Freshwater bacterial lifestyles inferred from comparative genomics

Joshua A. Livermore,1 Scott J. Emrich,1,2 John Tan3,4 and Stuart E. Jones1,4*

1Notre Dame Environmental Change Initiative, University of Notre Dame, 100 Galvin Life Sciences, Notre Dame, IN 46556, USA.
2Department of Computer Science and Engineering, University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46556, USA.
3Eck Institute for Global Health, University of Notre Dame, 107 Galvin Life Sciences, Notre Dame, IN 46556, USA.
4Department of Biological Sciences, University of Notre Dame, 264 Galvin Life Sciences, Notre Dame, IN 46556, USA.

Summary

While micro-organisms actively mediate and participate in freshwater ecosystem services, we know little about freshwater microbial genetic diversity. Genome sequences are available for many bacteria from the human microbiome and the ocean (over 800 and 200, respectively), but only two freshwater genomes are currently available: the streamlined genomes of Polynucleobacter necessarius ssp. asymbioticus and the Actinobacterium AcI-B1. Here, we sequenced and analysed draft genomes of eight phylogenetically diverse freshwater bacteria exhibiting a range of lifestyle characteristics. Comparative genomics of these bacteria reveals putative freshwater bacterial lifestyles based on differences in predicted growth rate, capability to respond to environmental stimuli and diversity of useable carbon substrates. Our conceptual model based on these genomic characteristics provides a foundation on which further ecophysiological and genomic studies can be built. In addition, these genomes greatly expand the diversity of existing genomic context for future studies on the ecology and genetics of freshwater bacteria.

Introduction

Human survival and prosperity require freshwater for basic functions including drinking, agriculture, recreation and sanitation, but freshwater availability is threatened by overuse and pollution. While micro-organisms underpin the majority of these ecosystem services, we know little about freshwater microbial genetic diversity (Reid et al., 2005). Genome sequences of over 200 marine bacteria (Yooseph et al., 2010) have enabled great advances in our understanding of the functional capacity and ecology of marine bacteria, including the roles they play in nutrient recycling and carbon fixation among others (DeLong and Karl, 2005). Additionally, the human microbiome project has revealed much about genetic diversity of human symbionts generating greater than 800 reference bacterial genomes with a goal of 3000 (Methé et al., 2012). In stark contrast, the only lake bacteria with published genome sequences are Polynucleobacter necessarius ssp. asymbioticus (Pnec; Newton et al., 2011; Hahn et al., 2012) and the Actinobacterium AcI-B1 (Garcia et al., 2012).

Polynucleobacter has been cultured from virtually every type of freshwater habitat (Hahn, 2003) with abundance ranging from the detection limit to 67% of bacterioplankton cells, and they, on average, comprise approximately 20% of all cells in any given freshwater body (Jezberová et al., 2010). Recent work indicates that one strain of this broader taxonomic group is indeed passive (no known genomic mechanisms for response to the environment) and highly substrate specialized with a small streamlined genome (2.1 Mb) and no obvious genetic markers for motility or cell-cell signalling (Hahn et al., 2012). Similar to Pnec, the Actinobacteria tribe AcI-B1 is also highly abundant in lakes (up to 70% of all cells; Warnecke et al., 2005) with evidence of ubiquity by diversification through niche partitioning (Newton et al., 2007). Additionally, recent genome analysis of a single amplified genome from tribe AcI-B1 suggests a passive specialized lifestyle with a streamlined genome (Garcia et al., 2012).

Based on these two genome sequences, one might assume all lake bacteria possess streamlined genomes and tend to have a passive and planktonic lifestyle. However, spatial and temporal surveys (Newton et al., 2006; Eiler and Bertilsson, 2007) as well as ecological theory would suggest that a single lifestyle strategy for all
Diverse freshwater bacteria genomes

<table>
<thead>
<tr>
<th>Freshwater isolate</th>
<th>Class</th>
<th>Most specific classification level</th>
<th>Classification</th>
<th>Predicted characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLX17</td>
<td>Actinobacteria</td>
<td>Tribe</td>
<td>acTH2 Myco</td>
<td>Closely affiliated phylogenetically with Mycobacterium. Known clones are from Lake Taihu in China.</td>
</tr>
<tr>
<td>L41A</td>
<td>Alphaproteobacteria</td>
<td>Tribe</td>
<td>alfI Brev</td>
<td>Associated with promotion or inhibition of Cyanobacteria</td>
</tr>
<tr>
<td>LLX12A</td>
<td>Alphaproteobacteria</td>
<td>Tribe</td>
<td>alfIV alfIV-B M-L-85</td>
<td>Predicted capacity to utilize humic substance or other difficult compounds based on isolation from humic lakes using phenol enrichment.</td>
</tr>
<tr>
<td>WG36</td>
<td>Gammaproteobacteria</td>
<td>Clade</td>
<td>gamII gamII-A</td>
<td>Potentially abundant in lakes, but also associated with terrestrial animals in databases</td>
</tr>
<tr>
<td>L18</td>
<td>Gammaproteobacteria</td>
<td>Tribe</td>
<td>gamIV gamIV-A</td>
<td>Potentially abundant in lakes, but also associated with terrestrial animals in databases</td>
</tr>
<tr>
<td>WG21</td>
<td>Flavobacteria</td>
<td>Clade</td>
<td>bacII bacII-A</td>
<td>Flavobacteria like lineage enriched during periods of high-dissolved organic carbon</td>
</tr>
<tr>
<td>FWI2</td>
<td>Betaproteobacteria</td>
<td>Lineage</td>
<td>betII (Pnec)</td>
<td>Cosmopolitan freshwater group Best BLAST hits are to organisms in genus Vogesella, which are commonly isolated from freshwater bodies.</td>
</tr>
<tr>
<td>L13</td>
<td>Betaproteobacteria</td>
<td>Subphylum</td>
<td>Betaproteobacteria</td>
<td></td>
</tr>
</tbody>
</table>

Classifications determined by position of 16S rRNA gene sequences in a RAxML maximum likelihood phylogeny with query sequences aligned to a hand-curated reference freshwater 16S rRNA gene alignment using the Silva incremental aligner (Sina). Where available, predicted characteristics are from summaries in Newton et al. 2011.

Freshwater bacteria is highly unlikely. For example, Bacteroidetes taxa can numerically dominate lakes following an influx of dissolved organic carbon, many Gammaproteobacteria use freshwater bodies as a transmission vector and freshwater Verrucomicrobia are globally pervasive, although scarce in any given water body. We also know that much of the bacterial diversity in lakes lies dormant, roused by an influx of resources or shifts in other environmental features (Jones and Lennon, 2010).

To expand our understanding of freshwater bacterial lifestyles and their genetic underpinnings, we sequenced the genomes of eight lake bacterial isolates. To maximize capture of novel freshwater genetic diversity, we selected a phylogenetically diverse set of previously isolated strains. We contrasted our eight draft genome sequences with the Pnec genome and the single-cell amplified genome of Aci-B1. In addition, we compare these 10 freshwater bacterial genomes with phylogenetically similar bacterial genome sequences from soil and marine environments. We conclude that distribution of genomic characteristics among these freshwater isolates suggests a trade-off between carbon substrate diversity and growth rate as isolates with augmented substrate capabilities exhibited slow-predicted growth rate, whereas fast-growing isolates had reduced substrate utilization capacity. Both of these strategies contrasted with the passive, slow and specialist habits of the previously sequenced Pnec and Aci-B1 genomes.

Cross-ecosystem comparison of genome content

We compared subsystem category distributions [assigned by Rapid Annotation using Subsystem Technology (RAST)] of freshwater bacteria genomes with close phylogenetic relatives from marine and soil environments. Principal components analysis of the full set of 30 genomes subsystem categories indicated separation of soil and marine environments (freshwater overlapping) in the first dimension (28% of variance) and taxonomy in the second (14% of variance) illustrated as plots of 95% confidence interval ellipses based on environment (Fig. 1A) and class (Fig. 1B). The RAST subsystem that contributed most to PC 1 was category Protein metabolism (Subsystem and Subsystem category names are noted with italics) with marine and freshwater bacteria having a higher rank (lower numbers indicate higher rank) in this category than the soil bacteria (average ranks 4.3 ± 2.0, 4.4 ± 2.2, and 5.3 ± 1.1, respectively; Fig. 1C). DNA metabolism, Cell division and cell cycle and Nucleoside and nucleotides were
Fig. 1. Summary of protein subsystem categories from RAST annotations of genomes from 10 freshwater bacterial genomes with phylogenetic pairings of both soil and marine bacteria.
A. The first two axes of a principal component analysis (PCA) on subsystem category proportions with ellipses on 95% confidence intervals of environment classifications (proportion of variance: PCA 1- 0.28, PCA 2- 0.14).
B. The same PCA with ellipses on class taxonomic level.
C. Rank of subsystem categories each genome arranged by a 16S rRNA gene PhyML phylogeny (GTR model). Solid dots on a dendrogram branch indicate greater than 90% approximate likelihood ratio test support, open dots indicate greater than 70% support, and remaining values are shown. The most abundant categories were assigned value ‘1’ to least abundant assigned ‘25’. Eigenvalues of first two principal component axes are also shown.

also strongly correlated to the first principal component. However, there were no compositional differences between the three environments for subsystems in these categories, indicating that within our genome set, marine bacteria allocate a greater overall proportion of genomic space to Protein metabolism, DNA metabolism, Cell division and cell cycle and Nucleosides and nucleotide, but not necessarily to any specific process or processes therein. Correlated to PC 1 in the opposite direction were Sulfur metabolism, Regulation and cell signalling and Membrane transport, indicating greater importance of these categories in soil bacteria as a whole and also some freshwater bacteria. The greatest subsystem category contributor to the second principal component is Cell wall and capsule with Actinobacteria and Alphaproteobacteria allocating proportionally less than other classes (average ranks Actinobacteria-12.1 ± 2.4, Alphaproteobacteria-11.3 ± 2.1, Betaproteobacteria-8.9 ± 3.3, Flavobacteria-7.0 ± 0.8, Gammaproteobacteria-7.3 ± 1.3) but again without evident compositional differences within the

© 2013 Society for Applied Microbiology and John Wiley & Sons Ltd, Environmental Microbiology, 16, 746–758
category. RNA metabolism was also correlated with PC 2 and Fatty acids, lipids and isoprenoids were negatively correlated.

Subsystem assignments of freshwater bacterial open reading frames

Most freshwater genomes indicate heavy investment in subsystem category Amino acids and derivatives, followed by Carbohydrates; Cofactors, vitamins, prosthetic groups and pigments; and Protein metabolism (Fig. 1C). All of the freshwater bacteria (and 29 of 30 total taxa examined), except Pnec, have Amino acids and derivatives or Carbohydrates as the highest-ranked category, and five of the freshwater genomes have both as the top 2 (24 of 30 total taxa examined, Fig. 1C). Polynucleobacter necessarius ssp. asymbioticus has lower investment in Amino acids and derivatives (ranked 3rd) and Carbohydrates (ranked 5th) with proportionally higher investment in Protein metabolism and Cofactors, vitamins, prosthetic groups and pigments (tied for 1st).

Closely examining Acl-B1 and Pnec proportional subsystem category investment, it is clear that both invest outside the range exhibited by the eight new genomes (Fig. 2). While the freshwater isolates we sequenced uniformly allocate about 14 ± 1% of genome space to category Amino acids and derivatives, Pnec and Acl-B1 allocate 11% and 20%, respectively. In contrast, Pnec invests relatively heavily in Cofactors, vitamins, prosthetic groups and pigments (11%), while Acl-B1 has low outlying investment in this category (6%; average for all freshwater genomes is 9 ± 1.7%). Both Acl-B1 and Pnec have disproportionately high allocation in Protein metabolism (12% and 11%, respectively, versus 7% ± 2.4% average) and in Respiration (5.4% each versus 4 ± 0.8% average in the other eight genomes). Polynucleobacter necessarius ssp. asymbioticus and Acl-B1 tended to have disproportionately high or low allocation to the majority of subsystems.

Regarding carbohydrate pathways, genome content for carbohydrate usage was highly variable among the genomes, although several metabolic systems are conserved throughout [e.g. Methylglyoxal metabolism, Pyruvate metabolism, Glycolysis and gluconeogenesis and TCA cycle (Tricarboxylic acid cycle)]. Also interesting are the substrate markers present in all but one or two genomes. FIGfams (the finest unit in the RAST hierarchy specifying individual homologous proteins) associated with the Serine-glyoxylate cycle are present in all but two of the genomes (24 or more FIGfams) the exceptions being FWI2 and L41A. Another example is Glycerate metabolism absent only in WG21. The isolates fell into three levels of potential for metabolism of aromatic compounds with FWI2 and L18 having large number of FIGfams (145 and 172, respectively), WG36 and LLX12A having an intermediate number (46 and 64), and the remainder having less than 20 FIGfams in the category.

Carbohydrate and Metabolism of aromatic compound subsystem categories likely serve as important indicators
of lifestyle for heterotrophic bacteria and an ordered presence-absence plot of those subsystems indicates two different strategies (Fig. 3). All the bacteria share a core set of substrates and processes for which each is present in at least 7 of the 10 genomes. Bacteria with small genomes (< 4 Mb; L13, LLX17, L41A and Pnec) have very few additional substrates and processes supplementing their core (5 ± 1 additional subsystems). In contrast, the large genomes (> 5 Mb; FWI2, WG21, LLX12A, L18 and WG36) each carry a large and unique substrate complement (24 ± 7 additional subsystems).

Examination of subsystem categories for use of other elements (S, N, P and Fe) reveals further variation in lifestyle characteristics (Table S2). In the Sulfur metabo-
lism category, 7 of 10 genomes have open reading frames (ORFs) classified as inorganic sulfur-assimilation genes (excluding Pnec, Acl-B1 and LLX12A). In contrast, the only two genomes with classified ORFs for Sulfur oxidation were Pnec (18 FIGfams) and FWI2 (9 FIGfams). In the Nitrogen metabolism category, all but two genomes had ORFs for both Nitrate/nitrite utilization and Ammonia assimilation the exceptions being L41A and Acl-B1, which lack FIGfams for use of nitrate/nitrite. Additionally, genomes for FWI2 and L13 had apparent ORFs for denitrification (13 and 8 FIGfams, respectively). For Phosphate metabolism, genome WG21 lacked High affinity phosphate transporter FIGfams, which were found in the nine other genomes. FWI2, L18 and WG36 had greater than nine FIGfams for Alkylphosphonate utilization or Phosphonate metabolism, while L41A, LLX17 and Pnec each had none and the remaining three genomes had four or less FIGfams. Iron acquisition and metabolism FIGfams were concentrated most heavily in FWI2, L18, WG36 and L13 (38, 70, 36 and 27 FIGfams, respectively) with remaining genomes harbouring eight or less. WG36 had FIGfams for Aerobactin and Enterobactin siderophores, while L18 had Enterobactin only. It is important to note that while assertions of presence and absence are tenuous with draft genomes, a survey of 31 marker genes (Wu and Scott, 2012) yields full recovery in six of the eight genomes and greater than 94% in the other two (30 in WG21 and 29 in L13). Therefore, it is reasonable to make inferences based on gene presence and absence in these genomes. ORFs associated with motility (LLX12A, L41A, L18, L13 and FWI2), while the remainder have less than 17 (two in Acl-B1 and one in Pnec). Polynucleobacter necessarius ssp. asymbioticus also possessed the second longest predicted minimum generation time of all the genomes, exceeded only by FWI2 (which clusters with the Pnec phylogenetically). The predicted fastest growing bacteria were L18 and WG36, which are both Gammaproteobacteria.

Discussion
We sequenced eight phylogenetically diverse isolates from freshwater lakes to contribute context and examine bacteria spanning a wide spectrum of ecophysiological lifestyles. Polynucleobacter necessarius ssp. asymbioticus and Acl-B1 epitomize the passive, oligotrophic specialist lifestyle (slow growth, small genome, non-motile and low signal transduction (Garcia et al., 2012; Hahn et al., 2012), while our newly sequenced genomes represent a spectrum of predicted lifestyle characteristics.

Comparison of freshwater genomes with those from soil and marine environments
Comparison of subsystem category distributions suggests that the average lifestyle of a freshwater bacterium is somewhere between a soil and marine bacterium. Specifically, freshwater genomes significantly overlap with the relatively distinct soil and marine genomes in the first two axes of a principal component analysis based on subsystem category distributions. In some sense, this observation is intuitive, given that soil and marine environments represent two extremes of a hydrological continuum, while freshwater lakes are one intermediate step between them as run-off from the land accumulates in lakes before draining to oceans. Taxonomic discrimination at the class level
in the second principal component indicates that phylogeny is somewhat predictive of microbial lifestyle, with properties conserved across the three environments. However, recall that to minimize the effect of phylogenetic variation, we deliberately chose closely related genomes for cross-ecosystem comparison. Therefore, this analysis does not necessarily imply that environmental variation plays a greater role in determining gene content than phylogeny. Indeed, expected variation in a randomly selected set of genomes would likely be dominated by phylogenetic differences irrespective of environmental origin.

Correspondence between previously predicted characteristics and genome content

Having determined the position of each of these eight genomes within a freshwater taxonomy (Newton et al., 2011), we can compare previously hypothesized characteristics of the group, based on occurrence patterns in temporal and spatial surveys, with the genomic characters observed here.

LLX17 is an Actinobacteria belonging to lineage ActIV-B M–L–85 of which members have been isolated under phenol enrichment suggesting a lifestyle based on degrading humic substances or other organic compounds (Hutalle-Schmelzer et al., 2010). This is supported by subsystem assignments of ORFs in LLX12A indicating many FIGfams unique to LLX12A are categorized as involved with aromatic compound degradation (Table S1). The other Alphaproteobacteria, L41A, is in clade alfII Brev (Brevundimonas in Linnaean taxonomy) and are associated with cyanobacterial blooms (Berg et al., 2008). This agrees with the observation that most of the carbohydrate utilization subsystems in L41A involve small carbon compounds that might be associated with algae.

Representing the Flavobacteria, isolate WG21 is in lineage BacII clade A, which are suspected as organic scavengers and degraders of polymers like cellulose and chitin (Kirchman, 2002) and sometimes dominate lakes following a period of high productivity (Newton et al., 2011). Concordant with predictions based on phylogeny, substrate utilization characteristics of WG21 indicate Chitin and N-acetylglucosamine utilization subsystems as well as Cellulosomes and a large range of other carbohydrates and aromatic compound utilization subsystems (Fig. 3 and Table S1).

The other genome with predicted Chitin and N-acetylglucosamine function is L13, which is a Betaproteobacterium. Our phylogeny was unable to resolve L13 any better than ‘Betaproteobacteria’, but a Basic Local Alignment Search Tool (BLAST) search of the L13 16S rRNA gene against the National Center for Biotechnology Information (NCBI) 16S database indicates that it is closely related to species in the freshwater Vogesella genus, which are known to utilize chitin (Jørgensen et al., 2010). The other Betaproteobacterium is FWI2 clustering with BetII, which is the cosmopolitan Pnec clade (named after Polynucleobacter). In agreement with this relationship, FWI2 shares several metabolic capabilities with Pnec not found in any of the other genomes examined, including FIGfams for Sulfur oxidation, Cyanate hydrolysis and a Siderophore assembly kit.

Of the two Gammaproteobacteria, WG36 clusters in GamII-A (unclear if A1 or A2), while L18 is in lineage GamIV clade A tribe Pseudo A1 (Pseudomonas). While found in lakes, both of these organisms are also at times associated with a terrestrial environment (Newton et al., 2011). Both of these genomes (and none of the others) have Carbon storage regulator FIGfams that are associated with infection and suggests a host component in their life cycle. (Sze and Li, 2011). However, they also share the unique functionality (among the genomes analysed) to utilize the plant-associated compounds xylose and arabinose, which also indicates capacity to metabolize carbon compounds from leaf litter that may accumulate in lakes.

Genome characteristic distributions

Carbon metabolism. Based on substrate utilization pathway distributions (Fig. 3 and Table S1), there appear to be three general categories of organisms in this study. Both Acl-B1 and Pnec are highly reduced and have extremely limited substrate range, while the next set (L41A, LLX17 and L13) are also reduced but have ancillary metabolic capabilities. All five of these organisms have small genomes (< 4 Mb). Members of the final group (WG21, WG36, LLX12A, FWI2 and L18) have larger genomes (> 5 Mb) with a core carbon utilization set augmented by a unique set of substrate utilization FIGfams, often including larger carbon compounds like polymers and aromatics. Predictably, a number of pathways are highly conserved among the 10 freshwater genomes, such as the TCA cycle, pyruvate metabolism and subsystems for utilization of small simple compounds such as lactate and glycerol (near the top of Fig. 3). However, all 10 of the genomes we consider also have genes for utilization of salicylate, quinate and gentisate, which are compounds derived from terrestrial plant matter. Conserved terrestrial carbon pathways suggest the importance of carbon availability in freshwater systems where leaf litter and other
decaying terrestrial plant matter is a consistently available carbon source.

**Environmental response.** In general, the smaller genomes tend to be associated with lower proportion of ORFs associated with signal transduction ($R^2 = 0.5$, $P < 0.03$) and three of the five appear to be largely non-motile (Fig. 4). In contrast, the augmented organisms have genomes greater than 5 Mb, and three of five appear to be motile. Number of membrane-transport genes detected correlates with genome size ($R^2 = 0.62$, $P < 0.007$), but this may simply be a genome size effect as it is not related to proportion of genomic-membrane transport investment ($R^2 = 0.0081$, $P > 0.8$). This may indicate that reduction in number of membrane transporters is proportional to total reduction in genome size. In other words, membrane transport proteins do not appear to be selectively retained or lost during genome streamlining. In contrast, signal transduction proteins disappear at a disproportionately high rate compared with the genome as a whole (i.e. larger genomes have greater proportion of ORFs in signal transduction as described earlier).

While we do not fully understand the process of genome ‘streamlining’, observations of distinct subsystem category distributions in AcI-B1 and Pnec indicate that the process may not be uniform. For example, both AcI-B1 and Pnec have disproportionately high investment in subsystem category Protein metabolism suggesting that reduction in this category may always pose a penalty to fitness. In contrast, in category Amino acids and derivatives, Pnec has heavily reduced investment and AcI-B1 has disproportionately high investment. Additionally, in category Carbohydrates, Pnec has heavily reduced investment, while AcI-B1 maintains investment comparable with other taxa. Such discrepancies in how these two genomes appear to have streamlined may indicate fundamental underlying differences in their lifestyles or perhaps simply different random paths taken in discarding superfluous genes. It may also be that Pnec simply has not fully streamlined to the extent of organisms like AcI-B1 or the marine *Pelagibacter ubique* (Pnec genome size 2 Mb compared with 1.2 Mb in AcI-B1 and 1.3 Mb in *Pelagibacter ubique*), which have the smallest genomes of any known free-living bacteria.

**Genome-informed lifestyles of freshwater bacteria**

An advantage of analysing the genomes of a diverse set of freshwater bacteria is the opportunity to generate hypotheses on the spectrum of lifestyles suggested by their genome composition. Based on the genomes presented here, we conceptualize freshwater organisms as utilizing a number of discrete lifestyle strategies based on a three-axis continuum with the first axis defined by predicted growth rate (estimated from codon usage bias, see Methods), the second axis by the number of identified carbon substrate utilization genes, which is correlated with utilization of larger carbon compounds/aromatics and genome size, and the third axis defined by the proportion of a genome coding for motility and signal transduction, as surrogate measures for environmental responsiveness (Fig. 5). Based on these semimetric axes, there appear to be four general lifestyle strategies that we call ‘passive and streamlined’, ‘slow, augmented responders’, ‘fast, reduced responders’ and ‘vagabonds’.

The streamlined passive group includes the cosmopolitan Pnec and AcI-B1. These might also be called ‘oligotrophs’ and appear near the origin of the graph with slow predicted growth rate (validated in Pnec culture; Hahn, 2003), very small genomes (2.0 and 1.2 Mb), few genes predicted for utilization of carbon substrates, little to no genomic evidence of motility and few genes with predicted signal transduction function. We also know that Pnec and AcI-B1 have very small cell volumes (Hahn, 2003; Hahn et al., 2003). Organisms conforming to such a lifestyle strategy might be thought of as survivalists persisting by streamlining their genomes (Lynch, 2006) and eliminating extraneous cell surface receptors to minimize phage-attack concomitant with likely loss of transporter function or signal transduction capacity (Bohannan and Lenski, 2000). Reduction of cell size and overall activity is also known to reduce predation risk from ciliate and flagellate grazers (Cole, 1999; Pernthaler et al., 2001; Pernthaler, 2005). Summarily, organisms in this category would passively persist in the water column growing slowly by gleaning a single or few carbon substrates while maintaining characteristics to minimize predation.

The remaining groups of organisms might be characterized as copiotrophic. These organisms generally use a greater variety of substrates and respond to environmental changes. The first of these groups we call the slow, augmented responders. Members of this group include FWI2, LLX12A and WG21, which all have larger genomes (7.2, 6.0 and 5.2 Mb, respectively) and higher diversity of carbon substrate utilization pathways but low growth rates. Additionally, all three have characteristics indicating active response to their environment including genes for motility and signal transduction. The second group, we call the fast, reduced responders that includes the organisms L13, L41A and LLX17. These organisms have intermediate genome sizes (~3.8, 3.1 and 3.0 Mb, respectively) and substantially fewer genes predicted for carbon substrate utilization but higher predicted growth rates. They also appear to sense and respond to their environment evidenced by motility and signal transduction markers. Finally, each of these three (L13, L41A and LLX17) has below average investment in the Membrane transport subsystem.
category. The distinction between the slower growing augmented group and the faster growing reduced group suggests a trade-off between substrate flexibility and maximum growth rate among these more environmentally responsive organisms. It appears the fast, reduced responders prefer a small set of substrates and either delay growth until their preferred substrate is available (Jones and Lennon, 2010) or they actively seek specific resources chemotactically. In exchange for this reduction, genomic evidence suggests that they grow faster. In contrast, the slow, augmented responders have the benefit of being able to use a large and diverse array of substrates but suffer lower growth rate either as a consequence of their larger genomes or some other burden related to an excess of metabolic capacity. While genome size and predicted growth rate are not correlated at the course scale (Mira et al., 2001), such a relationship may hold when examining only free-living copiotrophs as observed here (R = 0.87, P = 0.02 for these six genomes). It is also important to recognize the distinction between genomically predicted growth rates, which are a reflection of evolutionary history of growth habit, and measured growth rates, which describe propensity to grow in a defined liquid medium. These two measures are not always perfectly correlated (Fig. S2), and one or the other may prove a more reliable indicator of lifestyle.

Our distinction between responders and passive streamlined organisms can serve as an explanation for differing adaptations to temporal variation in resource availability but is easily extended to spatial variation. Recent reviews on bacterial adaptation to spatial nutrient heterogeneity in aquatic environments (Stocker, 2012; Grossart, 2010) distinguish between organisms that chemotactically respond to the nutrient gradients and hot spots intrinsic to all aquatic ecosystems and those that persist passively on the constant background concentrations of resources. These two lifestyles correspond directly to what we referred to as responders and passive streamlined, respectively.

The final group observed among available freshwater genomes we call ‘vagabonds’ represented by WG36 and L18 (both Gammaproteobacteria). A review on freshwater bacterial taxonomy (Newton et al., 2011) uses the term vagabond for those bacteria that appear in lakes but also may have a wider distribution in the biosphere. In other words, while these organisms are probably active freshwater inhabitants (Nercessian et al., 2005), a free-living lifestyle in freshwater may be only a portion of their life
cycle. The two representatives we identified here both possess extremely high-predicted growth rates relative to the other genomes, show indications of motility and appear to use a large number of complex carbon substrates paradoxical to the paradigm described earlier (Fig. 5). This would make sense if these types of organisms use freshwater bodies primarily as a maintenance and transmission vector but propagate primarily in a host or different freshwater-associated environment.

In examining the variation in lifestyle characteristics, it is worth noting the influence of isolation method used to obtain a genome from an organism of a given lifestyle type. The AcI-B1 genome was obtained from the amplified genome of a single cell (Garcia et al., 2012), while Pnec was isolated using a filtration-acclimatization method (Hahn et al., 2003). In contrast, the organisms presented here, with the exception of FWI2, were isolated and selected based on their propensity to grow on culture plates with relatively rich medium [Wright’s Chu #10 (WC) medium amended with thiamine, biotin and inorganic phosphate.]. We preferentially selected for more copiotrophic organisms by using traditional culturing techniques as opposed to the more laborious filtration-acclimatization procedures or culture-independent approaches, such as single-cell genome amplification. Passive streamlined specialists like AcI-B1 and Pnec were selected as freshwater models because they are cosmopolitan and frequently have high cell abundances. However, our genomic comparisons suggest this abundance may be a product of diminished growth and activity (non-motile, long-predicted minimum generation time, few substrate propensities). In contrast, the population dynamics of rarer members of the freshwater community, selected for by traditional culturing techniques (Shade et al., 2012), are likely characterized by punctuated periods of high growth and activity. These distinct lifestyle strategies (consistent, slow growth versus short-lived periods of rapid growth) may result in equivalent contributions to important ecosystem processes, such as nutrient recycling and carbon processing.

In summary, by examining a diverse group freshwater genomes, we are beginning to identify contrasting genomic characteristics that are indicative of ecologically relevant lifestyle strategies. This framework will be further strengthened by field surveys and experimentation validating the lifestyle inferences gained from genome characteristics. As evidence mounts, such a paradigm as presented here can guide analysis and classification of other organisms based on their genome sequences and also informs interpretation of environmental shotgun metagenomic surveys. Finally, the availability of genomes from a variety of native lake organisms will be beneficial for contextual analysis of new and novel genes and gene pathways identified in freshwater environments.

**Methods**

Eight freshwater bacteria previously isolated from Little Long Lake, Wintergreen Lake (MI, USA), or Pleasant Lake (IN, USA) were selected for genome sequencing based on taxonomic classification inferred from partial 16S rRNA gene sequences. Seven of the eight were isolated from Little Long and Wintergreen (five and two isolates, respectively) using a modified WC minimal medium (Stemberger, 1981 with 4.2 mM of glycine, 5.3 mM of acetate, 5.2 mM of dextrose and 3.5 mM of succinate). Additional isolation information for these seven is available in (Bird, 2012). The eighth isolate originates from Pleasant Lake and was isolated using the filtration-acclimatization technique previously used during the isolation of Pnec except that humic acids were added to the inorganic basal medium instead of nutrient broth, soytone and yeast extract (Hahn, 2003). Bacteria for sequencing were grown in R2A medium in 200 ml glass serum bottles on a shaker table. Once cultures reached early stationary phase, we reduced cell density to an OD600 (optical density measured at 600 nm) of 1.2 and isolated DNA with a cetyltrimethyl ammonium bromide and phenol chloroform total nucleic acid extraction followed by an RNAse treatment. DNA was quantified using a Qubit fluorometer (Invitrogen, Life Technologies, Grand Island, New York, USA).

DNA samples were prepared for sequencing adhering to the Roche 454 titanium chemistry protocol for rapid library preparation. Approximately 500 ng of genomic DNA was nebulized, end-repaired and ligated to 454 adaptors (containing barcode sequences for multiplexing) to generate the sequencing library. Sequencing libraries were attached to DNA capture beads, titrated and amplified through emulsion polymerase chain reaction (emPCR). The beads from the emPCR reaction were enriched, loaded onto a 454 sequencing plate and sequenced on a 454 GS-FLX sequencer (454 Life Sciences, Branford, CT, USA).

We also sequenced 3 kb paired-end libraries for two isolate genomes (L18, LLX17). Approximately 5 µg of genomic DNA were fragmented on a Covaris S220 high-intensity acoustic transducer (Covaris, Inc., Woburn, MA, USA), end-polished, ligated to a circularization adaptor, size-selected for 3 kb fragments, then circularized. After circularization, the protocol follows the general library preparation and sequencing steps as discussed earlier.

Demultiplexing of FLX titanium sequence data for each genome was performed using standard 454 software [standard flowgram format (SFF) file] and barcodes. Next, sequence data present in the resulting individual flowgram (SFF) files were assembled using 454 Newbler version 2.5.3 (454 Life Sciences) using multiple processors and default parameters. Newbler was used to estimate the empirical size of any mate-pair libraries present in SFF.
data during the assembly process and subsequent scaffolding. All analysis was performed on contigs or scaffolds where available. Assembly statistics are in Table S1 and draft genome assemblies can be accessed from the NCBI Bioproject repository (Accs. PRJNA174161, PRJNA174163, PRJNA174212 and PRJNA174214-PRJNA174218).

The complete nucleotide genome sequence of Polynucleobacter necessarius ssp. asymbioticus QLW-P1DMWA-1 (Meincke et al., 2012), hereafter referred to as ‘Pnec’, was obtained from GenBank (Acc. NC_009379.1). The nucleotide contig sequences of the single-amplified genome Actinobacteria Acl-B1 SCGC AAA027-L06 (hereafter, Acl-B1) (Garcia et al., 2012) were kindly provided by Katherine McMahon of the University of Wisconsin-Madison. ORF finding and translation for Pnec, Acl-B1 and all of the sequences collected in this study were conducted using Glimmer3 (Delcher et al., 2007) and a custom Python script.

To compare genomes of freshwater isolates with isolates from other environments while controlling for phylogenetic differences, we collected the phylogenetically closest available genome from both soil and marine environments. Phylogenetic pairings were identified with a RAxML (Stamatakis et al., 2008) maximum likelihood phylogeny of full-length 16S rRNA gene alignment from publically available genomes. We selected the closest soil and marine genome to each freshwater isolate.

To classify the freshwater genomes in an existing freshwater bacteria phylogenetic classification system (Newton et al., 2011), we identified 16S rRNA genes from the genome using RNAMMER (Lagesen et al., 2007). Resultant sequences (a single gene was identified for each genome) were aligned to a full-length 16S rRNA gene freshwater bacteria reference alignment (Newton et al., 2011) using the Silva incremental aligner (Sina) (Pruesse et al., 2012). Phylogenetic classifications were defined by positions of isolates in phylum-specific RAxML maximum likelihood phylogenies.

To holistically compare freshwater genomes with each other and soil/marine genomes we used the RAST pipeline [settings: ORFs were identified by Glimmer3, FIGfam release 45, all optional steps disabled (Aziz et al., 2008)] in which genes are assigned to orthologous groups called ‘FIGfams’ which are assigned functional categories called ‘subsystems’. We computed principal components on proportional subsystem category distributions using Vegan (Oksanen et al., 2007) in R (R Development Core Team, 2010) to identify clustering relationships between genomes and reveal the categories with greatest contribution to inter-genomic variation. We further examined Carbohydrate metabolism and Metabolism of aromatic compounds subsystem categories using Vegan in R to visualize a non-metric dimensionally scaled ordination with Raup–Crick dissimilarity, which is a presence-absence-based measure (Raup and Crick, 1979).

In addition to annotation in RAST, we identified known protein domains from all translated ORFs of the freshwater isolates using the Pfam-A database (Bateman et al., 2004). Annotation was implemented with the ‘pfam_scan.pl’ perl script from the Pfam website (http://pfam.sanger.ac.uk). Annotations permitted identification of marker domains for signal transduction (Ulrich and Zhulin, 2010) and motility (Lingner et al., 2010). Supplementary to signal transduction domain identification, a previously published hidden Markov model (Staron et al., 2009) identified likely ECFs, which are additional markers of signal transduction. Probable membrane transporters were identified from amino acid sequences using TransportTP (Li et al., 2009) with Escherichia coli serotype O157:H7 strain EDL933 as the reference organism and an E-value threshold of 0.1. Finally, minimum generation time of each isolate was estimated based on a previously developed model utilizing codon usage bias as a predictor (Vieira-Silva and Rocha, 2010).

References


**Supporting information**

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

**Fig. S1.** Maximum likelihood phylogeny (PhyML, GTR model, SPR moves) of 16S rRNA genes from eight newly sequenced freshwater bacterial genomes (closed circles), two previously sequenced freshwater bacterial genomes (open circles), and closely related sequences for context. Dot on branches indicate greater than 80% support (aLRT). Key: blue = marine, red = host-associated, green = freshwater, brown = soil.

**Fig. S2.** Correspondence between genome predicted growth rate and observed growth rate is noisy especially for Gammaproteobacteria (L18 and WG36). The line marks one to one correspondence.

**Table S1.** Genome assembly statistics and isolate origins for eight freshwater bacterial genomes.

**Table S2.** Nutrient and carbohydrate FIGfams identified in the genomes of eight freshwater bacteria isolates and two previously sequenced genomes (Pnec and Acl-B1).