of 200 bp. Recruitments were not competitive, meaning that each read could recruit to multiple

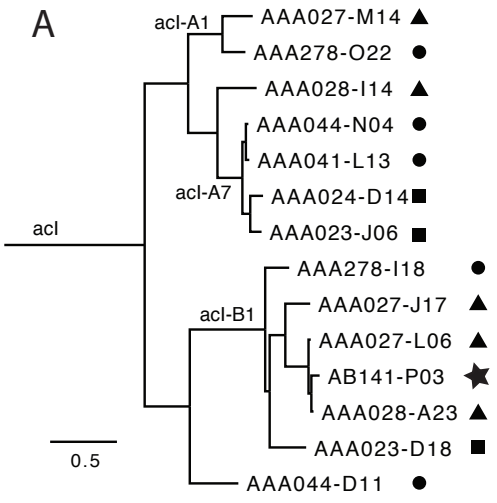
SAGs. Analogous competitive recruitments that required each read to recruit to only one SAG are presented in Figure S2. Each panel represents a different sub-set of the SAGs: A. acI from Mendota, B. acI not from Mendota, C. LD12 from Mendota, D. LD12 not from Mendota, E. other freshwater groups from Mendota

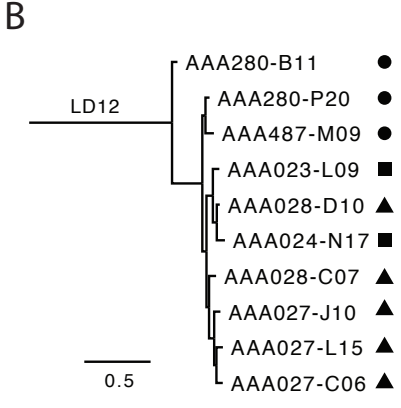
Figure 4. Sequence-discrete population abundance in Lake Mendota, as measured by the relative number of reads recruited to each SAG using blastn. All SAGs and samples are from Lake Mendota. Timepoints are pooled by month. Filtering criteria: ≥97.5% ANI and ≥200 bp alignment length. Blastn was done competitively, only counting the read for the best hit genome with the exception of when it hit equally well, then counted for all best hit genomes. Colors for each SAG are the same as in Figures 2 and 3. A. Sum of reads recruited for each lineage. B. Sum of reads recruited for each acI-SAG. C. Sum of reads recruited for each LD12 SAG.

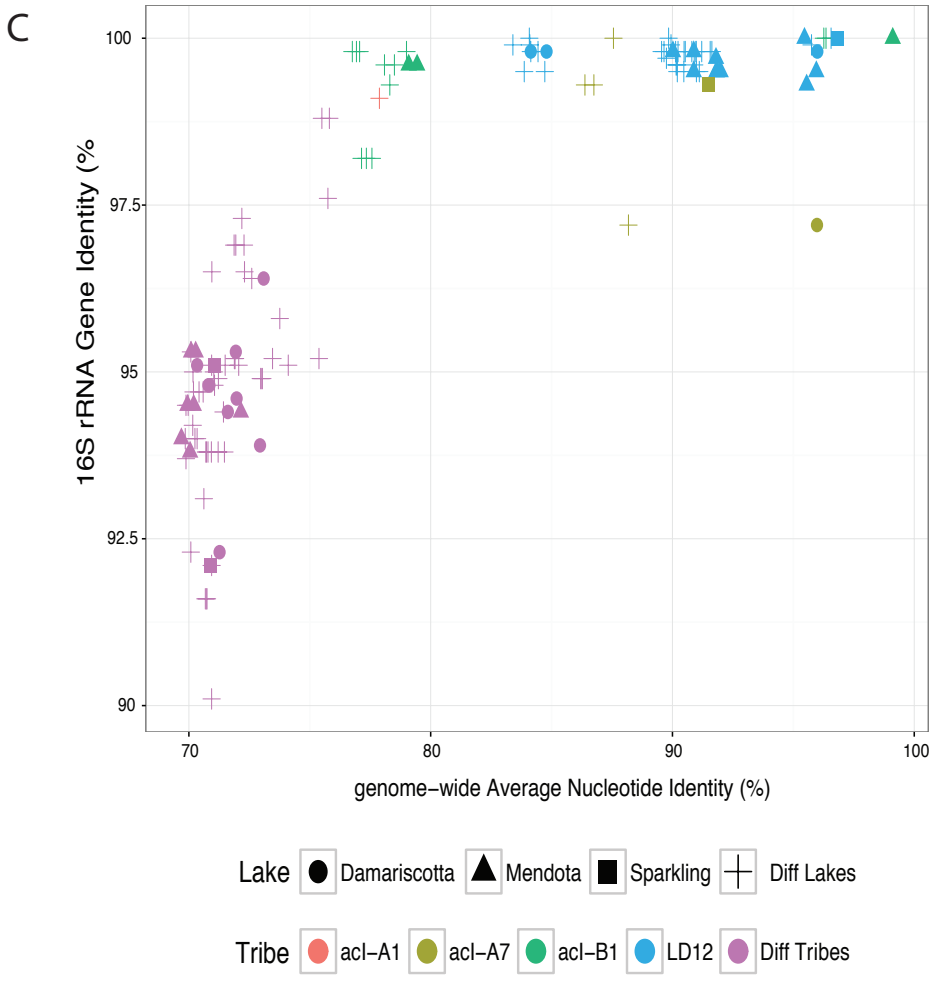
Figure 5. A. Metagenomic read recruitment using the SAGs from Lake Mendota. SAGs are in rows with bubbles representing all metagenomes from a particular month recruited against SAG. Filtering criteria are the same as those noted in Figure 4. Color scale indicates the ANI of the recruited metagenome reads. Bubble size represents the average coverage per base in the reference SAG divided by the size of the metagenome, multiplied by the average size of all metagenomes (1.34 Gigabases). Note that the resulting values do not represent a true measure of absolute abundance, but allow for quantitative comparison of month-to-month variation in population-level abundance. Recruitments were not performed competitively, meaning that each read could be recruited by multiple SAGs. B. Variation in ANI for each SAG, across all 30 metagenomes from throughout the five years. The data underlying these plots can be found in Table S5.

Tables

Table 1. Metadata for the 33 SAGs.


A
acl-A1
AAA027-M14
AAA278-O22
AAA028-I14
AAA044-N04
AAA041-L13
acl-A7
AAA024-D14
AAA023-J06
acl
AAA278-I18
acl-B1
AAA027-J17
AAA027-L06
AB141-P03
AAA028-A23
AAA023-D18
0.5
AAA044-D11


B
AAA280-B11
LD12
AAA280-P20
AAA487-M09
AAA023-L09
AAA028-D10
AAA024-N17
AAA028-C07
AAA027-J10
AAA027-L15
0.5
AAA027-C06


C
16S rRNA Gene Identity (%
100
97.5
95
92.5
90
70
80
90
100
genome−wide Average Nucleotide Identity (%)
Lake
Damariscotta
Mendota
Sparkling
Diff Lakes
Tribe
acl−A1
acl−A7
acl−B1
LD12
Diff Tribes

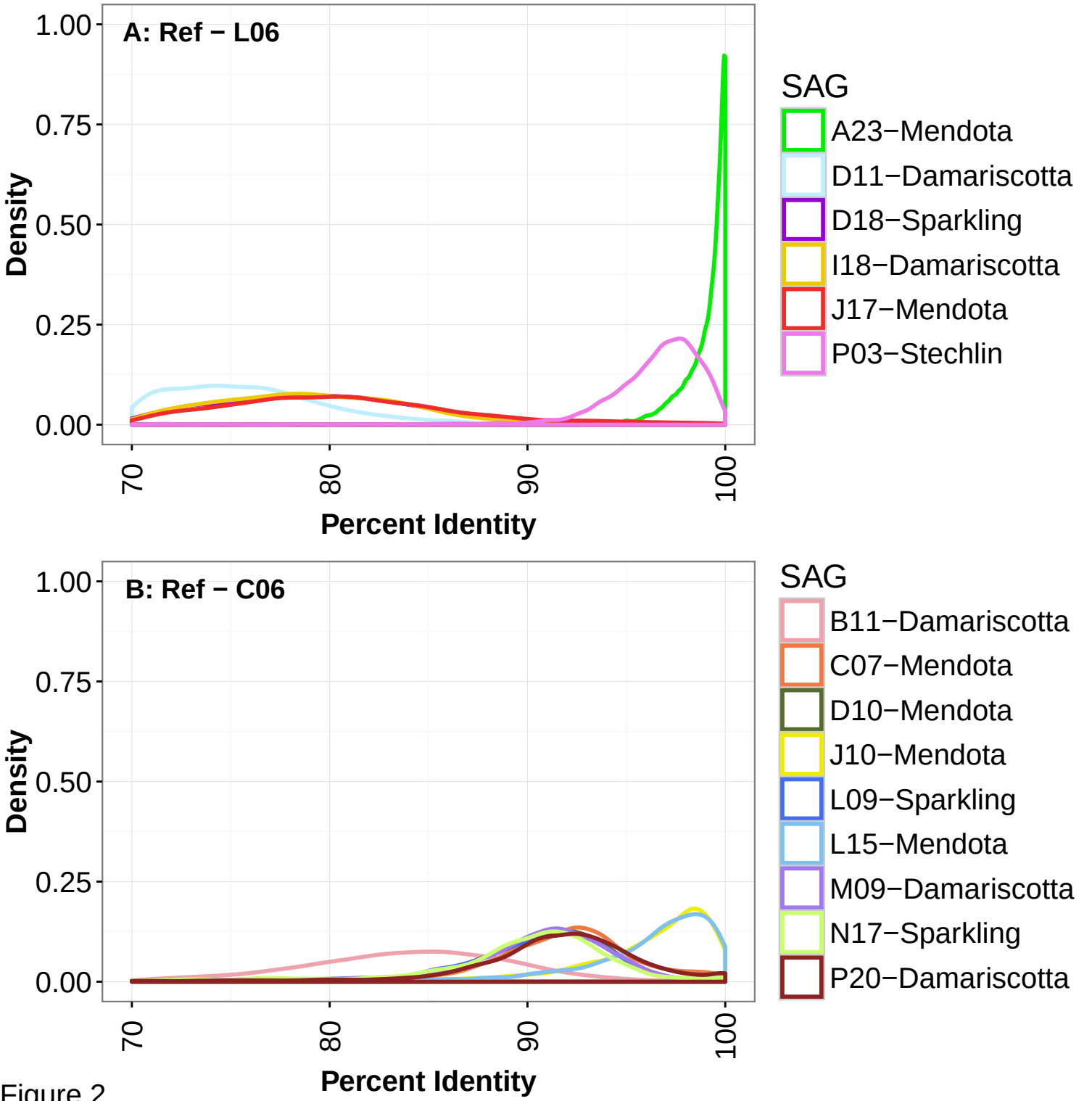
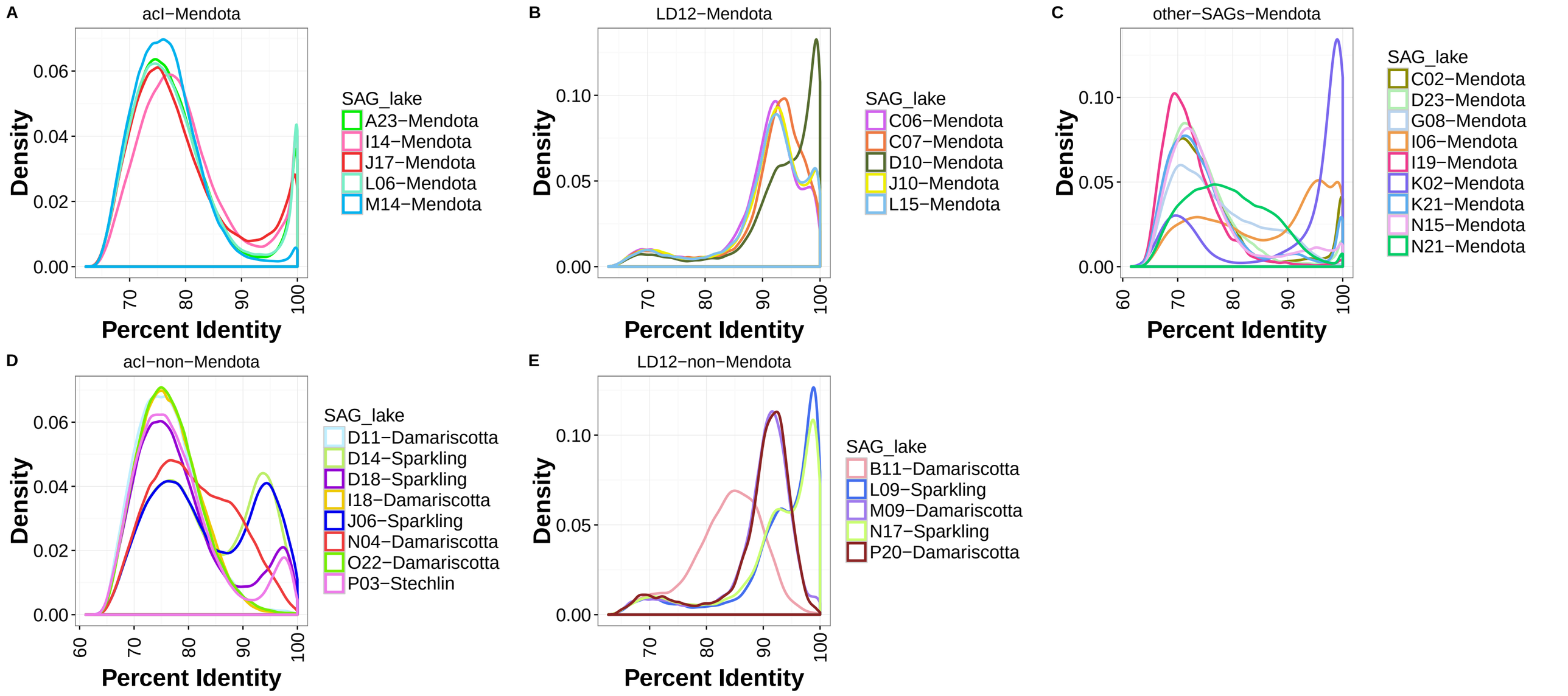

A: Ref − L06
Density
1.00
0.75
0.50
0.25
0.00
70
80
90
100
Percent Identity
SAG
A23−Mendota
D11−Damariscotta
D18−Sparkling
I18−Damariscotta
J17−Mendota
P03−Stechlin
B: Ref − C06
Density
1.00
0.75
0.50
0.25
0.00
70
80
90
100
Percent Identity
SAG
B11−Damariscotta
C07−Mendota
D10−Mendota
J10−Mendota
L09−Sparkling
L15−Mendota
M09−Damariscotta
N17−Sparkling
P20−Damariscotta

Figure 2


A
acI−Mendota
Density
0.00
0.02
0.04
0.06
70
80
90
100
Percent Identity
SAG_lake
A23−Mendota
I14−Mendota
J17−Mendota
L06−Mendota
M14−Mendota
B
LD12−Mendota
Density
0.00
0.05
0.10
70
80
90
100
Percent Identity
SAG_lake
C06−Mendota
C07−Mendota
D10−Mendota
J10−Mendota
L15−Mendota
C
other−SAGs−Mendota
Density
0.00
0.05
0.10
60
70
80
90
100
Percent Identity
SAG_lake
C02−Mendota
D23−Mendota
G08−Mendota
I06−Mendota
I19−Mendota
K02−Mendota
K21−Mendota
N15−Mendota
N21−Mendota
D
acI−non−Mendota
Density
0.00
0.02
0.04
0.06
60
70
80
90
100
Percent Identity
SAG_lake
D11−Damariscotta
D14−Sparkling
D18−Sparkling
I18−Damariscotta
J06−Sparkling
N04−Damariscotta
O22−Damariscotta
P03−Stechlin
E
LD12−non−Mendota
Density
0.00
0.05
0.10
70
80
90
100
Percent Identity
SAG_lake
B11−Damariscotta
L09−Sparkling
M09−Damariscotta
N17−Sparkling
P20−Damariscotta

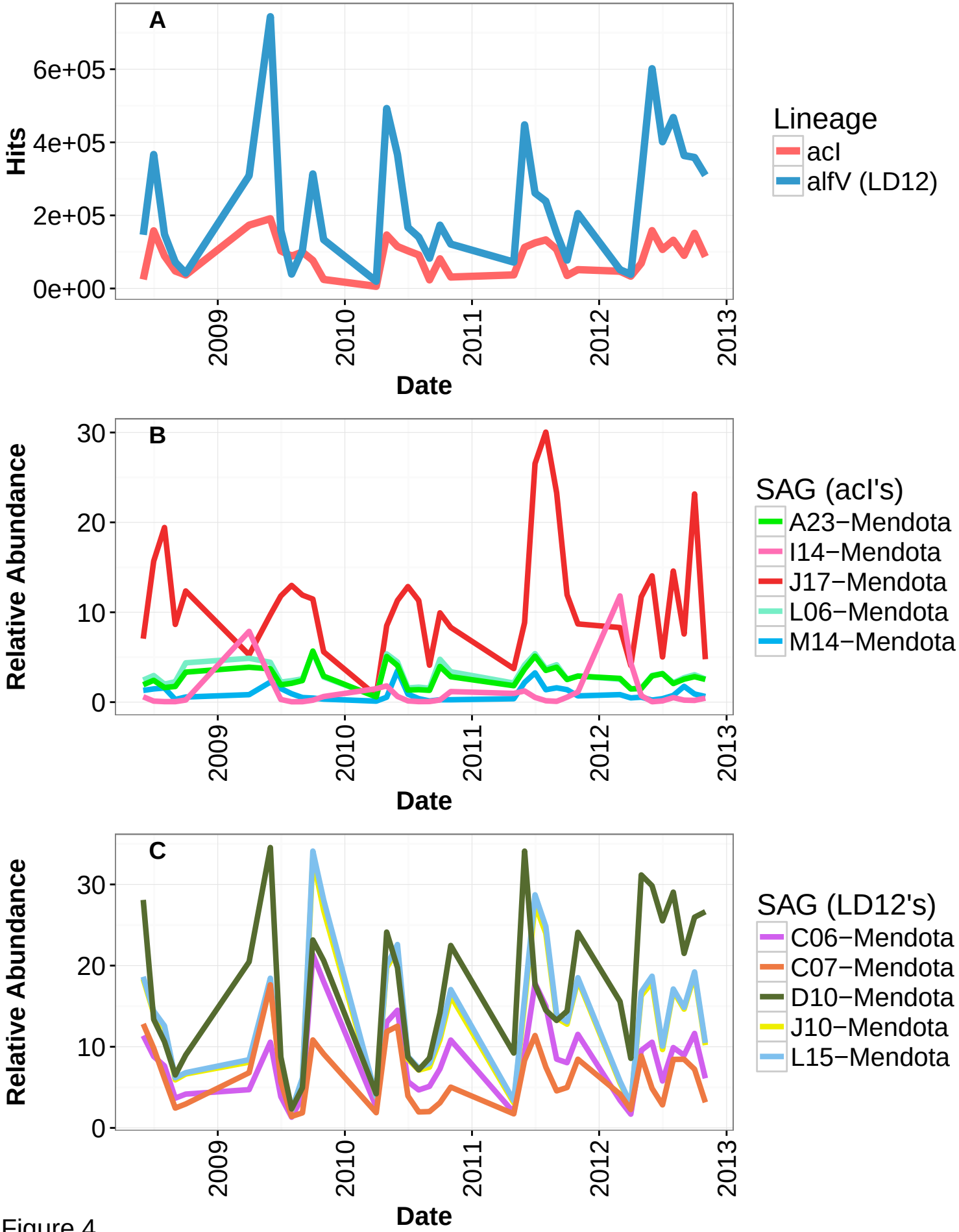

A
Hits
6e+05
4e+05
2e+05
0e+00
2009
2010
2011
2012
2013
Date
Lineage
acI
alfV (LD12)
B
Relative Abundance
30
20
10
0
2009
2010
2011
2012
2013
Date
SAG (acI's)
A23−Mendota
I14−Mendota
J17−Mendota
L06−Mendota
M14−Mendota
C
Relative Abundance
30
20
10
0
2009
2010
2011
2012
2013
Date
SAG (LD12's)
C06−Mendota
C07−Mendota
D10−Mendota
J10−Mendota
L15−Mendota

Figure 4

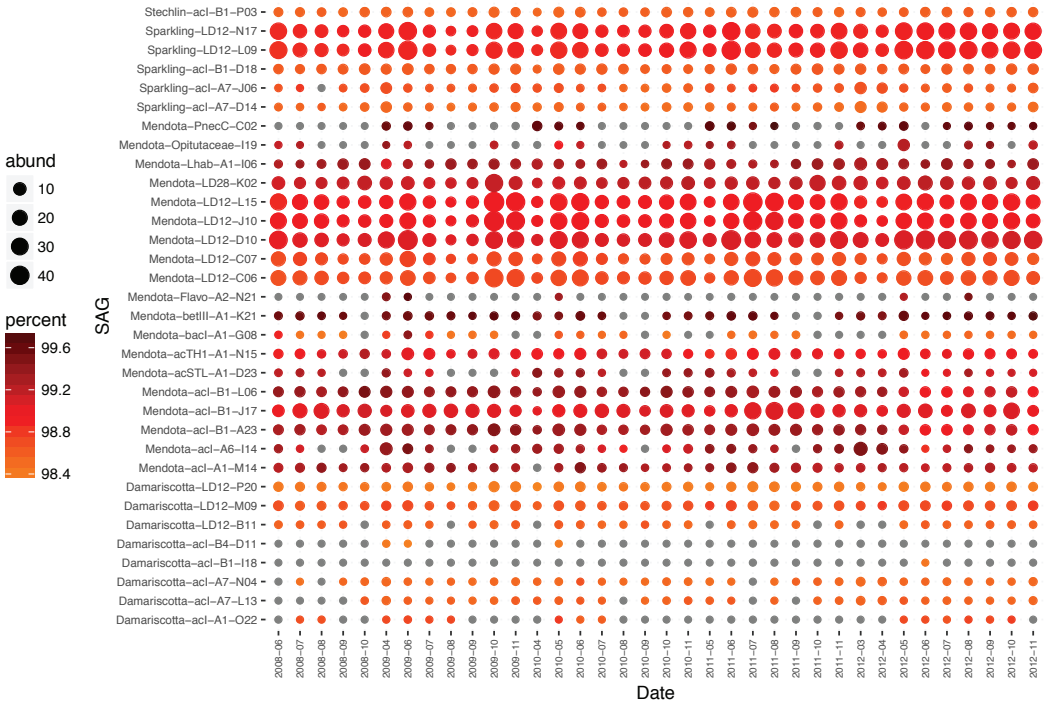


Figure 5A

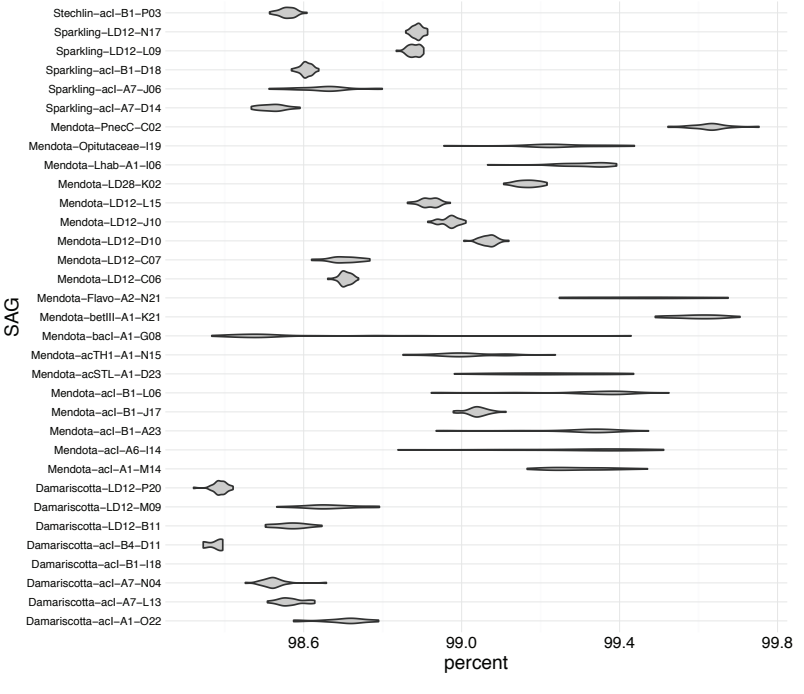


Figure 5B

Genome name	Phylum/Class	Tribe	Lake	Collection date (M/D/Y)	Assembly size (Mb)	Est. Genome Comp. (checkm)
AAA278-O22	Actinobacteria	acI-A1	Damariscotta	09/18/09	1.14	74.4
AAA027-M14	Actinobacteria	acI-A1	Mendota	12/5/09	0.82	43.1
AAA028-I14	Actinobacteria	acI-A6	Mendota	12/5/09	0.78	39.66
AAA041-L13	Actinobacteria	acI-A7	Damariscotta	04/28/09	1.38	74.14
AAA044-N04	Actinobacteria	acI-A7	Damariscotta	04/28/09	1.29	79.59
AAA023-J06	Actinobacteria	acI-A7	Sparkling	05/28/09	0.70	34.48
AAA024-D14	Actinobacteria	acI-A7	Sparkling	05/28/09	0.78	48.4
AAA278-I18	Actinobacteria	acI-B1	Damariscotta	09/18/09	0.94	63.73
AAA027-J17	Actinobacteria	acI-B1	Mendota	12/5/09	0.97	65.26
AAA027-L06	Actinobacteria	acI-B1	Mendota	12/5/09	1.16	76.59
AAA028-A23	Actinobacteria	acI-B1	Mendota	12/5/09	0.83	57.56
AAA023-D18	Actinobacteria	acI-B1	Sparkling	05/28/09	0.75	44.22
AB141-P03	Actinobacteria	acI-B1	Stechlin	05/25/10	0.66	45.98
AAA044-D11	Actinobacteria	acI-B4	Damariscotta	04/28/09	1.15	66.18
AAA027-D23	Actinobacteria	acSTL-A1	Mendota	12/5/09	0.94	44.01
AAA028-N15	Actinobacteria	acTH1-A1	Mendota	12/5/09	0.83	45.98
AAA280-B11	Alphaproteobacteria	LD12	Damariscotta	09/18/09	0.67	51.24
AAA280-P20	Alphaproteobacteria	LD12	Damariscotta	09/18/09	0.72	65.06
AAA487-M09	Alphaproteobacteria	LD12	Damariscotta	09/18/09	0.63	53.15
AAA027-C06	Alphaproteobacteria	LD12	Mendota	12/5/09	0.78	82.29
AAA027-J10	Alphaproteobacteria	LD12	Mendota	12/5/09	0.79	69.56
AAA027-L15	Alphaproteobacteria	LD12	Mendota	12/5/09	0.72	68.68
AAA028-C07	Alphaproteobacteria	LD12	Mendota	12/5/09	0.85	74.12
AAA028-D10	Alphaproteobacteria	LD12	Mendota	12/5/09	0.93	81.64
AAA023-L09	Alphaproteobacteria	LD12	Sparkling	05/28/09	0.77	68.1
AAA024-N17	Alphaproteobacteria	LD12	Sparkling	05/28/09	0.33	30.19
AAA028-K02	Alphaproteobacteria	LD28	Mendota	12/5/09	0.56	34.48
AAA027-G08	Bacteroidetes	bacI-A1	Mendota	12/5/09	1.32	59.36
AAA027-N21	Bacteroidetes	Flavo-A2	Mendota	12/5/09	2.21	92.44
AAA027-K21	Betaproteobacteria	betIII-A1	Mendota	12/5/09	1.38	42.1
AAA027-I06	Betaproteobacteria	Lhab-A1	Mendota	12/5/09	1.52	39.38
AAA027-C02	Betaproteobacteria	PnecC	Mendota	12/5/09	1.27	61.93
AAA027-I19	Verrucomicrobia	Opiputaceae	Mendota	12/5/09	2.42	54.58