1  **Contrasting patterns of genome-level diversity across distinct co-occurring bacterial**
2  **populations**
3
4  Sarahi L Garcia[a,b,*], Sarah L R Stevens[a,*], Benjamin Crary[c], Manuel Martinez-Garcia[d], Ramunas
5  Stepanauskas[e], Tanja Woyke[f], Susannah G Tringe[f], Siv Andersson[g], Stefan Bertilsson[b], Rex R.
6  Malmstrom[f], Katherine D McMahon[a,c]
7
8  a Department of Bacteriology, University of Wisconsin-Madison Madison, WI, USA
9  b Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala,
10  Sweden
11  c Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA
12  d Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain.
13  e Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA
14  f DOE Joint Genome Institute, Walnut Creek, CA, USA
15  g Department of Molecular Evolution, Uppsala University Uppsala, Sweden
16  * Equal contributors
17
18
19
20
21  Corresponding Author:
22  Katherine McMahon
23  1550 Linden Drive
24  Madison, WI  53711
25  trina.mcmahon@wisc.edu
26
27

1

## 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

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

101    "tribes").

102         The SAGs in this study originated from four lakes geographically isolated from one

103    another and represent a rich source of reference genomes that can be used to recruit

104    metagenomic reads in order to study population structure and dynamics through time in

105    naturally assembled communities. In particular, the contrasting origin of these SAGs provide

106    the opportunity to assess the differences in populations belonging to the same "tribe" while

107    having evolved in different island-like habitats (i.e. lakes). Two of the lineages featured in the

108    present study are the abundant and ubiquitous freshwater Actinobacteria acI and

109    Alphaproteobactera alfV containing the freshwater SAR11 sister-clade, LD12. Members of

110    these lineages are intriguing in their own right, as they represent groups of free-living

111    ultramicrobacteria that dominate many freshwater ecosystems [22-28]. They differ markedly

112    with respect to within-lineage diversity: LD12 is the sole tribe defined within the freshwater

113    alfV lineage, while the acI lineage is comprised of 13 tribes [21]. The acI and LD12 have no

114    axenic cultured representatives and share a large number of genomic and cellular traits. First,

115    both lineages have genomes with GC content values lower than 40% and estimated sizes of

116    about 1.5 Mb or less [29-31]. These genome characteristics are all the more striking since most

117    cultivated species in the Alphaproteobacteria and Actinobacteria have GC-rich genomes up to

118    10 Mb in size. Second, both lineages have evolved by massive gene loss [30,32]. Third, the

119    fraction of gained genes is only about 10% of the lost genes. Fourth, both groups of bacteria

120    have small cell volumes[27,28]. However, acI and LD12 seem to employ different substrate

121    niche specialization. While acI is thought to primarily use polyamines, oligopeptides and

122    carbohydrates, LD12 specializes in carboxylic acids and lipids [29,33].

123         By combining genome information from twenty-one previously published [29,30] and

124    twelve new SAGs from different freshwater lineages and an extensive five-year time series of

125   lake metagenomes (94 samples), we investigated the population structure of such ubiquitous

126   freshwater bacteria for the first time. Our results confirm the existence of coherent sequence-

127   discrete populations within these ubiquitous freshwater bacterial groups in natural

128   communities and we could trace the abundance and gANI of these populations over monthly

129   to seasonal time scales. Our work demonstrates the power of combining time-series

130   metagenomics and single cell genomics for studying bacterial diversification and for

131   describing ecologically meaningful population structure within the uncultured majority

132   inhabiting natural ecosystems.

133

## Results

135   *The SAG collection represents multiple clades within cosmopolitan freshwater lineages*

136   We analyzed 33 SAGs from four different freshwater lakes. Twenty-one of these

137   SAGs were previously analyzed for their genomic features and phylogenetic relationships [29-

138   31,34]. The 33 SAGs had total assembly sizes between 0.33 and 2.42 Mbp and were organized

139   into 8 to 103 contigs with GC contents between 29.1% and 51.7% (**Table 1**). Estimated

140   genome completeness, calculated using two different methods, ranged between 30% and

141   99%. Throughout the paper we will use mostly the shorter name version to facilitate reading,

142   for example, M14 in place of AAA027-M14.

143   The 33 SAGs in the study represent fifteen different previously defined freshwater

144   "tribes" that are each monophyletic and defined by >97.9% within-clade 16S rRNA gene

145   sequence identity, measured across the nearly full-length 16S rRNA gene [20,21]. Freshwater

146   microbial ecology researchers generally discuss and track these tribes as if they were

147   coherent units that are ecologically distinct from one another. Ten tribes are represented by

148   only one SAG each, while four tribes (LD12, acI-A1, acI-A7 and acI-B1) have more than one

149   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

175    the most complete SAGs from the acI-B1 and LD12 tribes as reference genomes and

176    calculated nucleotide identity using a sliding window with other SAGs from the same

177    respective tribe and visualized the results as a frequency distribution (**Figure 2**). The acI-B1

178    SAGs featuring the highest gANI (L06 and A23) were both from Lake Mendota and shared

179    nucleotide identity consistently greater than 95% with a peak at 99-100%. The acI-B1 SAG

180    P03 recovered from a lake in Germany had a frequency distribution with a peak more near

181    97% and a distinctly different shape. Other acI-B1 SAGs shared genomic regions with

182    primarily 80-85% nucleotide identity. This was even true for J17, which was also collected

183    from Lake Mendota and shared an average gANI of 79% with L06/A23 (**Table S1**),

184    suggesting that cells belonging to the same tribe (acI-B1) and living in the same environment

185    can have substantial genetic differences. The LD12 SAGs, which all belonged to the same

186    tribe, also displayed three distinct patterns, with one peak near 85%, several near 91%, and

187    two near 97%. Lake origin did not appear to explain these differences. That is, some LD12

188    cells from Lake Mendota were more similar to LD12 cells from Sparkling Lake than to other

189    LD12 cells from Lake Mendota.

190

191    *Diversity and structure of wild populations inferred using SAGs*

192        The variety of patterns observed in **Figure 2** indicates substantial within-tribe

193    variability even among cells recovered from the same lake.  This made us wonder if tribes

194    were composed of genetically and ecologically distinct populations coexisting in the same

195    environment. SAGs can serve as relevant reference points to study the diversity of uncultured

196    populations sampled using shotgun metagenomics by recruiting metagenomic reads and

197    examining the extent of nucleotide identity for each aligned read [8]. The results can also be

198    used to identify sequence-discrete populations whose boundaries are revealed by recruitment

199    patterns and specifically the dramatic drop in coverage observed around 95% sequence

8

200 identity [7,18,19]. To examine the diversity and structure of wild freshwater bacterial populations,

201 metagenomic reads from Lake Mendota, WI, USA, were mapped to the 33 SAGs, 19 of

202 which were collected from this lake.

203       Each of the SAGs was first used to recruit reads from a single metagenomic dataset

204 collected from Lake Mendota on 29 April 2009 (**Figure S1**). This time point was chosen

205 because it was the sample collected closest to the date on which the single cells were

206 collected (12 May 2009). Frequency distribution plots of the same data (**Figure 3**) revealed

207 patterns that were similar to those obtained with SAG pairs (**Figure 2**). The five acI-SAGs

208 from Lake Mendota (J17, L06, A23, M14 and I14) recruited more reads than the acI-SAGs

209 from other lakes, with many reads recruiting at nucleotide identity greater than 97.5%

210 (**Figure 3A**). All of the acI-SAGs also recruited many reads at 60 – 90% identity (**Figure 3A**

211 **and D**), creating the characteristic bimodal distribution observed in previous work [7]. Based

212 on these results, we hereafter consider reads sharing > 97.5% nucleotide identity as coming

213 from the same, operationally defined *population* as the reference SAG.  Thus, the acI lineage

214 in Lake Mendota was composed of multiple sequence-discrete populations.  Interestingly, the

215 acI-B1 tribe in Lake Mendota, a subset of the acI lineage, appeared to be composed of at least

216 two coexisting and genetically distinct populations, one represented by SAG J17 and the

217 other by SAGs A23 and L06, consistent with the pairwise gANI observed using only the

218 SAGs (**Figure 2**).

219       To determine if we recovered representative SAGs from all acI populations in Lake

220 Mendota, we next performed recruitments competitively, allowing each read to only map to

221 the SAG with the greatest % identity (**Figure S2**). As the patterns in **Figure 3** were generated

222 by non-competitive mapping, some reads mapping with 100% similarity to one SAG might

223 for example also have mapped with 60-90% similarity to SAGs from different sequence-

224 discrete populations.  After competitive mapping the resulting frequency distributions

9

225    changed and the fraction of reads mapping with 60-90% identity to each acI SAG dropped

226    dramatically (**Figure S2**). However, a secondary peak around 80% identity still remained in

227    most cases, and it is possible these reads originated from cells belonging to other acI

228    populations lacking a representative SAG.

229         LD12 SAGs collected from Lake Mendota (C06, J10, L15, C07 and D10) also had a

230    distinctive peak of recruited reads at >97.5% sequence identity (**Figure 3B**), although the

231    overall shape of the recruitment patterns differed dramatically from those of the acI lineage.

232    For example, LD12 SAGs had a secondary recruitment peak at ~92% identity whereas the acI

233    SAGs had secondary peaks at ~75% with non-competitive mapping. This suggests the

234    sequence-discrete populations within the LD12 tribe were more similar genetically than

235    populations comprising the acI-B1 tribe. In fact, the populations were sufficiently similar that

236    the hallmark coverage discontinuity below 97% similarity was not particularly pronounced

237    (**Figure 3B**). Under competitive recruiting conditions, the LD12 recruitment distribution

238    plots had remarkably different shapes (**Figure S2B and D**), as compared to the uncompetitive

239    recruiting conditions (**Figure 3B**), and each SAG had only a single peak at >97.5% identity.

240    This suggests the majority of LD12 cells in Lake Mendota belong to sequence-discrete

241    populations represented by the SAGs in our collection.

242         All but one (I06) of the other freshwater SAGs in this study that were collected from

243    Lake Mendota generated the distinctive read recruitment frequency peak above 97.5%

244    identity (**Figure 3C**) that was observed for acI (**Figure 3A**). A negligible number of reads

245    recruited to the SAGs collected from other lakes under the competitive recruiting conditions

246    (data not shown). Since each of these SAGs represent just one tribe, it is not appropriate to

247    infer any general conclusions for these populations or tribes, but we present them here to

248    show the intriguing diversity of recruitment patterns. We finally underscore the need to more

249    deeply sample individual population members using SAGs, to better capture and describe the

10

250     range of variation in population structure.

251

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

253     Results from a single metagenome sample suggested that individual tribes were

254     composed of multiple genetically distinct populations that could be delineated and tracked

255     using metagenomic read recruitment. Next we hypothesized that these populations might also

256     be ecologically distinct and fill different realized niches. If so, we might expect these

257     populations to display different temporal abundance patterns. We followed changes in

258     population abundance through time by recruiting reads from a five-year metagenomic time-

259     series applying a nucleotide identity cutoff of 97.5%. SAGs from the LD12 tribe recruited

260     more reads than all of the acI SAGs summed together, on almost all sample dates (Figure

261     4A).

262     Using the relative number of reads recruited as a proxy for abundance, we found the

263     J17 population, which belonged to the acI-B1 tribe, to be the most abundant acI population in

264     almost every sample (**Figure 4B and 5A**). The abundance of the J17 population was poorly

265     correlated over time with the other acI-B1 population represented by L06 and A23 (maximum

266     Spearman rank correlation = 0.294), indicating each population had a different temporal

267     abundance pattern.  This suggests the two sequence-discrete populations comprising the acI-

268     B1 tribe were also ecologically distinct. The different tribes of acI, which were more distantly

269     related than the populations within the B1 tribe, also displayed different abundance patterns

270     in Lake Mendota. For example, the acI-A1 I14 SAG population peaked in spring, but at

271     markedly higher levels in 2009 and 2012 than in other years (**Figure 4B**). The acI-A6 I14

272     SAG population was consistently in low abundance compared to other tribes, but had small

273     peaks in June and July.

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

11

275  similar abundance patterns. (**Figure 4C and S3**). The abundances of J10, L15, and C06

276  populations were strongly correlated (Spearman rank correlation = 0.997-0.999) and tended

277  to peak both in Spring and Fall (Figure S3).  The D10 population was the most abundant in

278  the dataset but its abundance was not as strongly correlated to the other LD12 populations

279  (Spearman rank correlation = 0.712-0.725). The C07 population was the least abundant but

280  was also correlated to both the J10-L15-C06 populations and the D10 population (Spearman

281  rank correlation = 0.861-0.873). Based on the similar temporal abundance patterns, the

282  ecological differences among genetically discrete LD12 populations might be small, at least

283  compared to the presumed major differences among acI-B1 populations that resulted in

284  substantially different abundance patterns.

285

286  *Does the genetic diversity of populations change over time?*

287  We also examined the extent to which within-population diversity varied through time

288  by quantifying changes in population-wide ANI, i.e. the average identity of all reads mapping

289  with at least 97.5% identity (**Figure 5B**). For Mendota SAGs, the more abundant populations

290  (such as LD12 and acI-B1 J17) generally had lower population-wide ANI variance through

291  time compared to some less abundant populations (such as acSTL-A1-D23 and acI-A6-I14).

292  For example, the SAG bacI-A1 G08 population had relatively high population-wide ANI in

293  June 2009, around the time when the sample was collected for SAG library collection, but

294  had markedly lower ANI on all other dates. One interesting exception to this observation was

295  a significantly lower ANI for the relatively abundant acI-B1 L06-A23 population in 2012, as

296  compared to 2007-2011 (Mann-Whitney U test p=1.4e-06). However, we note that for those

297  populations recruiting a very low number of reads, it is possible that the sampling is not deep

298  enough to reflect the true ANI value of the populations thus leading to a higher observed

299  variance.

12

300

## Discussion

301

302 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).

312 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.

13

325        Pair-wise gANI analysis of SAGs and metagenomic read recruitment indicated that

326    cells belonging to the same tribe but inhabiting different lakes were usually genetically

327    distinct. For example, SAGs collected from other lakes generally recruited very few reads

328    from Lake Mendota at ANI >97.5% (**Figure 5**) while many recruited a substantial number of

329    reads in the 89-92% range (**Figure 3**). However there were two prominent exceptions: LD12

330    N17 and L09, both of which are from Sparkling Lake. N17 and L09 share 97% gANI with

331    Mendota SAG D10, which is substantially higher than the average (88%) and median (90%)

332    within-tribe gANI (Table S1). These SAGs also recruited roughly the same number of reads

333    with >97.5% identity as did the LD12 SAGs from Lake Mendota, though around 20% of the

334    base pairs in the genomes did not recruit any reads (17% for L09 and 23% for N17). This

335    implies that some gene content was present in the Sparkling Lake populations but missing in

336    Lake Mendota. However, 10% of the base pairs in the D10 genome also did not recruit any

337    reads, even though it was from Lake Mendota. This rare genome content could represent

338    flexible or low frequency genes in the population, or contamination in the SAG preparation.

339    We examined the phylogenetic distribution of low-coverage contigs and did not discern any

340    evidence of contamination.

341        In Lake Mendota, acI cells are organized into genetically discrete populations, but the

342    forces creating this organization remain a mystery. The consistent lack of coverage around

343    90-97% identity in recruitment plots indicates Lake Mendota lacks acI genotypes sharing this

344    degree of sequence similarity with our SAGs, or at least that these putative genotypes were

345    consistently at much lower abundances than their close relatives over the five years surveyed.

346    This raises the questions "how do sequence discrete populations persist?" and "why don't we

347    detect a continuum of genotypes across the range of 90-97% identity?". The P03 SAG from

348    Stechlin Lake shares gANI of 96% with acI-B1 SAGs from Mendota, indicating that

349    genotypes within this locally excluded sequence space do exist, at least as long as they are

14

350  from different environments. We infer the persistence of the coverage discontinuity between

351  populations to be less a factor of dispersal limitation and more likely the result of competitive

352  exclusion and barriers to recombination within Mendota populations.  Additional SAG and

353  metagenomic studies are necessary to determine if the same forces maintain the coverage

354  discontinuities among sequence discrete populations observed in other phylogenetic groups

355  and in different environments.

356      We know that both acI tribes and LD12 vary in abundance over seasonal and annual

357  time-scales, based on previous work using 16S rRNA gene sequencing, quantitative PCR, and

358  FISH [27,28,37,38]. Here we used our SAGs to track such populations at monthly intervals over

359  five years. The results confirmed prior work that showed acI tribes and LD12 are among the

360  most abundant non-cyanobacterial groups in Lake Mendota [21,39] but also revealed dynamics

361  at unprecedentedly high phylogenetic resolution. Based on our extensive comparison of how

362  SAGs recruited relative to one another, we are confident that our metagenomic recruitment

363  filters allowed us to delineate discrete populations that would not be possible to resolve using

364  more traditional and widely used methods (e.g. 16S rRNA gene sequencing or FISH).

365  Specifically, metagenomic recruitments to LD12 SAGs revealed strikingly different patterns

366  compared to the acI lineage, suggesting fundamental differences in evolutionary history

367  and/or lifestyles among such abundant and ubiquitous freshwater bacteria.  We discovered

368  that LD12 populations were not as strongly genetically separated as acI populations; pair-

369  wise gANIs between SAGs were higher and recruitment plots showed secondary peaks

370  between 90-95% identity (**Figure 3B**), the same range where coverage of acI SAGs was at a

371  minimum (**Figure 3A**). Under a competitive recruitment analysis, wherein each read is

372  counted only once and attributed to the best match SAG, the secondary peaks disappear

373  (**Figure S2**), indicating the LD12 SAGs represent highly similar, but still genetically discrete,

374  populations. Temporal abundance patterns of these LD12 populations were strongly

15

375  correlated over five years, whereas acI populations showed much lower correlation within

376  tribes. This suggests that the acI-B1 populations are ecologically distinct (i.e. occupying

377  temporally discrete niches) while LD12 populations are less differentiated genetically, and

378  possibly ecologically neutral, leading to co-occurrence and synchronization of temporal

379  abundance patterns. LD12 is a particularly fascinating group because it is also a subclade of

380  the broader SAR11 clade, with hypothesized ancient transition from marine to freshwater [40]

381  followed by specialization through gene flux and mutation, with comparatively low

382  recombination rates [30]. Over time, low recombination rates should lead to large genetic

383  divergence among coexisting populations. Thus, we propose that LD12 populations are

384  simply at earlier stages of differentiation as compared to acI populations while we cannot

385  exclude that something fundamental about their lifestyle is "holding" the populations together

386  genetically and ecologically. In any case, the lack of coherence among acI-B1 populations

387  challenges our concept of tribes as ecologically coherent units and will force freshwater

388  microbial ecologists to re-examine conventions for tracking these units through space and

389  time.

390      Our observations stand in contrast to those reported for the marine species *Vibrio*

391  *cyclitrophicus* [41]. Shapiro and colleagues examined strains inhabiting large-size (L) and

392  small-size (S) particles and considered their particle-association to represent ecological

393  differentiation. These L and S strains had an average across-group gANI of 99.0% (**Table S6**)

394  and would not appear genetically distinct using the metagenomic read recruitment method

395  applied in our study. That is, the *V. cyclitrophicus* L and S strains appear to be less genetically

396  differentiated than LD12 but possibly more ecologically differentiated. Thus, our work

397  provides further evidence for conceptual models of bacterial evolution in which different

398  lineages can diversify in different ways, and no single mode will explain all extant diversity.

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

16

400    diversity varied within and among populations as well as how diversity changed over time.

401    We calculated the population-wide ANI for reads that recruited only above 97.5% and found

402    the resulting value was remarkably stable through time for most of the abundant populations

403    (**Figure 5B**). This was particularly true for the LD12 populations. However, one striking

404    contrast was the acI-B1 population represented by L06/A23, which had consistent

405    population-wide ANI of 99.3% during 2008-2011 but 99.0% during 2012 (Mann-Whitney U

406    test p=1.4e-06).   This change suggests a substantial shift in the relative abundance of

407    genotypes comprising the population. Similar shifts were observed previously in sequence-

408    discrete populations inhabiting Trout Bog Lake, indicating this could be a common

409    phenomenon among freshwater clades [19]. Unlike the genome-wide selective sweep observed

410    in one *Chlorobium* population from Trout Bog Lake, the distribution of single nucleotide

411    polymorphisms within the L06/A23 population before and after 2012 exhibited no clear

412    pattern of gene- or genome-wide sweep (data not shown). It is difficult or impossible to

413    separate genotypes within sequence-discrete populations using short-read shotgun

414    sequencing, so further work using long-read technologies will be needed to link SNPs in

415    populations to individual genomes. This kind of approach will likely be required to tease

416    apart the paths leading to diversification within and among populations.

417

## Methods

418

419    *Single amplified genomes (SAGs)*

420    Water samples (1-ml) were collected from the upper 0.5m to 1m of each of four lakes

421    (Mendota, Sparkling, Damariscotta, Stechlin) and cryopreserved, as previously described

422    [31,42]. These lakes were originally selected because they represent different freshwater trophic

423    status (eutrophic, oligotrophic, mesoeutrophic, and oligotrophic, respectively) and geographic

424    regions (Wisconsin and Maine, USA, and Germany). Bacterial single amplified genomes

17

425 (SAGs) were generated by fluorescence-activated cell sorting (FACS) and multiple

426 displacement amplification (MDA), and identified by PCR-sequencing of their 16S rRNA

427 genes at the Bigelow Laboratory Single Cell Genomics Center (SCGC; http://scgc.bigelow.org).

428 Thirty-two SAGs from lakes Mendota, Sparkling and Damariscotta were selected for

429 sequencing based on the previously sequenced 16S rRNA gene as well as the kinetics of the

430 MDA reactions [42]. The one SAG from Lake Stechlin was selected from a separate library

431 because its 16S rRNA gene was 100% identical to an acI-B1 SAG previously analyzed

432 (AAA027-L06) [31]. In the present study we analyze 21 previously published and 12 new

433 SAGs. All 33 SAGs were analyzed (**Table 1**) after genome sequencing, assembly,

434 contamination removal and annotation as previously described (Ghylin et al. 2014).

435 Estimation of completeness was done using CheckM [43] and the gene markers from a recent

436 study examining a large collection of draft environmental genomes [44].

437

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

439 A phylogenomic analysis was conducted using PhyloPhlAn [35]. ANI was calculated by

440 using the method described in [11] with fragment size of 1000, minimum alignment length of

441 700 bp, percent identity of 70, and e-value of 0.001. AAI was calculated by averaging the

442 identity of the reciprocal best hits from the BLASTP searches of the predicted proteins for

443 each pair of genomes. 16S rRNA gene similarity for each pair was calculated using the

444 overlapping region in an alignment created using a multiple alignment (default options) in

445 Geneious Version R6 [45].

446

447 *SAG-to-SAG recruitments*

448 SAG pairs from the same tribe were used to examine the frequency distribution of

449 nucleotide identities across homologous regions of the two genomes. In order to create a

450 sliding window for comparison, the contigs of all SAGs were shredded into 301bp fragments

18

451   with 150 bp overlap. Two SAGs were selected as reference genomes: L06 as the most

452   complete from the tribe acI-B1 and C06 as the most complete LD12. The contigs of each of

453   the two selected SAGs were used as a reference for recruiting from the shredded SAGs using

454   Blast 2.2.28 [46]. Ribosomal RNAs were masked from the SAGs prior to performing blast.

455

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

457

458   Samples were collected from Lake Mendota over the course of five years, as

459   previously described [47,48]. Lake Mendota, Madison, Wisconsin, (N 43°06, W 89°24) is one of

460   the most well-studied lakes in the world, and is a Long Term Ecological Research site

461   affiliated with the Center for Limnology at the University of Wisconsin Madison [49]. It is

462   dimictic and eutrophic with an average depth of 12.8 m, maximum depth of 25.3 m, and total

463   surface area of 3938 ha. Depth integrated water samples were collected from 0 to 12 m of the

464   epilimnion (upper mixed layer) at 94 different time points during ice-free periods from

465   summer 2008 to summer 2012, and filtered onto 0.2 µm pore-size polyethersulfone filters

466   (Supor, Pall) prior to storage at -80°C. DNA was later purified from these filters using the

467   FastDNA kit (MP Biomedicals). DNA sequencing was performed at the Department of

468   Energy Joint Genome Institute using standard protocols (Walnut Creek, CA, USA). DNA

469   from the 94 samples was used to generate libraries that were sequenced on the Illumina

470   HiSeq 2000 platform. Paired-end sequences of 2 X 150bp were generated for all libraries.

471   Adapter sequences, low quality reads (i.e. ≥80% of bases had quality scores <20), and reads

472   dominated by short repeats of ≥3 bp were removed. The remaining high quality reads were

473   merged with the Fast Length Adjustment of Short Reads [50] with a mismatch value of ≤0.25

474   and a minimum of ten overlapping bases from paired sequences, resulting in merged read

475   lengths of 150 to 290 bp (**Table S3**). Metagenomes were pooled by month to reduce the time-

476   series data to 30 observations and increase coverage.

19

477
478       All contigs from each of the 33 SAGs were used as a reference to recruit reads from

479 the Mendota metagenomes using blastn. Metagenome reads that recruited to the SAGs were

480 filtered and only alignments 200bp or longer were considered. An additional filter requiring

481 an alignment percent identity of at least 97.5% was applied when analyzing the metagenome

482 time series. Ribosomal RNAs were masked from the SAGs prior to performing the

483 recruitments.

484

485 *Statistics, Visualization, Reproducible Methods*

486       Datasets were analyzed and results were visualized using custom scripts written in R

487 [51] and python. Pipeline and scripts for analysis can be found at

488 https://github.com/sstevens2/blast2ani.

489
490

## References

492
493 1    Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically
494      cohesive units. *Trends in microbiology*, (2014).
495 2    Luo, C. W. *et al.* Genome sequencing of environmental Escherichia coli expands
496      understanding of the ecology and speciation of the model bacterial species. *Proceedings of*
497      *the National Academy of Sciences of the United States of America* **108**, 7200-7205, (2011).
498 3    Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria.
499      *Bmc Biol* **3**, (2005).
500 4    Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of
501      individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143-169, (1995).
502 5    Kaeberlein, T., Lewis, K. & Epstein, S. S. Isolating "Uncultivable" Microorganisms in Pure
503      Culture in a Simulated Natural Environment. *Science* **296**, 1127-1129, (2002).
504 6    Little, A. E. F., Robinson, C. J., Peterson, S. B., Raffa, K. E. & Handelsman, J. Rules of
505      Engagement: Interspecies Interactions that Regulate Microbial Communities. *Annual Review*
506      *of Microbiology* **62**, 375-401, (2008).
507 7    Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal.
508      *Environmental microbiology* **14**, 347-355, (2012).
509 8    Stepanauskas, R. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol*
510      **15**, 613-620, (2012).
511 9    Hunt, D. E. *et al.* Resource partitioning and sympatric differentiation among closely related
512      bacterioplankton. *Science* **320**, 1081-1085, (2008).
513 10   Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and
514      their ecological interpretation. *Nat Rev Microbiol* **13**, 133-146, (2015).
515 11   Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for
516      prokaryotes. *Proceedings of the National Academy of Sciences of the United States of*
517      *America* **102**, 2567-2572, (2005).

| | | |
|---|---|---|
| 518 | 12 | Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Research* **43**, 6761-6771, (2015). |
| 520 | 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). |
| 523 | 14 | Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344**, 416-420, (2014). |
| 525 | 15 | Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43, (2004). |
| 527 | 16 | Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304**, 66-74, (2004). |
| 529 | 17 | Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77, (2007). |
| 531 | 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). |
| 533 | 19 | Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*, (2016). |
| 535 | 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). |
| 537 | 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). |
| 539 | 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). |
| 542 | 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). |
| 545 | 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). |
| 548 | 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). |
| 551 | 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). |
| 554 | 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). |
| 557 | 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). |
| 560 | 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). |
| 562 | 30 | Zaremba-Niedzwiedzka, K. *et al.* Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biology* **14**, (2013). |
| 564 | 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). |
| 566 | 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). |
| 569 | 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). |
| 571 | 34 | Eiler, A. *et al.* Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *ISME J*, (2016). |
| 573 | 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). |
| 576 | 36 | Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**, 784-798, (2010). |

578  37  Eiler, A., Heinrich, F. & Bertilsson, S. Coherent dynamics and association networks among
579      lake bacterioplankton taxa. *ISME J* **6**, 330-342, (2012).
580  38  Allgaier, M. & Grossart, H.-P. Diversity and Seasonal Dynamics of Actinobacteria Populations
581      in Four Lakes in Northeastern Germany. *Appl. Environ. Microbiol.* **72**, 3489-3497, (2006).
582  39  Newton, R. J. & McMahon, K. D. Seasonal differences in bacterial community composition
583      following nutrient additions in a eutrophic lake. *Environmental Microbiology* **13**, 887-899,
584      (2011).
585  40  Logares, R., Brate, J., Heinrich, F., Shalchian-Tabrizi, K. & Bertilsson, S. Infrequent
586      Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial
587      Lineages (SAR11). *Mol Biol Evol* **27**, 347-357, (2010).
588  41  Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of
589      bacteria. *Science* **336**, 48-51, (2012).
590  42  Martinez-Garcia, M. *et al.* High-throughput single-cell sequencing identifies photoheterotrophs
591      and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**, 113-123, (2011).
592  43  Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
593      assessing the quality of microbial genomes recovered from isolates, single cells, and
594      metagenomes. *Genome research*, (2015).
595  44  Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter.
596      *Nature* **499**, 431-437, (2013).
597  45  Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for
598      the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, (2012).
599  46  Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421,
600      (2009).
601  47  Shade, A. *et al.* Interannual Dynamics and Phenology of Bacterial Communities in a
602      Eutrophic Lake. *Limnology and Oceanography* **52**, 487-494, (2007).
603  48  Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L. & McMahon, K. D. A decade of seasonal
604      dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic
605      Lake Mendota, WI, USA. *Isme J* **7**, 680-684, (2013).
606  49  Carpenter, S. R. *et al. The ongoing experiment: Restoration of lake mendota and its
607      watershed". Long term dynamics of lakes in the landscape.* . (Oxford Press, 2006).
608  50  Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome
609      assemblies. *Bioinformatics* **27**, 2957-2963, (2011).
610  51  R Core Team. *R: A language and environment for statistical computing.*, <http://www.R-
611      project.org/.> (2014).
612
613

## Acknowledgements

## 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

23

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

640
641

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

643
644

## 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

25

671 SAGs. Analogous competitive recruitments that required each read to recruit to only one

672 SAG are presented in Figure S2. Each panel represents a different sub-set of the SAGs:  A.

673 acI from Mendota, B. acI not from Mendota, C. LD12 from Mendota, D. LD12 not from

674 Mendota, E. other freshwater groups from Mendota

675

676 **Figure 4.** Sequence-discrete population abundance in Lake Mendota, as measured by the

677 relative number of reads recruited to each SAG using blastn. All SAGs and samples are from

678 Lake Mendota. Timepoints are pooled by month. Filtering criteria: $\geq$97.5% ANI and $\geq$200 bp

679 alignment length. Blastn was done competitively, only counting the read for the best hit

680 genome with the exception of when it hit equally well, then counted for all best hit genomes.

681 Colors for each SAG are the same as in Figures 2 and 3. A. Sum of reads recruited for each

682 lineage. B. Sum of reads recruited for each acI-SAG. C. Sum of reads recruited for each

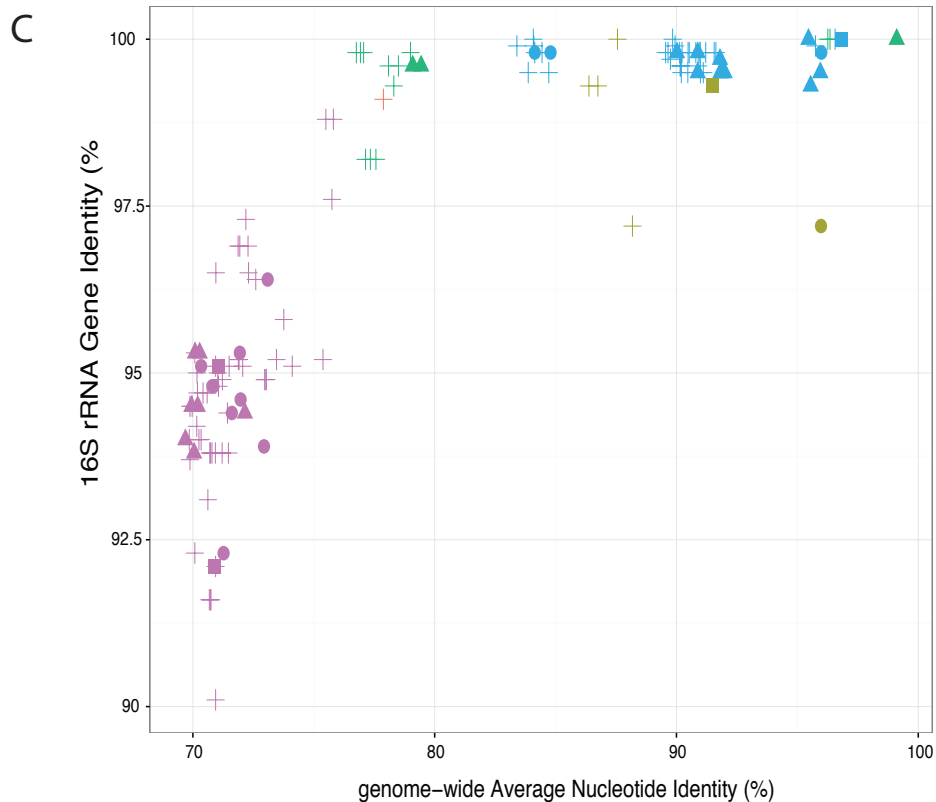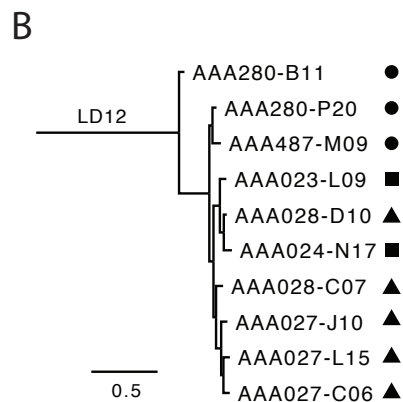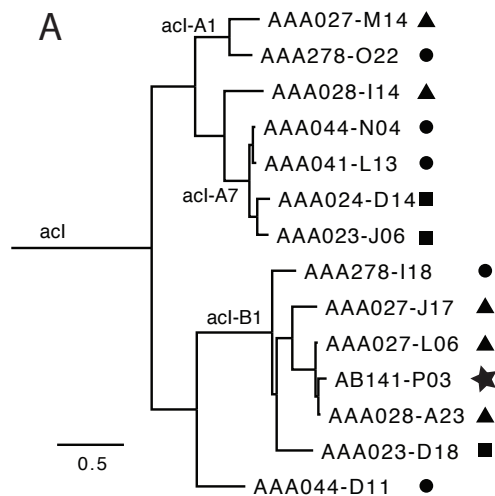683 LD12 SAG.

684
685

686 **Figure 5.** A. Metagenomic read recruitment using the SAGs from Lake Mendota. SAGs are

687 in rows with bubbles representing all metagenomes from a particular month recruited against

688 SAG. Filtering criteria are the same as those noted in Figure 4. Color scale indicates the ANI

689 of the recruited metagenome reads. Bubble size represents the average coverage per base in

690 the reference SAG divided by the size of the metagenome, multiplied by the average size of

691 all metagenomes (1.34 Gigabases). Note that the resulting values do not represent a true

692 measure of absolute abundance, but allow for quantitative comparison of month-to-month

693 variation in population-level abundance. Recruitments were not performed competitively,

694 meaning that each read could be recruited by multiple SAGs. B. Variation in ANI for each

695 SAG, across all 30 metagenomes from throughout the five years. The data underlying these

696 plots can be found in Table S5.

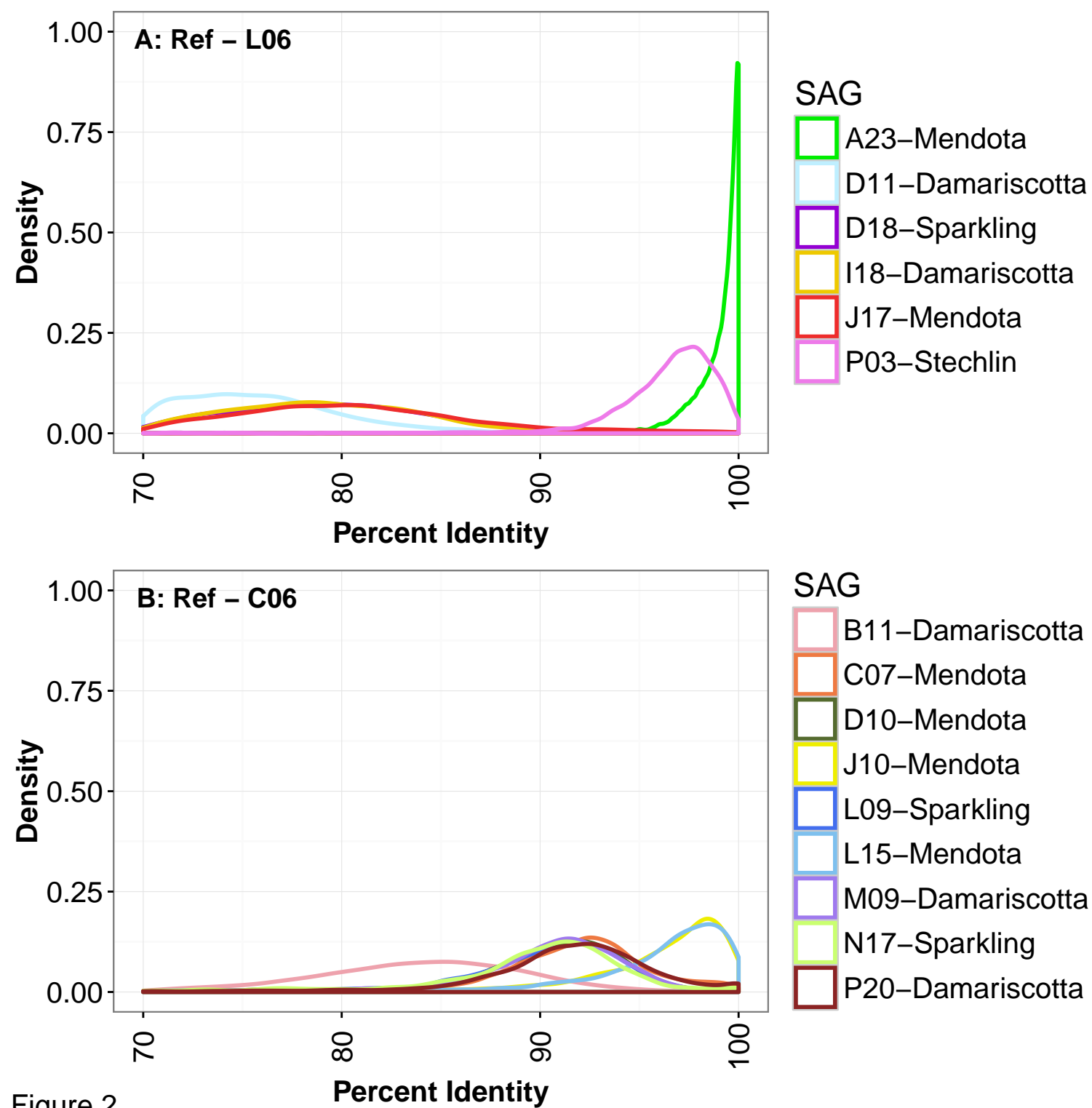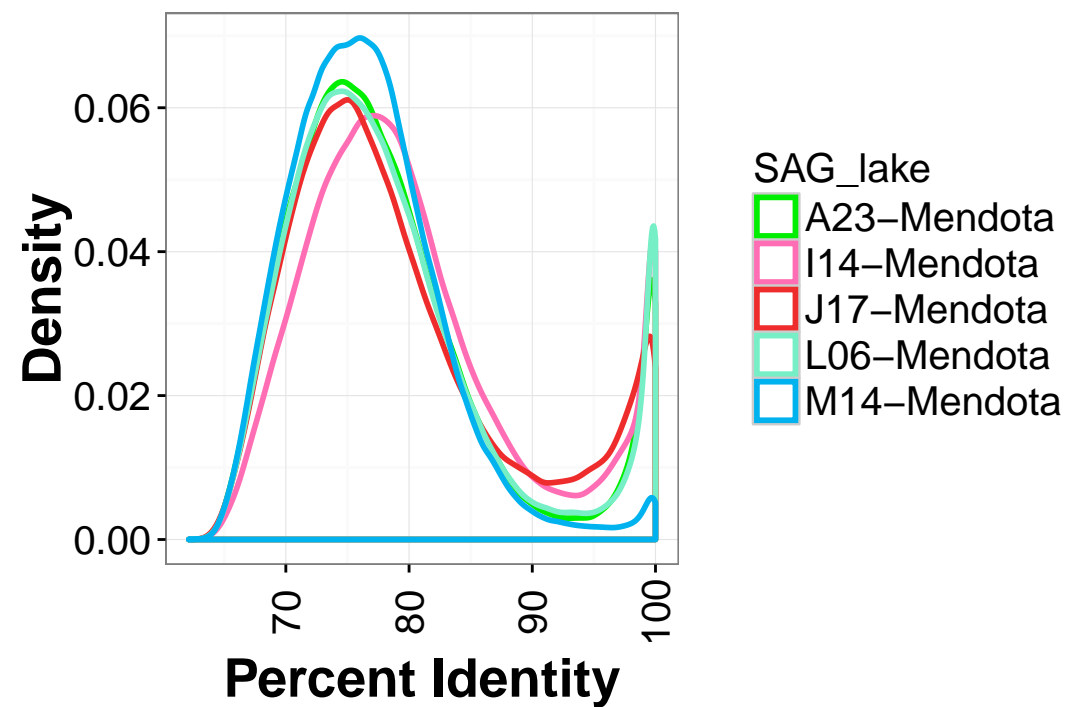26

697
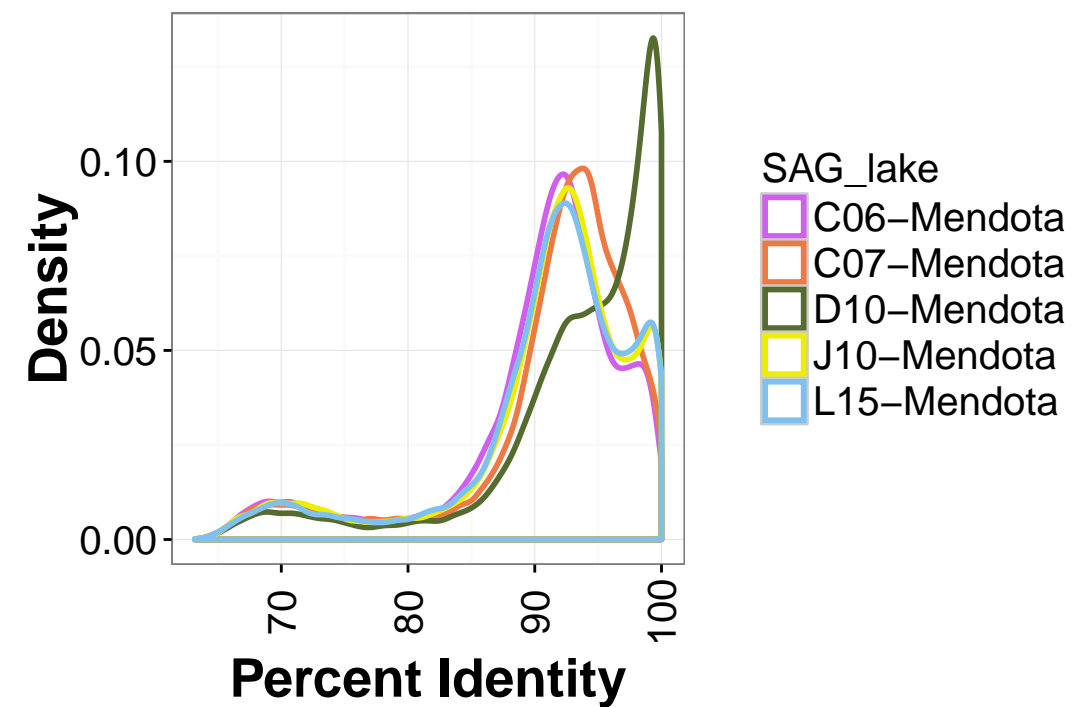
**Tables**

699

Table 1. Metadata for the 33 SAGs.

A

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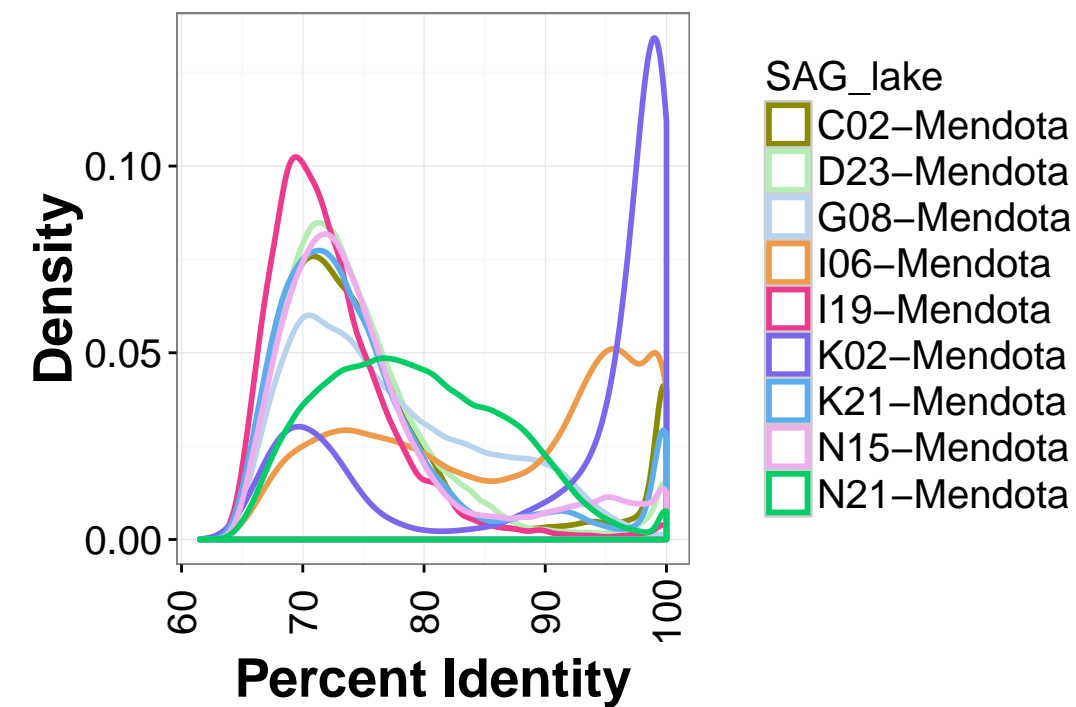        AAA278-O22 ●
        AAA028-I14 ▲
        AAA044-N04 ●
acl-A7  AAA041-L13 ●
        AAA024-D14 ■
        AAA023-J06 ■
acl ─
acl-B1  AAA278-I18 ●
        AAA027-J17 ▲
        AAA027-L06 ▲
        AB141-P03 ★
        AAA028-A23 ▲
        AAA023-D18 ■
        AAA044-D11 ●
0.5

B

LD12 ─ AAA280-B11 ●
       AAA280-P20 ●
       AAA487-M09 ●
       AAA023-L09 ■
       AAA028-D10 ▲
       AAA024-N17 ■
       AAA028-C07 ▲
       AAA027-J10 ▲
       AAA027-L15 ▲
       AAA027-C06 ▲
0.5

C

Lake: Damariscotta ● Mendota ▲ Sparkling ■ Diff Lakes +

Tribe: acl–A1 ● acl–A7 ● acl–B1 ● LD12 ● Diff Tribes ●
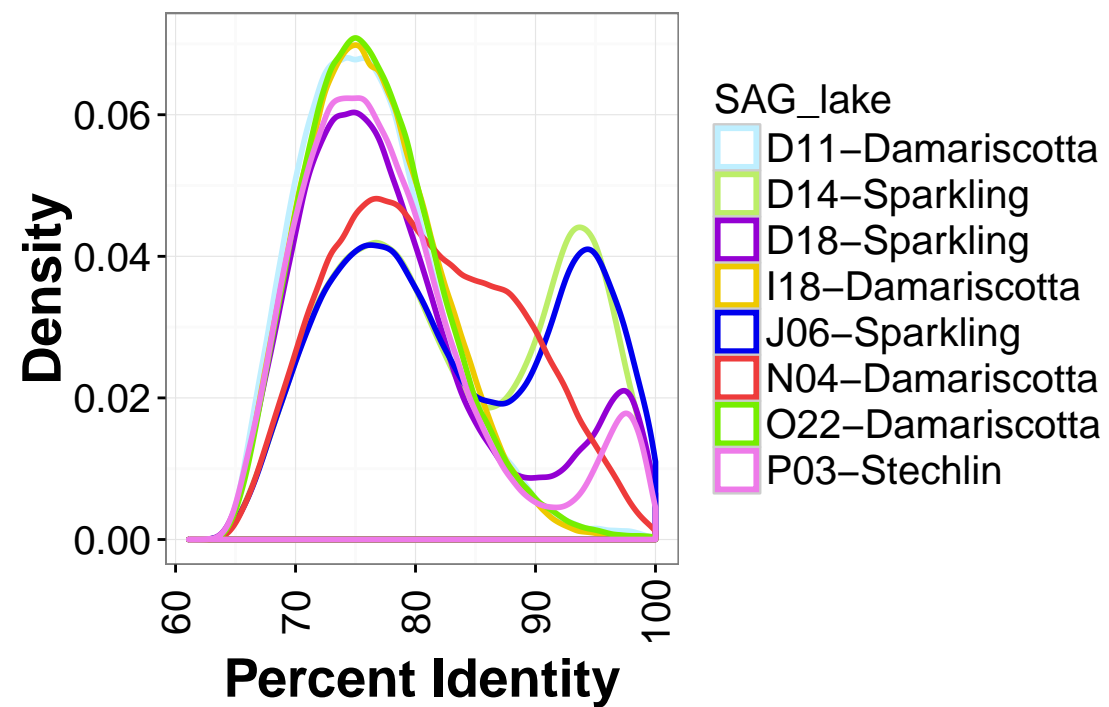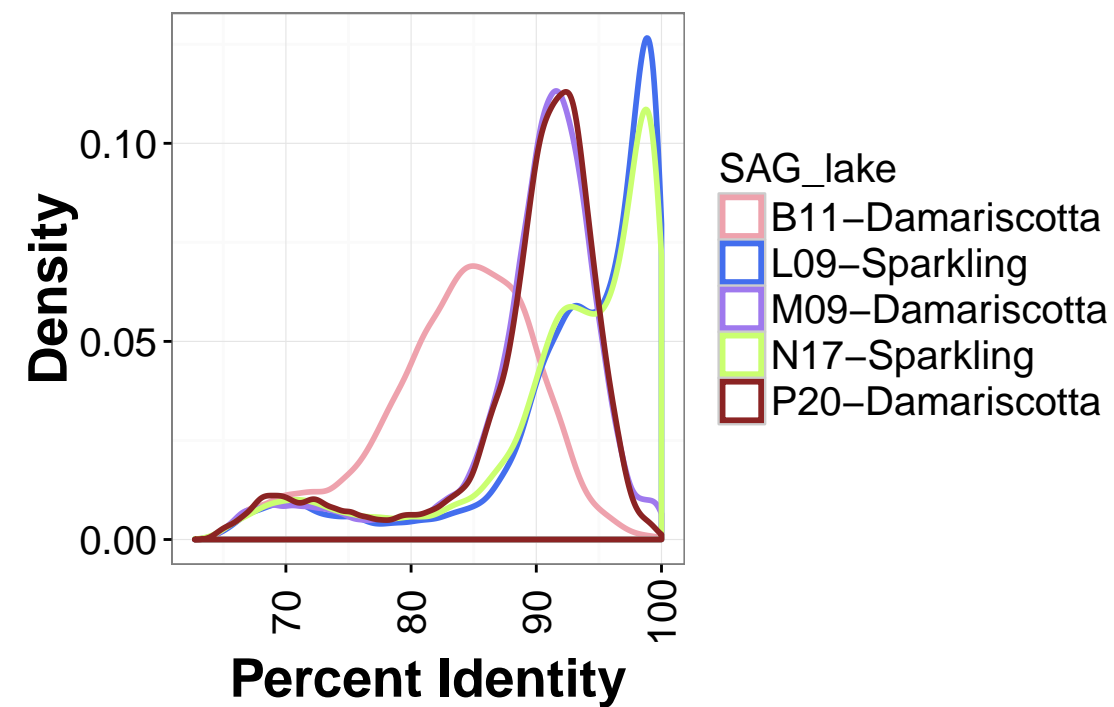
Figure 2
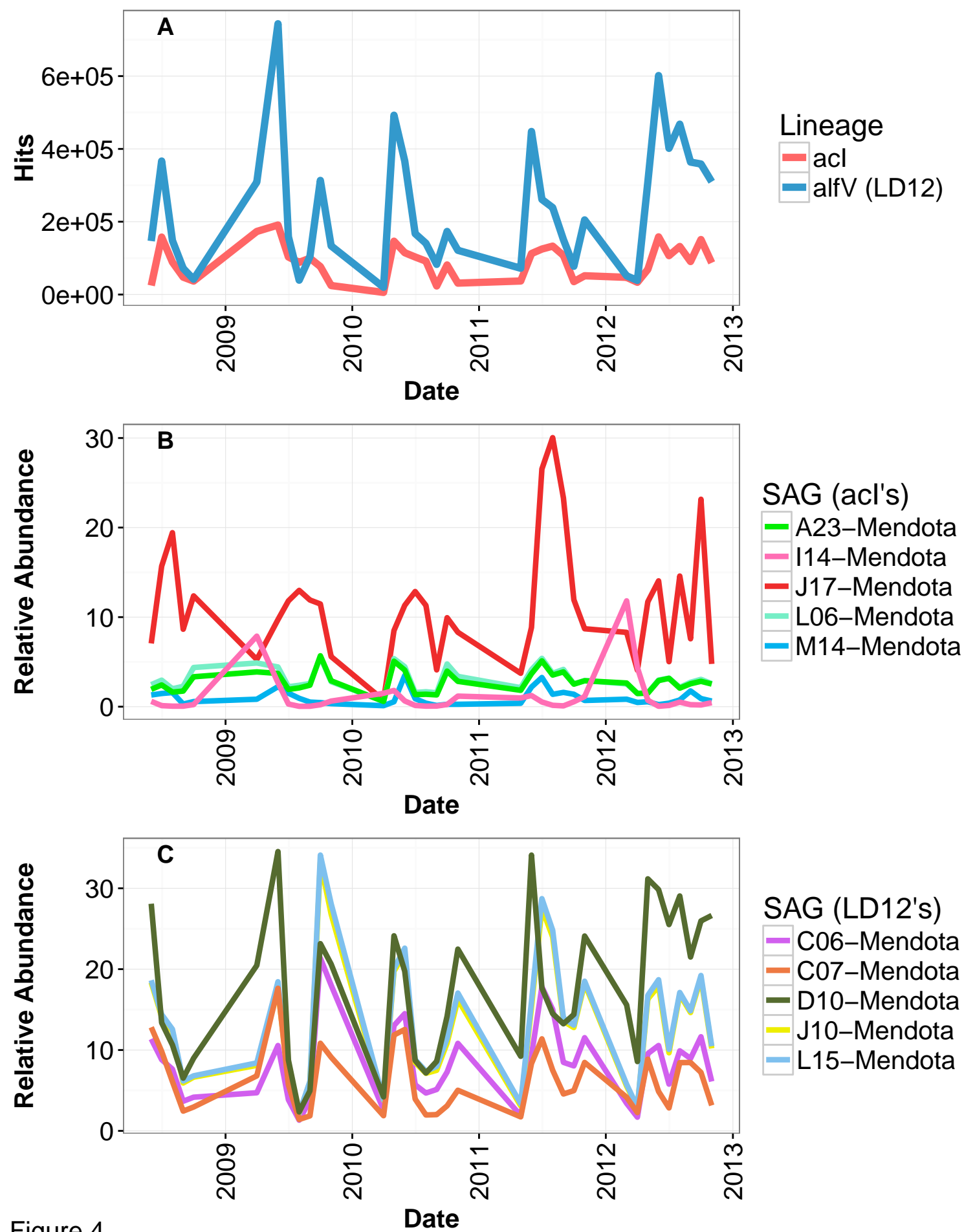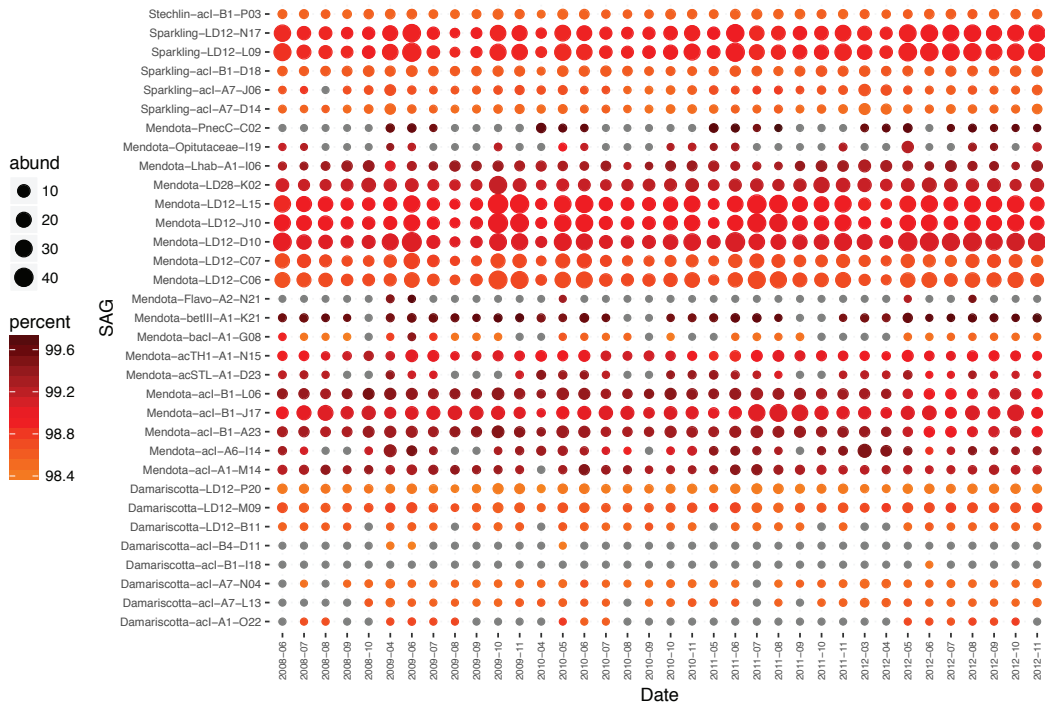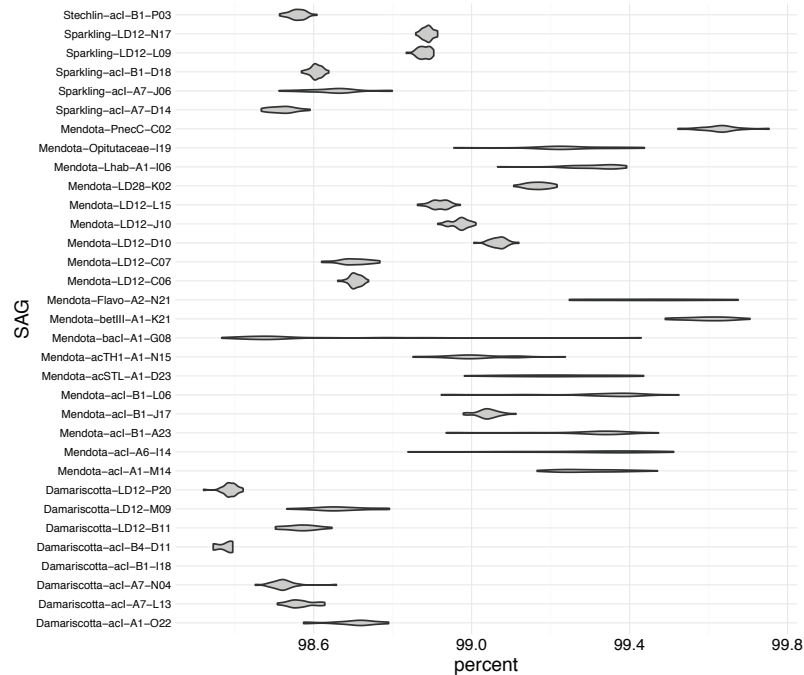
Figure 4

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 | acl-A1 | Damariscotta | 09/18/09 | 1.14 | 74.4 |
| AAA027-M14 | Actinobacteria | acl-A1 | Mendota | 12/5/09 | 0.82 | 43.1 |
| AAA028-I14 | Actinobacteria | acl-A6 | Mendota | 12/5/09 | 0.78 | 39.66 |
| AAA041-L13 | Actinobacteria | acl-A7 | Damariscotta | 04/28/09 | 1.38 | 74.14 |
| AAA044-N04 | Actinobacteria | acl-A7 | Damariscotta | 04/28/09 | 1.29 | 79.59 |
| AAA023-J06 | Actinobacteria | acl-A7 | Sparkling | 05/28/09 | 0.70 | 34.48 |
| AAA024-D14 | Actinobacteria | acl-A7 | Sparkling | 05/28/09 | 0.78 | 48.4 |
| AAA278-I18 | Actinobacteria | acl-B1 | Damariscotta | 09/18/09 | 0.94 | 63.73 |
| AAA027-J17 | Actinobacteria | acl-B1 | Mendota | 12/5/09 | 0.97 | 65.26 |
| AAA027-L06 | Actinobacteria | acl-B1 | Mendota | 12/5/09 | 1.16 | 76.59 |
| AAA028-A23 | Actinobacteria | acl-B1 | Mendota | 12/5/09 | 0.83 | 57.56 |
| AAA023-D18 | Actinobacteria | acl-B1 | Sparkling | 05/28/09 | 0.75 | 44.22 |
| AB141-P03 | Actinobacteria | acl-B1 | Stechlin | 05/25/10 | 0.66 | 45.98 |
| AAA044-D11 | Actinobacteria | acl-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 | bacl-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 |