Environmental Microbiology (2013) doi:10.1111/1462-2920.12301

Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics

Alexander Eiler,[1*†] Katarzyna Zaremba-Niedzwiedzka,[2†] Manuel Martínez-García,[4,5] Katherine D. McMahon,[3] Ramunas Stepanauskas,[4] Siv G. E. Andersson[2] and Stefan Bertilsson[1]

[1]*Department of Ecology and Genetics, Limnology and* [2]*Department of Cell and Molecular Biology, Molecular Evolution, Uppsala University, Uppsala, Sweden.*
[3]*Departments of Civil and Environmental Engineering, and Bacteriology, University of Wisconsin-Madison, Madison, WI, USA.*
[4]*Single Cell Genomics Center, Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA.*
[5]*Department of Fisheries, Genetics and Microbiology, University of Alicante, Alicante, Spain.*

Summary

Little is known about the diversity and structuring of freshwater microbial communities beyond the patterns revealed by tracing their distribution in the landscape with common taxonomic markers such as the ribosomal RNA. To address this gap in knowledge, metagenomes from temperate lakes were compared to selected marine metagenomes. Taxonomic analyses of rRNA genes in these freshwater metagenomes confirm the previously reported dominance of a limited subset of uncultured lineages of freshwater bacteria, whereas Archaea were rare. Diversification into marine and freshwater microbial lineages was also reflected in phylogenies of functional genes, and there were also significant differences in functional beta-diversity. The pathways and functions that accounted for these differences are involved in osmoregulation, active transport, carbohydrate and amino acid metabolism. Moreover, predicted genes orthologous to active transporters and recalcitrant organic matter degradation were more common in microbial genomes from oligotrophic versus eutrophic lakes. This comparative metagenomic analysis allowed us to formulate a general hypothesis that oceanic- compared with freshwater-dwelling microorganisms, invest more in metabolism of amino acids and that strategies of carbohydrate metabolism differ significantly between marine and freshwater microbial communities.

Received 6 December, 2012; accepted 27 September, 2013.
*For correspondence. E-mail alexander.eiler@ebc.uu.se; Tel. (+46) 18 471 2700; Fax (+46) 18 531134. †These authors contributed equally.

Introduction

Lakes are systems of enhanced biological activity and are central to many biogeochemical processes (Battin *et al.*, 2009; Tranvik *et al.*, 2009). Lakes also represent a critical natural resource for human societies (Downing *et al.*, 2006). Although bacteria are known to perform many critical biogeochemical processes and thus also have the potential to modify and control water quality in these ecosystems, we have limited understanding of their functional potential, genetic variability and community interactions. This is partly because most abundant lake bacteria are notoriously difficult to culture in isolation (Newton *et al.*, 2011). The first sequenced genomes of abundant freshwater bacteria (Garcia *et al.*, 2012; Hahn *et al.*, 2012) and recent metagenomic characterization of microorganisms from Lake Gatun (Rusch *et al.*, 2007), Lac du Bourget (Debroas *et al.*, 2009) and Lake Lanier (Oh *et al.*, 2011) have provided some first snapshots of the functional diversity of freshwater bacterioplankton in single lake ecosystems. These studies have corroborated findings based on 16S rRNA amplicon surveys with regards to the composition of freshwater bacterial communities and the existence of a phylogenetically distinct freshwater microbiota (reviewed in Newton *et al.*, 2011). Nevertheless, because of the often substantial genomic variation among even closely related strains, it is challenging to predict community metabolism solely from taxonomic markers and the often rather limited metabolic and functional information derived from reference isolates.

In contrast with such marker gene approaches, metagenomic analysis has the potential to summarize the combined genetic blueprint of all organisms in a given community (Riesenfeld *et al.*, 2004). By sequencing all genetic information in a community, the relative abundance of all represented genes can, at least in theory, be determined and used to provide a synoptic description of

Table 1. Description of lakes used in this study.

ID	Sample location	Country	Date	Location	Sample depth	T (°C)	Size fraction (μm)	Habitat type	Tot P
DamariscottaSP	Lake Damariscotta	USA	20090528	44°10′n; 69°29′w	0.5–1	12.1	> 0.2	Mesotrophic lake	10
DamariscottaSU	Lake Damariscotta	USA	20090819	44°10′n; 69°29′w	0.5–1	12.1	> 0.2	Mesotrophic lake	10
Ekoln	Lake Ekoln	Sweden	20070731	59°45′n; 17°36′e	0–2	19.0	0.2–100	Eutrophic lake	50
Erken	Lake Erken	Sweden	20070620	59°25′n; 18°15′e	0–2	18.7	0.2–100	Mesotrophic lake	33
Lanier	Lake Lanier	USA	20090827	34°12′n; 83°59′w	0–5	28.5	0.22–1.6	Mesotrophic lake	30
MendotaSP	Lake Mendota	USA	20090512	43° 6′n; 89°24′w	0.5–1	12.68	> 0.2	Eutrophic lake	118
MendotaSU	Lake Mendota	USA	20090823	43° 6′n; 89°24′w	0.5–1	23.07	> 0.2	Eutrophic lake	100
Spark	Sparkling Lake	USA	20090528	46° 0′n; 89°42′w	0.5–1	13.97	> 0.2	Oligotrophic lake	0.3
Trout	Trout Bog Lake	USA	20090528	46° 2′n; 89°41′w	0.5–1	20.71	> 0.2	Dysotrophic lake	7.8
Vattern	Lake Vättern	Sweden	20070717	58°24′n; 14°36′e	0–2	17.0	0.2–100	Oligotrophic lake	3
Yellowstone1	Yellowstone Lake	USA	20080916	44°28′n; 110°22′w	0–2	46	0.1–0.8	Eutrophic lake	80
Yellowstone2	Yellowstone Lake	USA	20080915	44°28′n; 110°22′w	0–2	12.3	0.1–0.8	Eutrophic lake	80

Tot P, total phosphorus concentration (μg l^{-1}); T, temperature.

the functional potential of communities under scrutiny (i.e. Fierer *et al*., 2007; Rusch *et al*., 2007; Debroas *et al*., 2009; Oh *et al*., 2011). By annotating and comparing multiple such data sets, differences in the metabolic profiles across environments can furthermore be identified (Dinsdale *et al*., 2008), and it is also possible to identify specific genomic adaptations to life in contrasting habitats. Such metagenomic studies have previously revealed significant relationships between the environmental conditions and the functional composition of microbial communities in a wide range of habitats (Tringe *et al*., 2005; DeLong *et al*., 2006; Dinsdale *et al*., 2008; Kunin *et al*., 2008; Gianoulis *et al*., 2009; Raes *et al*., 2011) including a first comparison between metagenomes from freshwater lake and marine samples (Oh *et al*., 2011).

Here, we use metagenomic sequence data from marine and freshwater systems to identify general differences in functional gene profiles and the variability in metabolic profiles among lakes of different trophic status. Comparative analyses of freshwater bacterial communities based on taxonomic markers have previously revealed differences in bacterial community composition across trophic gradients, where specific lineages respond either positively or negatively to high productivity (Kolmonen *et al*., 2011). Microbial community structure is not only determined by environmental characteristics (Newton *et al*., 2011) and contemporary biotic interactions (Eiler *et al*., 2012) but also by a complex combination of historical factors such as dispersal limitation, past environmental conditions and evolution (Martiny *et al*., 2006). In comparison with oceans, inland waters are much more directly influenced by the surrounding terrestrial landscape and coupled to inputs of organisms and chemical constituents from the catchment. Such external influences are likely to have a profound influence on the phylogenetic composition of bacterioplankton communities (see for example Lindström, 2000; Lindström *et al*., 2005; Yannarell and Triplett, 2005; Eiler and Bertilsson, 2007; Newton *et al*., 2011; Peura *et al*., 2012).

To better understand factors controlling and shaping the community-level functional traits of freshwater microplankton, nine planktonic DNA samples from seven different lakes were analysed by pyrosequencing-enabled metagenomics. In addition, three available freshwater metagenome data sets from National Center for Biotechnology Information-Short Read Archive (NCBI-SRA) were included in the analysis, resulting in a combined freshwater data set from altogether 12 freshwater metagenomes. As marine references, we used 13 marine metagenomes comprising samples from the open and coastal ocean. One further aim was to corroborate that lake systems are not only different from marine systems in their phylogenetic but also in their functional gene composition. By comparing lakes of contrasting productivity, we further aimed at revealing functional differences related to nutrient and energy acquisition as well as substrate preferences.

Results and discussion

General description of the sampling sites and sequence data

DNA samples were collected from seven lakes, whereof two lakes were sampled twice (in Spring and Summer) (Table 1). These nine samples were subject to whole-community genome shotgun 454 pyrosequencing using Titanium chemistry. An additional three freshwater lake metagenomes and 13 marine metagenomes were obtained from public databases. The latter included samples from open-ocean and coastal habitats (Table S1). We selected these 16 metagenomes available at the time of analysis because they were of sufficient size to be compared with our data and processed in the most similar fashion to the nine new freshwater metagenomes with regards to sample handling, DNA extraction, library preparation and sequencing. Still, we want to make the reader aware that they were not processed in an identical way, which might influence the comparison and our interpretations (Carrigg *et al*., 2007). In addition, with the limited number of samples and shallow sequencing at hand, we can never cover the entire functional diversity dwelling in both marine and freshwater biomes, and this adds some uncertainties to generalizations of the major findings from this study.

The nine lakes included in the analysis represent a wide range of trophic states, including oligotrophic, mesotrophic and eutrophic systems (Table 1). They range from 0.3 to 120 $\mu g\ l^{-1}$ in total phosphorus (TP) and are all situated in the temperate climate zone. On average, 325 000 high-quality reads with mean length of 330 bp were obtained for the lake metagenomes and slightly lower numbers of 280 000 sequences with mean read length of 270 bp for the marine data sets. No quality files were available for the marine data sets, but quality filtering (mean read quality > 21) affected the lake metagenomes very little (1–2% for two data sets, 0% for the majority). To match the quality filtering step as best possible, marine metagenomes had an extra upper length filter added because many long sequences were observed to be of poor quality. Lower length limit (> 150 bp) and clustering to remove artefacts were performed in the same way on all data sets, resulting in over 8.2 million reads in total (Table 2; for detail about the removed sequences in the preprocessing steps, see Table S2). Five samples yielded much lower total sequenced nucleotides than average (84%): marine sample from Sargasso Sea (depth 40 m, 67%) and four lake samples from Yellowstone Lake (sample 1, 79%), Lake Mendota (spring sample, 76%), Trout-Bog Lake (75%) and Sparkling Lake (58%). These samples were also among the most extreme outliers in terms of the eukaryotic content. To ensure robustness of the results, the impact of including/excluding those samples from the statistical analyses was investigated.

In order to investigate the genomic similarity between and within freshwater and marine samples, DNA sequences were first evaluated for features that did not a priori require any taxonomic or functional annotation. Sequences were evaluated for Guanine and Cytosine (GC) content, isoelectric point and amino-acid usage. The GC content of the freshwater metagenome samples was 46.6% on the average (Table 2), ranging from 35% to 60% for the large majority of reads in each of the individual samples (Fig. S1). This was not significantly different to the average GC content of the marine metagenome samples (Wilcoxon test; $P = 0.406$) where for example the Sargasso Sea samples (46.6–48.6% on the average) had higher GC content than the Western English Channel (below 40%). The isoelectric points were not significantly different (Wilcoxon test; $P = 0.624$) between freshwater and marine metagenomes using Open Reading Frames (ORFs) of at least 50 aa in length predicted from six frame translation procedures (Table 2). Nor did the inferred amino acid usage differ between marine and freshwater samples [permutational multivariate analysis of variance (PERMANOVA); $P = 0.432$]. Specifically, we observed no difference in the usage of sulphur-containing amino acids, methionine and cysteine, for which an increased cost could be expected in freshwater environments. Hence, there was no convincing evidence for 'elemental sparing', which has been described as adaptive selection pressure on amino acid usage when cellular maintenance costs for protein synthesis are assumed to affect fitness (Bragg and Wagner, 2009).

Taxonomic composition

The microbial diversity captured in the metagenomic sequences from the 25 different metagenomes was analysed using rRNA hidden Markov models (hmm) and tblastx against Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). Identification and analysis of rRNA genes with hmm identified 16 743 small subunit (SSU) rRNA hits, applying an e-value cut-off of 1e-10 for a hit (Table 2). From these, 33% were of bacterial origin, whereas 2.2% and 0.6% were annotated to eukaryotes and archaea, respectively, with the rest being unclassified (64%) using the SILVA database (Quast *et al*., 2013) in combination with the naïve Bayesian classifier (Wang *et al*., 2007). Two lake metagenomes (Spring sample from Lake Mendota and Trout Bog Lake) had more than 20% of eukaryotic (18S rRNA) reads annotated as mainly algal-derived. Comparing marine and freshwater metagenomes, archaeal 16S rRNA were more common in marine systems (on average, 3.8% of the annotated SSUs in the marine vs. 0.4% in the freshwater metagenomes) when compared with freshwaters where the proportion of eukaryotic 18S rRNA hits was higher (on average, 3.2% of the annotated SSUs in the marine vs. 10.2% in the freshwater metagenomes). Possible explanations are upwelling events at marine sites that may contribute Archaea to surface communities, but also general physicochemical differences between marine and freshwaters could select for the observed patterns. The taxonomic composition of bacteria in each individual sample was also determined by annotating 16S rRNA genes using a custom curated freshwater database (Newton *et al*., 2011) (Fig. 1A). Whatever database used, Proteobacteria was the dominant bacterial phylum in all marine metagenomes. Conversely, all but five of the lake metagenomes instead featured Actinobacteria as the most abundant phylum. In marine environments, alpha-Proteobacteria was the dominant class within the Proteobacteria, whereas beta-Proteobacteria were

Table 2. Characteristics of lake and marine metagenomes.

ID	Reference/site	Size after QC (Mb)	Reads after QC	GC content (%)	Isoelectric point	SSU rRNA genes	% Eukaryotic SSU rRNA genes	Reads with STRING hit	Reads with COG assignment	% Reads with COG assignment	Average number of single copy COGs	% Bacteria among single copy COGs	Simple EGS Mb/single copy COG
DamariscottaSP	Martinez-Garcia *et al.* 2012	121	281 625	48.6	9.75	531 (185/0/5)	2.6	149 906	135 640	42	78	94	1.55
DamariscottaSU	Martinez-Garcia *et al.* 2012	140	323 939	48.2	9.82	666 (200/0/27)	11.9	156 281	140 701	50	93	93	1.51
Ekoln	this study	115	284 609	46.0	9.49	622 (209/0/13)	5.9	107 593	94 783	33	69	95	1.67
Erken	this study	233	554 862	44.9	9.41	1170 (399/0/5)	1.2	273 058	250 931	45	196	93	1.19
Lanier	Oh *et al.* 2011	449	1 078 031	47.1	9.77	1989 (714/0/20)	2.7	440 459	399 647	37	252	93	1.78
MendotaSP	Martinez-Garcia *et al.* 2012	133	319 321	45.7	9.52	1118 (242/0/149)	38.1	124 837	111 654	25	77	97	1.73
MendotaSU	Martinez-Garcia *et al.* 2012	192	447 054	47.7	9.75	795 (247/0/31)	11.2	173 517	146 222	46	76	95	2.53
Spark	Martinez-Garcia *et al.* 2012	26	66 160	52.5	10.01	108 (28/0/5)	15.2	22 364	19 857	30	8	87	3.25
Trout	Martinez-Garcia *et al.* 2012	60	150 515	46.5	9.59	335 (63/0/21)	25.0	46 795	41 628	28	26	88	2.31
Vattern	this study	117	285 637	47.4	9.67	540 (177/0/15)	7.8	116 970	103 047	36	66	93	1.77
Yellowstone1	SRR077348	181	416 139	43.7	9.34	541 (212/0/2)	0.9	152 376	136 972	33	83	93	2.18
Yellowstone2	SRR078855	107	346 239	41.4	9.03	754 (256/1/0)	0.0	91 459	86 132	25	75	97	1.43
FRESHWATER (Mean)		156	379 511	46.6	9.60	764 (244/0/24)	10.2	154 635	138 935	37	92	93	1.91
BATS0	Sargasso Sea	118	478 976	48.0	9.74	1137 (431/0/13)	2.9	142 979	131 449	27	104	97	1.14
BATS200	Sargasso Sea	134	525 891	48.3	9.70	1049 (310/38/13)	3.6	133 259	121 763	23	97	85	1.38
BATS250	Sargasso Sea	115	456 677	46.6	9.63	606 (183/20/9)	4.2	95 919	88 658	19	70	89	1.65
BATS40	Sargasso Sea	95	394 461	48.1	9.78	675 (227/0/17)	7.0	86 262	79 155	20	67	96	1.42
EqDP35155	Equatorial Pacific	56	219 390	45.4	9.70	508 (164/10/3)	1.7	62 135	57 103	26	53	91	1.05
NPTG35179	North Pacific Tropical Gyre	45	181 907	44.8	9.53	656 (253/4/4)	1.5	55 589	51 145	28	45	95	1.00
PNEq35163	Pacific North Equatorial	55	221 925	49.8	9.94	790 (300/6/5)	1.6	59 337	53 915	24	52	92	1.06
PNEqCc35171	Pacific North Equatorial	13	50 267	42.5	9.38	101 (31/3/2)	5.6	15 791	14 620	29	13	92	0.97
SPSG35131	South Pacific Subtropical Gyre	36	155 219	47.7	9.77	583 (225/1/4)	1.7	46 502	42 726	28	39	96	0.94
SPSG35139	South Pacific Subtropical Gyre	16	61 766	41.9	9.33	169 (71/1/0)	0.0	23 083	21 352	35	19	97	0.85
SPSG35147	South Pacific Subtropical Gyre	21	80 088	43.2	9.47	259 (97/3/1)	1.0	28 681	26 504	33	25	93	0.83
WChannelApr	Gilbert *et al.* 2010	102	278 931	39.2	9.01	317 (64/0/6)	8.6	82 819	67 968	24	35	90	2.91
WChannelJan	Gilbert *et al.* 2010	208	548 680	38.4	8.97	724 (195/17/6)	2.8	180 844	153 475	28	100	74	2.08
MARINE (Mean)		78	281 091	44.9	9.53	583 (196/8/6)	3.2	77 938	69 987	25	55	91	1.33

The isoelectric point represents the average pH at which predicted genes from a specific metagenome carry no electric charge. In column seven, numbers in parentheses represent the number of SSU rRNA genes annotated to Bacteria, Archaea and Eukaryota, respectively.
COG, clusters of orthologous groups; EGS, effective genome size; QC, quality filtering; mb, megabases; SSU rRNA, small subunit of the ribosomal RNA.

more abundant in freshwaters. Other abundant phyla in the lakes were Verrucomicrobia, Planctomycetes, Cyanobacteria, and Bacteroidetes. Furthermore, we observed significant differences between marine and freshwater metagenomes in community composition analysed at the phylum level (PERMANOVA, $R^2 = 0.34$, $P < 0.001$). Resolving sequences to a finer taxonomic level (roughly comparable with genus-level) revealed a dominance of previously identified typical freshwater bacteria in the 12 lake samples, including the freshwater SAR11 (LD12), taxa within the Actinobacterial acI lineage (acI-B1, acI-A6, acI-C2) and the beta-Proteobacterial Polynucleobacter (Fig. 1A; Newton *et al*., 2011). Moreover, this reflects previously described patterns between systems of different trophic status where dystrophic (humic) systems such as Trout Bog Lake are lacking most typical freshwater taxa (Peura *et al*., 2012).

Using the taxonomic annotations of the best tblastx hit to STRING revealed patterns highly similar to that of the SSU rRNA taxonomy where hits to bacteria dominated (on average 92%) over hits matching archaea (2%) and eukarya (7.6%) (Fig. S2). As for most metagenomes, the dominant portion of the reads had no hits (on average 60% for the lake and 71% for the marine data sets) in the STRING database and could thus not be taxonomically assigned. Still, comparing freshwater and marine metagenomes revealed that hits to the bacterial phylum Actinobacteria were more abundant in freshwater metagenomic libraries (on average 31%) compared with marine metagenomes where hits to Proteobacteria, especially alpha-Proteobacteria, were dominant (on average, 38%; see Fig. 1B), thus corroborating observations made at the SSU rRNA level. Other prominent (sub)phyla in the freshwater metagenomes were beta-Proteobacteria (on average 24%), Bacteroidetes (on average 10%), Cyanobacteria (on average 21%), Verrucomicrobia/Chlamydia (on average 3%) and Planctomycetes (on average 2%). Overall, the metagenomic comparison revealed taxonomic distributions as expected from previous studies based on clone libraries (i.e. Zwart *et al*., 2002; Eiler and Bertilsson, 2004) and fluorescence *in situ* hybridization (Glöckner *et al*., 1999).

Comparative functional metagenomics between marine and freshwater systems

Functional assignment was made on the basis of the best tblastx cluster of orthologous genes (COGs) hit using an E-value threshold of $1e^{-10}$. To assure the best available taxonomic representation, the STRING database was used (Franceschini *et al*., 2013), as it comprises over 1000 genomes of bacteria, archaea and eukaryota compared with 66 genomes in the original COG database. The average percentage of the reads that could be annotated (had a COG annotation) was 37% for lake and 25% for marine metagenomes (range 19–50% per sample). The total number of annotations (COGs) per sample ranged from about 14 500 to almost 400 000 (Table 2). The relative abundance of best hits assigned to each major subsystem (orthologous gene classes, OGCs) in the marine versus freshwater system is summarized in Table 3, showing that 'Amino acid transport and metabolism' was the dominant OGC.

Counts for 35 marker COGs were used to approximate the average effective genome size in freshwater and marine microbial communities. The estimated average effective genome size for freshwaters (1.91) was slightly higher than for the selected marine systems (1.33, Table 2) (Wilcoxon test; $P < 0.003$) where the latter estimates are similar to previous estimates for marine plankton (Raes *et al*., 2007; Quaiser *et al*., 2011). These findings corroborate the widespread assumption that small and streamlined genomes are a more common feature of bacterioplankton from oligotrophic sites (Giovannoni *et al*., 2005; Grote *et al*., 2012) compared with those that reside in more productive waters such as eutrophic freshwater lakes (i.e. lakes Ekoln, Erken and Mendota; see also Oh *et al*., 2011). Discrepancies in estimated genome sizes to previously published estimates (Lake Lanier, our estimate 1.78 vs. published 2.2) are most likely due to differences in databases and quality filtering used.

COGs were normalized against best hits to 35 likely essential and single copy COGs (Table S3; Ciccarelli *et al*., 2006; Raes *et al*., 2007) without taking read length into account prior to statistical analyses. Each of these single copy COGs had, on average, 77 hits in the 25 metagenomes (range 11–279, representing averages from single metagenomes). To assess whether or not each biome had a distinct functional profile, an ordination was conducted using an occurrence matrix of COGs in nonmetric multidimensional scaling (metaMDS function in R; Oksanen *et al*., 2008). PERMANOVA (Anderson, 2001) corroborated the visual impression (Fig. S3) of a significant difference in functional beta-diversity between marine and freshwater systems (PERMANOVA; $P < 0.001$, $R^2 = 0.34$). These differences were maintained even if low-quality metagenomes were excluded (PERMANOVA; $P < 0.001$, $R^2 = 0.34$), or when only bacterial COGs where analysed (PERMANOVA; $P < 0.001$, $R^2 = 0.35$) and when specific OGCs were analysed separately (Table 3). The most pronounced difference in the composition of OGCs was observed for the OGC 'ion transport and metabolism' and 'transcription', whereas the composition within OGCs 'Cytoskeleton' and 'Cell motility' were the least separated. Moreover, we also looked for proportional differences at the level of OGC by using Wilcoxon test. Overall, OGCs 'energy production and conversion' and 'coenzyme transport and metabolism'

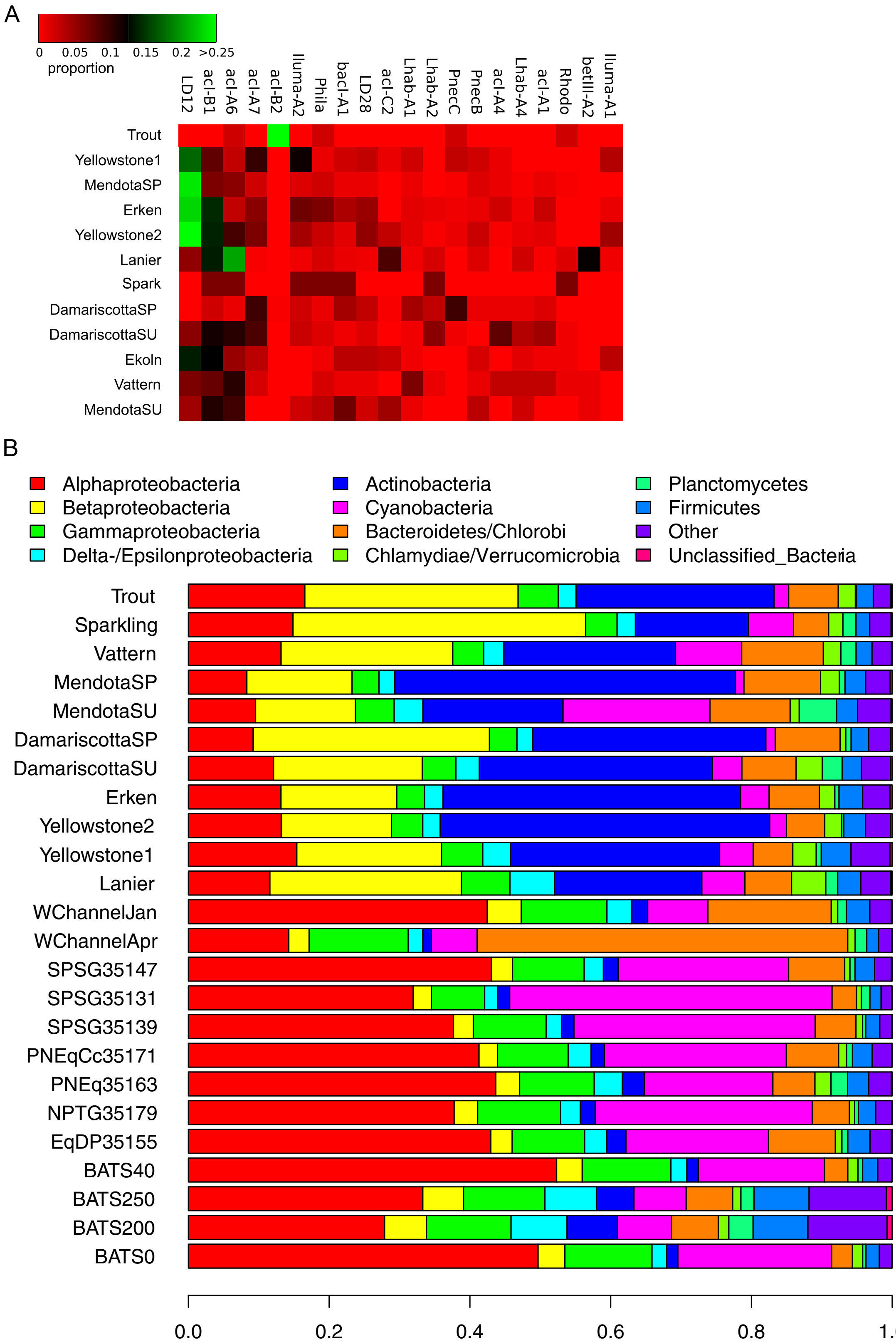


Fig. 1. Heatmap of the 20 most abundant typical freshwater taxa (A) in the metagenomics datasets as inferred from their proportion of SSU rRNA gene sequences. Typical freshwater taxa were defined previously using a well-curated freshwater-specific phylogeny (Newton *et al.*, 2011). (B) Barplot showing taxonomic classification of bacterial reads into phyla based on the best hit to STRING (Franceschini *et al.*, 2013).

Table 3. Summary statistics of each OGC and their comparison between freshwater and marine metagenomes.

		Median best[a] (all[b])		Wilcoxon test				Permanova					
				Best[a]		All[b]		Best[a]			All[b]		
		Fresh	Ocean	W	*P* value	W	*P* value	F	R2	*P* value	F	R2	*P* value
Translation, ribosomal structure and biogenesis	J	9.4%	10.4%	70	0.098	118	0.030	4.7	0.21	0.002	6.1	0.21	0.001
RNA processing and modification	A	0.0%	0.0%	38	0.473	57	0.270	8.1	0.31	0.001	10.3	0.31	0.001
Transcription	K	3.3%	3.1%	9	0.002	11	0.000	13.0	0.42	0.001	17.1	0.43	0.001
Replication, recombination and repair	L	8.4% (8.6%)	6.4%	4	0.000	6	0.000	10.2	0.36	0.001	13.8	0.38	0.001
Chromatin structure and dynamics	B	0.0%	0.0%	54	0.678	75	0.894	3.7	0.17	0.003	3.9	0.14	0.003
Cell cycle control, cell division, chromosome partitioning	D	1.4% (1.3%)	1.5%	66	0.181	107	0.123	11.4	0.39	0.002	15.1	0.40	0.001
Nuclear structure	Y	0.0%	0.0%	40	0.447	71	0.655	NA	NA	NA	NA	NA	NA
Defence mechanisms	V	2.0%	1.7%	20	0.031	25	0.003	5.7	0.24	0.005	7.9	0.26	0.001
Signal transduction mechanisms	T	1.9%	1.2%	1	0.000	1	0.000	12.3	0.41	0.001	16.9	0.42	0.001
Cell wall/membrane/envelope biogenesis	M	6.3%	5.3%	12	0.004	17	0.000	10.0	0.36	0.001	14.2	0.38	0.001
Cell motility	N	0.2%	0.2%	29	0.157	57	0.270	2.2	0.11	0.075	2.5	0.10	0.047
Cytoskeleton	Z	0.1%	0.3%	81	0.010	127	0.007	3.6	0.17	0.023	3.2	0.12	0.021
Extracellular structures	W	0.0%	0.0%	48	NA	78	NA	NA	NA	NA	NA	NA	NA
Intracellular trafficking, secretion and vesicular transport	U	1.3%	1.4%	73	0.057	114	0.052	5.3	0.23	0.002	7.6	0.25	0.001
Posttranslational modification, protein turnover, chaperones	O	5.0%	5.6%	82	0.007	137	0.001	10.4	0.37	0.001	13.0	0.36	0.001
Energy production and conversion	C	9.5%	11.2% (11.1%)	89	0.001	148	0.000	6.6	0.27	0.001	9.2	0.29	0.001
Carbohydrate transport and metabolism	G	5.5%	5.3%	22	0.047	48	0.110	7.2	0.28	0.001	8.8	0.28	0.001
Amino acid transport and metabolism	E	11.7%	13.7%	89	0.001	146	0.000	11.9	0.40	0.001	14.6	0.39	0.001
Nucleotide transport and metabolism	F	3.8%	4.2%	65	0.208	101	0.225	6.9	0.28	0.001	8.0	0.26	0.001
Coenzyme transport and metabolism	H	4.1%	4.9%	96	0.000	156	0.000	7.9	0.31	0.002	10.0	0.30	0.001
Lipid transport and metabolism	I	4.2%	4.4% (4.3%)	69	0.115	106	0.137	8.6	0.32	0.001	11.3	0.33	0.001
Inorganic ion transport and metabolism	P	4.2% (4.1%)	4.3%	58	0.473	101	0.225	13.0	0.42	0.001	16.8	0.42	0.001
Secondary metabolites biosynthesis, transport and catabolism	Q	1.8% (1.7%)	1.5%	27	0.115	51	0.152	10.5	0.37	0.001	13.8	0.38	0.001
General function prediction only	R	10.0%	9.2% (9.1%)	26	0.098	38	0.030	8.8	0.33	0.001	11.7	0.34	0.001
Function unknown	S	5.5%	3.9%	6	0.000	7	0.000	12.2	0.40	0.001	16.0	0.41	0.001

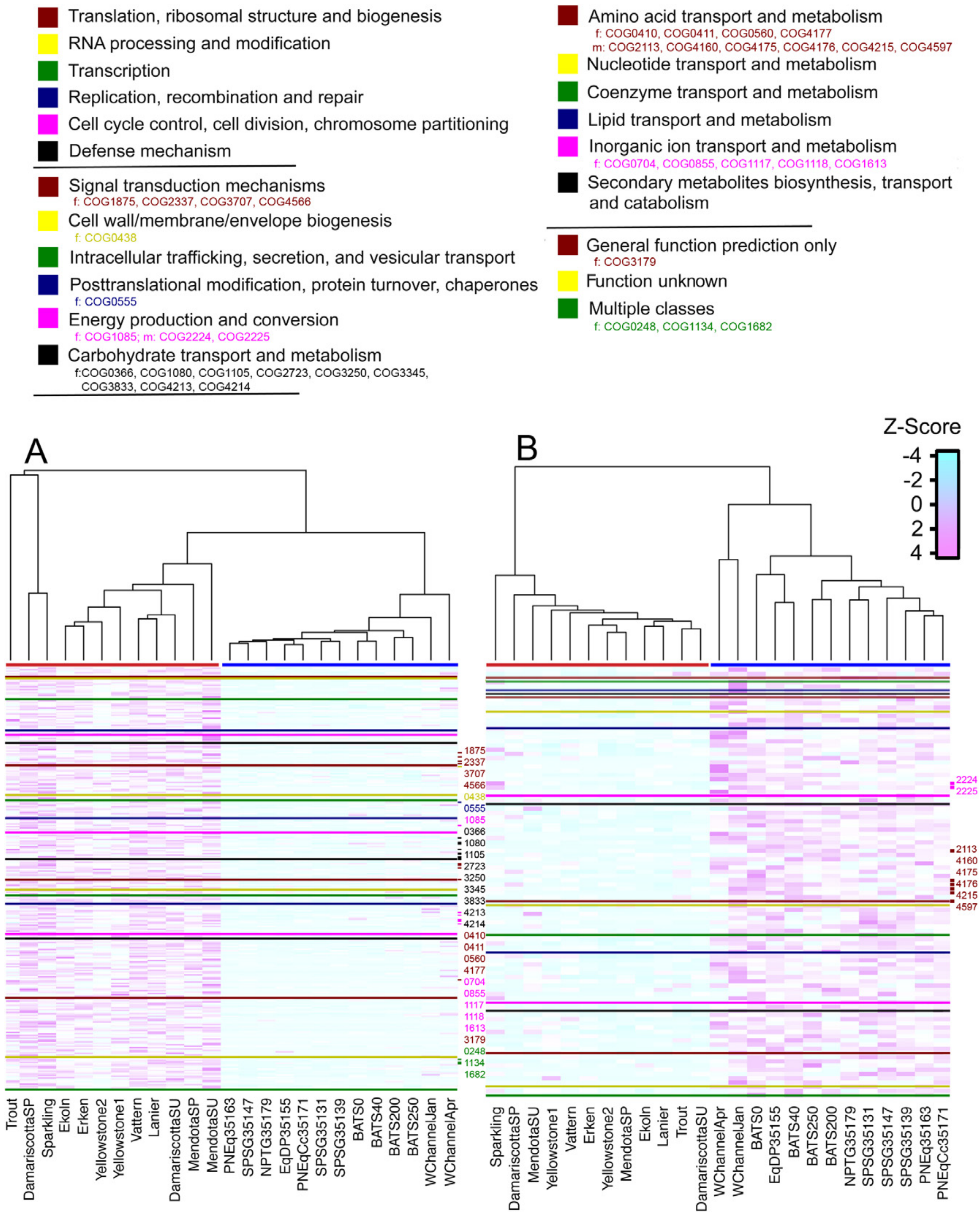
a. Smaller data set of eight lake and 12 ocean samples, excluding worst quality samples (see Table S1).
b. Full data set of 12 lake and 13 ocean samples (see Table S1).
Average and standard deviation are derived from the relative fraction of OCGs averaged over all marine and freshwater metagenomes, respectively. *P*-values and W statistics from Wilcoxon test on the contribution of ORFs to each OGC as well as results from PERMANOVA to test for differences in functional composition between marine and freshwaters using normalized COGs from each OGC.
NA, Not Assessed.

were under-represented in freshwater metagenomes, whereas core functions involved in 'transcription', and 'replication, recombination and repair' were over-represented when compared with marine samples (Table 3). The higher proportion of the OGC 'signal transduction' in freshwater than marine metagenomes suggest that freshwater microbial communities feature more complex interactions and cellular controls that may involve cell-to-cell communication.

This was also reflected in a more detailed analysis based on the Wilcoxon test where all COGs differing in resampled and normalized occurrence between marine and freshwater systems were tested. Out of 707 COGs identified as significantly different in their prevalence between the marine and freshwater metagenomes ($P < 0.01$ and false discovery rate < 0.027), and 560 significantly different ($P < 0.01$) when excluding low-quality metagenomes, limited the list to COGs significant for both all and best data sets to 102 COGs that were over-represented in the marine and 295 in the freshwater metagenomes (Fig. 2, Table S4). For example, core functions belonging to 'transcription' such as transcriptional regulators, for example arginine repressor (bacterial) was significantly over-represented in lakes ($P < 0.001$). 'Replication, recombination and repair' was represented by numerous transposases, several helicases, and the recombination repair proteins RecF and RecB, which were all significantly over-represented in the lake metagenomes (all bacterial, $P < 0.01$). Other COGs over-represented in freshwaters were related to a phosphorus starvation-inducible protein *phoH* (cog1875, $P < 0.001$), a growth inhibitor (cog2337, $P < 0.002$) and two response regulators (cog3707, $P < 0.001$; cog4566, $P < 0.001$). Homologues to subunits of archaeal polymerases such as COG1311 (archaeal DNA polymerase II, SSU/DNA polymerase delta, subunit B) and COG1933 (archaeal DNA polymerase II, large subunit) were over-represented in marine metagenomes ($P < 0.005$ and $P < 0.001$ respectively). With regards to metabolism, differences between freshwater and marine metagenomes were limited to few key enzymes (Fig. 2). Examples of this is the significant over-representation of malate synthase homologues (cog2225, $P < 0.001$) and isocitrate lyase (cog2224, $P < 0.002$) in the marine biome, both coding for enzymes with a central function in the glyoxylate cycle. The isocitrate lyase catalyses the cleavage of isocitrate to succinate and glyoxylate, and the malate synthase feeds glyoxylate into the tricarboxylic acid cycle (TCA) via oxalacetate (known as the glyoxylate shunt). This allows microorganisms to utilize simple carbon compounds as a carbon source when complex sources such as glucose are not available. In the absence of available carbohydrates, the glyoxylate cycle permits the synthesis of carbohydrates needed for cell-wall assembly from lipids via acetate. In contrast, reads annotated as being involved in carbohydrate metabolism (i.e. 'phosphoenolpyruvate-protein kinase' cog1080, $P < 0.001$; 'Fructose-1-phosphate kinase and related fructose-6-phosphate kinase' cog1105, $P < 0.002$) seem to be more common in freshwater as compared with marine metagenomes, where such genes were never significantly over-represented. This included galactose-1-phosphate uridylyltransferase (cog1085, $P < 0.001$) a putative enzyme central to the Leloir pathway involved in the catalyses between galactose and glucose. Another interesting finding was that homologues of enzymes that hydrolyse glycolipids, glycoproteins, lactose and galactosides to monosaccharides such as alpha-(cog3345, $P < 0.001$) and beta-galactosidases (cog3250, $P < 0.004$) were over-represented in freshwater metagenomes. Also, other homologues to enzymes catalysing the hydrolysis of glycosidic linkages were over-represented in the freshwaters metagenomes, including chitinase (cog3179, $P < 0.001$), glycotransferase (cog438, $P < 0.001$) and glycosidase (cog2723, $P < 0.001$; cog366, $P < 0.001$), known to mediate the production of oligosaccharide and monosaccharide from chitin, cellulose and hemicelluloses. This is consistent with a recent finding that the genomes of the abundant acI-B1 taxon of freshwater Actinobacteria are enriched with glycosidase homologues when compared with other bacterial genomes (Garcia *et al*., 2012).

Moreover, freshwater microbial genomes seem to harbour a higher proportion of certain putative genes involved in transport of sugars such as xylose (cog4213, $P < 0.001$; cog4214, $P < 0.001$) and various polysaccharides (cog1134, $P < 0.001$; cog1682, $P < 0.001$; cog3833 $P < 0.001$) (Fig. 2). A similar pattern was also observed for genes involved in transport of peptides (cog410, $P < 0.001$; cog411, $P < 0.001$; cog4177, $P < 0.002$). In contrast, ORFs putatively identified as ATP-dependent amino-acid transporters (cog2113, $P < 0.00009$; cog4160, $P < 0.001$; cog4175, $P < 0.001$; cog4176, $P < 0.002$; cog4215, $P < 0.001$; cog4597,

Fig. 2. Heatmap of COGs showing only those that were either significantly over- (A) and under-represented (B) in freshwater metagenomes when compared with marine metagenomes after resampling and normalization against single-copy core COGs. Significantly over- and under-represented COGs were identified by Wilcoxon test ($P < 0.01$) when testing all data sets, as well as the best data sets only, and the subsequent estimation of false discovery rate (q < 0.027). These lists are not exhaustive and only include well-characterized COGs. COGs mentioned in the text are indicated. Dendograms from hierarchical cluster analysis based on displayed COGs are shown at the top of each graph.


Translation, ribosomal structure and biogenesis
RNA processing and modification
Transcription
Replication, recombination and repair
Cell cycle control, cell division, chromosome partitioning
Defense mechanism
Signal transduction mechanisms
f: COG1875, COG2337, COG3707, COG4566
Cell wall/membrane/envelope biogenesis
f: COG0438
Intracellular trafficking, secretion, and vesicular transport
Posttranslational modification, protein turnover, chaperones
f: COG0555
Energy production and conversion
f: COG1085; m: COG2224, COG2225
Carbohydrate transport and metabolism
f:COG0366, COG1080, COG1105, COG2723, COG3250, COG3345, COG3833, COG4213, COG4214
Amino acid transport and metabolism
f: COG0410, COG0411, COG0560, COG4177
m: COG2113, COG4160, COG4175, COG4176, COG4215, COG4597
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Inorganic ion transport and metabolism
f: COG0704, COG0855, COG1117, COG1118, COG1613
Secondary metabolites biosynthesis, transport and catabolism
General function prediction only
f: COG3179
Function unknown
Multiple classes
f: COG0248, COG1134, COG1682
A
B
Z-Score
-4
-2
0
2
4
Trout
DamariscottaSP
Sparkling
Ekoln
Erken
Yellowstone2
Yellowstone1
Vattern
Lanier
DamariscottaSU
MendotaSP
MendotaSU
PNEq35163
SPSG35147
NPTG35179
EqDP35155
PNEqCc35171
SPSG35131
SPSG35139
BATS0
BATS40
BATS200
BATS250
WChannelJan
WChannelApr
Sparkling
DamariscottaSP
MendotaSU
Yellowstone1
Vattern
Erken
Yellowstone2
MendotaSP
Ekoln
Lanier
Trout
DamariscottaSU
WChannelApr
WChannelJan
BATS0
EqDP35155
BATS40
BATS250
BATS200
NPTG35179
SPSG35131
SPSG35147
SPSG35139
PNEq35163
PNEqCc35171

$P < 0.001$) were significantly over-represented in marine metagenomes. We propose that the compositional differences in amino acid and carbohydrate metabolism is a consequence of major differences in the overall composition of organic substrates available for heterotrophs in the respective biomes. Freshwater systems, including the temperate systems of this study, are highly influenced by allochthonous organic matter inputs from the catchment as well as plant-derived polysaccharides (e.g. xylose-containing hemicellulose) inputs from the littoral zone, whereas marine systems are less influenced by organic matter loadings from such terrestrial surroundings and littoral fringe zones and instead rely largely on autochthonous organic matter inputs from plankton rich in proteinaceous materials (Duarte and Cebrián, 1996; Bertilsson and Jones, 2003).

ORFs putatively involved in acquisition of phosphate (cog573/cog581, $P < 0.003/0.006$; cog1117, $P < 0.002$) including phosphate uptake regulators (cog704, $P < 0.003$) and sulphate (cog555, $P < 0.007$; cog1118, $P < 0.001$; cog1613, $P < 0.002$; cog4208, $P < 0.001$) were mostly over-represented in freshwater genomes. The over-representation of exopolyphospatase (cog248, $P < 0.001$) and polyphosphate kinase (cog855, $P < 0.001$) homologues supports the previously recognized role of polyphosphates as a form of phosphorus storage in freshwater environments (Broberg and Persson, 1988; Ilikchyan *et al.*, 2009). We did not observe any significant differences in nitrogen metabolism and uptake between the marine and often more productive freshwater systems. The previously inferred reliance on potassium instead of sodium for osmoregulation was a typical feature of the freshwater metagenomes as well as a higher representation of reads annotated as cobalt, magnesium and nickel transporter systems (Fig. 2A). In contrast, homologues of zinc and manganese transporters were over-represented in marine metagenomes (Fig. 2B). This confirms previously reported differences in osmoregulatory traits between freshwater and marine microorganisms inferred from comparative metagenomics of microbial communities (Oh *et al.*, 2011). These findings are also consistent with recent results based on comparisons of 16S rRNA gene libraries (Zwart *et al.*, 2002; Lozupone and Knight, 2007; Logares *et al.*, 2009; Newton *et al.*, 2011) where salinity was suggested to represent a strong environmental barrier for microorganisms. Our results also point to the importance of factors other than salinity, at least when comparing marine and freshwater environments with regards to substrate availability and substrate acquisition. As illustrated earlier, microbial communities in these contrasting biomes seem to have different metabolic capabilities as genes involved in amino acid metabolism were over-represented in marine metagenomes when compared with freshwater metagenomes, and clear differences in the strategies of carbohydrate metabolism were observed.

Comparing functional profiles among freshwater systems

When freshwater functional profiles were analysed by non-metric multidimensional scaling, it was apparent that Sparkling Lake and Trout Bog Lake metagenomes were rather distinct from the others (Fig. 3). This can at least partly be attributed to their high amounts of eukaryotic sequences. An additional non-exclusive explanation may be trophic status: Trout Bog Lake was the only humic (dystrophic) system, whereas Sparkling Lake was the most oligotrophic system in the study. Interestingly, we observed a significant correlation between the overall functional composition and TP, a widely used proxy for ecosystem productivity (Schindler, 1978) ($R^2 = 0.53$, $P = 0.029$; Fig. 3). The correlation was even more significant if only bacteria were taken into account ($R^2 = 0.52$, $P = 0.018$). The observation that the functional profile of one metagenome from Yellowstone Lake was very different from the others was probably caused by the proximity of this sample to a thermal vent and the associated higher temperature and different ion composition. For a more detailed analysis, we relied on maximal information-based non-parametric exploration (MINE; Reshef *et al.*, 2011) statistics for identifying and classifying relationships between the proportion of COGs and TP. We used a maximal information coefficient (MIC) > 0.54 (uncorrected $P < 0.05$ and false discovery rate < 2.56e-07) to identify COGs that were significantly related to productivity (TP) in the sampled lakes. A total of 183 COGs of 3335 COGs tested were identified using these criteria, whereof 34 COGs were positively related to TP (Table S5). An inverse relationship to TP was observed for certain active transporters of phosphonates (cog3454, cog4107) and organic compounds such as amino acids (cog559, cog1147, cog4177). Homologues to other active transporters such as permeases (cog2998, cog4603, cog5265) that facilitate the transport of for example nitrate and sulphate (cog619, cog659) were negatively related with TP. The number of predicted homologues to phosphoserine phosphatase (cog560) and serine acetyltransferase (cog1045) genes involved in amino acid metabolism was negatively correlated with TP as were genes with a crucial role in carbohydrate degradation (cog153, cog1082, cog3250). Other gene products that could be useful for diagnostics of metabolic processes were carbon-monoxide dehydrogenase CoxLMS subunits (CO oxidation) that were significantly negatively related to TP. These genes are involved in the oxidation of CO to CO_2 and represent an alternative or supplementary energy source that is widespread in marine bacteria (King and Weber, 2007;

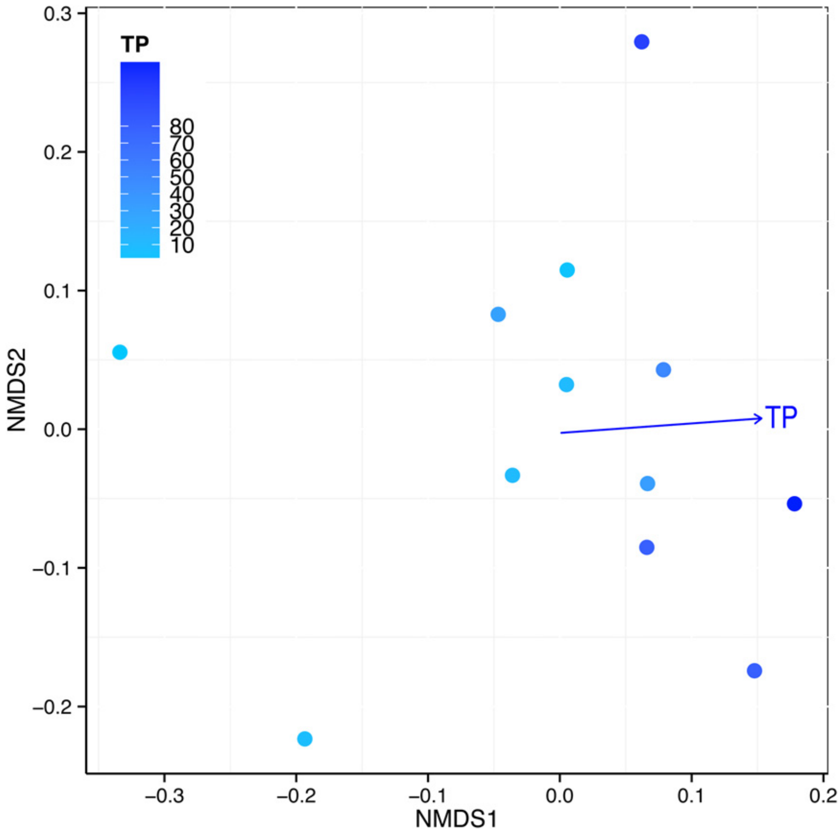


Fig. 3. Non-metric multidimensional scaling plot of microbial functional diversity along a productivity gradient (stress-value = 0.10). This plot is based on Horn–Morisita distances from COGs lists of 12 freshwater metagenomes. Total phosphors (TP) was mapped as en environmental variable vector onto the ordination using R function (TP) 'envfit'. NMDS, Non-parametric-Multi-Dimensional-Scaling.

Brinkhoff *et al.*, 2008). CO-dehydrogenase genes were detected at higher relative abundance in three lakes with low levels of TP: Trout Bog, Damariscotta and Vättern.

The significant relationships observed between TP and COG patterns inferred by MINE mainly provide new genome-level confirmation of earlier empirical findings of how microbial processes such as sugar, amino acid and phosphate acquisition strategies are structured along productivity gradients but also identify variations in the occurrence of response regulators that allow microbes to sense and to react to environmental stress (i.e. cog589).

Phylogenetic analyses of selected functional genes and the correspondence between functional and taxonomic composition

Phylogenetic trees were constructed for a selected number of proteins including the mmoA, nirK, pstA/B, RuBisCo and the nifH/bchL/chlL family, including Swiss-Prot references and their homologues in the metagenomes (Fig. S4). The selected genes are involved in key biogeochemical processes including methane oxidation, denitrification, phosphorus uptake, CO_2 fixation, nitrogen fixation and the synthesis of photopigments. Obtained phylogenies were analysed to infer the phylogenetic structuring between and within freshwater and marine sequences using PYLOCOM (Webb *et al.*, 2011). Resulting beta nearest taxon indexes (βNTIs; see *Experimental procedures*) from the functional genes were compared with βNTI derived for the 16S rRNA (Table 4). These comparisons revealed that proteins, similar to the 16S rRNA, exhibit phylogenetic overdispersion between biomes when compared with random phylogenetic structures. This infers that freshwater and marine protein

Table 4. Results from phylogenetic analyses estimating beta-NTI within and between marine and freshwater sequences.

BetaNTI		Freshwater	Marine
Freshwater	16S	−7.775	
	bcn	−1.989	
	nirK	0.036	
	RuBisCo	15.345	
	pstA/B	−3.156	
	mmoA	12.287	
Marine	16S	−27.123	21.883
	bcn	−4.866	1.754
	nirK	−3.736	0.278
	RuBisCo	−232.179	30.088
	pstA/B	−1.654	3.144
	mmoA	−90.028	22.273

Genes annotated as 16S rRNA and related to functional genes such as mmoA, nirK, pstA/B, RuBisCo, and the nifH/bchL/chlL family. Values above +2 indicate phylogenetic clustering, whereas a NTI below −2 indicates overdispersion.

sequences are more different from each other than expected by chance (Webb *et al*., 2011). This suggests that these key functional genes from marine and freshwater biomes are usually not closely related and often group into distinct marine and freshwater phylogenetic clusters, similar to what has been reported before for the 16S rRNA marker gene (Logares *et al*., 2009).

To determine if 16S rRNA-derived taxonomic and functional profiles among the metagenomes were coherent, a procrustes analysis was performed (Oksanen *et al*., 2008). Our results demonstrate that the known rRNA-inferred microbial community shifts across the freshwater to marine gradient are reflected also in cohesive shifts in community-level functions observed in the metagenomes. 16S rRNA taxonomy resolved to either genus/typical freshwater taxa or phylum levels were significantly correlated with the functional data based on COG annotations ($R = 0.95$ and $R = 0.83$, respectively, $P < 0.001$ using procrustes analysis). When lake data were analysed separately, the procrustes analyses between 16S rRNA community composition (both phylum and genus composition) and functional COG annotations revealed similarly high coefficient values ($R = 0.74$ and $R = 0.84$, respectively, using procrustes analysis), but because these analyses included fewer samples, *P*-values increased dramatically ($P < 0.033$ and $P < 0.11$ respectively). This suggests that the taxonomic composition as inferred by phylogenetic markers (i.e. 16S rRNA gene) and the functional potential of communities are linked through evolutionary history. Still, it remains to be shown whether this implies that differentiation at the fine-scale population level has only minor effects on the overall gene content and potential subsequent ecosystem function, or instead is mainly determined by distribution patterns of broad taxonomic groups.

Outlook

Our metagenomic analyses of pelagic microbial communities in lakes and oceans suggest that many core functions are shared across these two biomes. Although the functional overlap is substantial, our analyses also point to some profound functional differences. Because of the rather shallow coverage of the underlying genetic diversity in the metagenomes analysed here, many genes or gene categories were not sufficiently abundant in the data set to determine with any certainty, whether or not there were significant changes in their relative abundances across the freshwater marine boundary or across the freshwater productivity gradient. This applies to genes associated with less widespread metabolic processes that may nevertheless be of critical importance to carbon and nitrogen cycling in these aquatic systems (including genes associated with N cycling, chitin degradation and ammonia oxidation). Forthcoming deeper metagenomic sequencing will likely capture trends also in these genes across environmental gradients and will help build a more comprehensive understanding of how the functional capabilities of aquatic microbial communities change along salinity and productivity gradients. Nevertheless, the present comparison of freshwater and marine metagenomes based on whole-genome shotgun sequence data did provide functional, phylogenetic and taxonomic trends across these gradients and will help us design biogeochemical experiments to test metagenome-inferred predictions such as differences in substrate preferences. Examples are the inferred prevalence towards amino acids in marine systems and difference in carbohydrate metabolism between marine and freshwaters, and the over-representation of homologues involved in the oxidation of recalcitrant organic matter in oligotrophic lakes compared with eutrophic lakes.

Experimental procedures

Sample characterization and DNA extraction

For Lakes Vättern, Ekoln and Erken, integrated water samples from the upper 2 m were collected with a rinsed 2 m Polyvinyl chloride (PVC) tube. Samples were sieved through an autoclaved 100 μm nylon mesh prior to further processing. Samples were kept dark at near *in situ* temperature and upon return to the laboratory, microbial cells from between 0.5 and 1 l of water were collected on replicated 0.2 μm membrane filters (Supor 200, 47 mm diameter; Gelman) by vacuum filtration followed by freezing at −80°C until further analyses. Water temperature profiles measured on site at the time of sampling verified that the sampling was limited to the upper mixed layer (epilimnion). TP and dissolved organic carbon was measured using standard methods as previously described (Eiler *et al*., 2012). Community DNA was extracted from individual membrane filters using the FASTDNA spin kit for soil (QBiogene, Carlsbad, CA, USA) as recommended by the manufacturer. At least three membrane filters were extracted to recover sufficient DNA for 454 pyrosequencing. The amount and quality of recovered DNA was quantified by spectrophotometry at 260 and 280 nm, and aqarose gel electrophoresis revealed DNA with an average molecular weight exceeding 20 kb. All three metagenome samples were sequenced with 454 pyrosequencing with Titanium chemistry using half a chip for the Lake Erken metagenome and one quarter of a chip for Ekoln and Vättern (separated by sample specific molecular barcodes). Samples were collected from the epilimnia of Damariscotta Lake, Lake Mendota, Sparkling Lake, and Trout Bog Lake, and sequenced as described elsewhere (Martinez-Garcia *et al*., 2012). Sequences are publicly available through the European Nucleotide Archive under project PRJEB4844.

Data mining

Metagenome data from Lake Lanier (Oh *et al.*, 2011) and two samples from Yellowstone Lake (T. McDermott, unpubl. data) were acquired from SRA (fastq-files) and analysed following the quality control and annotation procedures as described later. Fasta files for all selected marine samples were downloaded from Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (Seshadri *et al.*, 2007). Annotations were performed as described later.

Sequence annotation and functional assignment

Preprocessing was performed to bring all data sets (fasta and quality files) to the same starting point. This procedure included the following steps: length filter (length > 150) and quality filter (mean quality > 21) for lake metagenomes and just a length filter for marine metagenomes (upper length filter as listed in Table S2), clustering artificial duplicates with cd-hit-454 (Beifang *et al.*, 2010) using 97% identity threshold and 80% of the sequence in the alignment, and finally creation of consensus sequences from the clusters with cdhit-cluster consensus ignoring terminal gaps (-maxlen = 1). Quality-filtered data sets were used in all analyses. Simple six-frame translation with 50 aa length threshold was used for non-annotation-based analyses (aa usage and isoelectric point). COG annotations of the reads were extracted from the best tblastx hit against STRING (Franceschini *et al.*, 2013), and rRNAs were identified using hmm rRNA to obtain annotations. An E-value threshold of 1e-10 was applied.

We also performed a second preprocessing and annotation procedure, and subsequent statistical analyses in which results supported the main findings presented earlier. In short, a more stringent quality filtering was performed with cutting reads when quality scores dropped below 21 and using a length cut-off of 150 bp. Clustering artificial duplicates was performed as described earlier. The quality-filtered data sets were then submitted to CAMERA using Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline (RAMMCAP) (Seshadri *et al.*, 2007; Weizhong, 2009) with the following parameters: six-frame translation, hmm rRNA and annotation (no clustering), which masks tRNAs and rRNAs before calling ORFs. Subsequently, Reversed Position Specific-Basic Local Alignment Search Tool (RPS-BLAST) was performed against COG (Tatusov *et al.*, 2003). An E-value threshold of 1e-10 was applied. Fasta files for the marine data sets downloaded from CAMERA were used without any quality filtering, except cd-hit-454 for artificial duplicate removal.

Statistical analyses

To ensure robustness of the statistical tests to outliers, we have compared the results using three types of data sets: all COGs from 12 lake and 13 marine samples; all COGs from 8 lake and 12 marine samples; and only bacterial COGs from 12 lake and 13 marine samples. The smaller number of samples for the second set resulted from excluding samples with the worst quality processing results. Bacterial COGs are used to address the issues of varied eukaryotic content between the samples. The abundance of individual reads matching a particular COG were normalized against the average abundance of 35 likely essential and single copy COGs (Ciccarelli *et al.*, 2006; Raes *et al.*, 2007) and used to generate a metabolic profile of the metagenome. This provides a proxy for the number of genomes harbouring a specific COG in the community. Core-gene normalized profiles were then used in statistical analyses such as metaMDS, PERMANOVA and procrustes test with Horn-Morisita distance measure using the functions in the 'ecodist' and 'vegan' libraries in the R-package (http://www.r-project.org; Goslee and Urban, 2007; Oksanen *et al.*, 2008). PERMANOVA (Anderson, 2001) was used to determine significant differences between freshwater and marine functional beta-diversity, and procrustes analysis was used to determine correspondence between taxonomic and functional composition. To fit TP as an environmental vector onto the ordination, we used the function 'envfit'. The fitted vector is an arrow that points to the direction of its most rapid change in the ordination space (direction of the gradient), and its length is proportional to the correlation between community composition and TP. Prior to applying the Wilcoxon test, COGs were resampled and then normalized against the single-copy core COGs to identify over- and under-represented COGs in freshwater compared with marine metagenomes. False discovery rate (q-value) was estimated after Storey (2002). MINE (Reshef *et al.*, 2011) was used with default settings for identifying and classifying relationships between the resampled and normalized COG abundances, and TP that was used as a proxy of lake productivity (Schindler, 1978). Relationships were defined as significant when the MIC was > 0.54 with a P-value < 0.01 and a false discovery rate < 2.56e-07.

Taxonomic assignments

Protein-based taxonomic assignments for domain and phylum were extracted from the best hit to the STRING database (E-value threshold 10^{-10}). In addition, SSU rRNAs were extracted by hmm rRNA. The Bayesian classifier (Wang *et al.*, 2007) (using bootstrap cut-off > 60) was used to annotate 16S rRNA genes against a custom curated freshwater-specific database (Newton *et al.*,

2011) and the SILVA database using taxonomy after SILVA (Quast *et al*., 2013). The number of reads annotated to the different bacterial phyla and bacterial 'genera' were extracted and ordinated using R (Oksanen *et al*., 2008). A procrustes test was used to compare the functional annotations with the taxonomic annotations at the genus level.

Phylogenetic analyses

Reference (master) sequences for mmoA, nirK, pstA/B, RuBisCo and the nifH/bchL/chlL family were obtained from Swiss-Prot. After six-frame-shift translation of the sequences (using a minimum length of 50 aa), homologous ORFs in the 29 metagenomes were identified based on blastp searches using an E-value threshold of 1e-10 and per cent identity of 40%. Alignments of master sequences were obtained for each of the five genes using Multiple sequence comparison by log-expectation (MUSCLE) (default settings; Edgar, 2004). Preliminary multiple sequence alignment were obtained for the metagenomic ORFs by MUSCLE using settings -maxiters 1 and -diags to increase speed. These 'slave' alignments were then aligned against the master alignment with muscle using function '-profile'. Bootstrapped Random Axelerated Maximum Likelihood (RAxML) trees (Stamatakis *et al*., 2008) were computed based on trimmed master alignments using standard model JTTF and default convergence criteria. Trees and alignments were imported into ARB (Ludwig *et al*., 2004), and the quick parsimony option was used to add the aligned metagenomic ORFs to the RAxML master trees. For 16S rRNA genes, the procedure outlined in Peura and colleagues (2012) was used to insert metagenomic 16S rRNA homologues into the SILVA106 reference tree. Phylogenetic trees were visualized using iTOL (Letunic and Bork, 2011) and analysed using PHYLOCOM (Webb *et al*., 2011). The phylocom function 'comdistnt' was used to infer if freshwater and marine sequences were phylogenetically distinct by estimating the βNTI. Here, we used both the marine and freshwater biomes as separate groups. Mean nearest taxon distance (MNTD) was estimated for within each biome and between biomes. To weigh phylogenetic distances by taxa abundances, the average distance among random individuals drawn from each of the two biomes was calculated. The NTI was quantified by the number of standard deviations that the observed MNTD is from the mean of the null distribution (999 randomizations; MNTDnull). MNTDnull is found by randomizing OTUs across the phylogeny and recalculating MNTD 999 times.

NTI = −1*(MNTDobs-mean MNTDnull/sdMNTDnull) (Webb *et al*., 2002).

For a single community, NTI greater than +2 indicates that coexisting taxa are more closely related than expected by chance (phylogenetic clustering). NTI less than −2 indicates coexisting taxa are more distantly related than expected by chance (phylogenetic overdispersion). βNTI is the between-group analogue of NTI (Fine and Kembel, 2011; Webb *et al*., 2011).

Acknowledgements

We want to thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for access to data storage and computing resources under project b2011105. This work was supported by the Swedish Foundation for Strategic Research (Grant Number ICA10-0015 to AE), the Swedish Research Council (Grant Numbers 349-2007-831, 621-2008-3259 and 621-2011-4669 to SGEA; 2009-3784, 2008-1923 and 2012-3892 to SB), the National Science Foundation [Awards CBET-0644949 (CAREER), MCB-0702653 (Microbial Observatories Program) to KD and DEB-841933 to RS], DEB-0822700 (Long Term Ecological Research, NTL LTER to KDM), the European Union (grant to SGEA), the Göran Gustafsson Foundation (grant to SGEA), the Knut and Alice Wallenberg Foundation (Grant Numbers KAW-2011.0148 and KAW-2012.0075 to SGEA), and the Swedish Wennergren Foundation (to KDM and SB). Pyrosequencing was partially supported by an instrument grant from the K&A Wallenberg foundation. Friederike Heinrich and Lorena Grubisic assisted with sampling and provided metadata for the Swedish lakes.

References

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Aust Ecol* **26:** 2–46.

Battin, T.J., Luyssaert, S., Kaplan, L.A., Aufdenkampe, A.K., Richter, A., and Tranvik, L.J. (2009) The boundless carbon cycle. *Nat Geosci* **2:** 598–600.

Beifang, N., Limin, F., Shulei, S., and Weizhong, L. (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11:** 187. doi:10.1186/1471-2105-11-187.

Bertilsson, S., and Jones, J.B., Jr (2003) Supply of dissolved organic matter to aquatic ecosystems: autochthonous sources. In *Aquatic Ecosystems: Interactivity of Dissolved Organic Matter*. Findlay, S.E.G., and Sinsabaugh, R.L. (eds). New York, NY, USA: Academic Press, pp. 3–24.

Bragg, J.G., and Wagner, A. (2009) Protein material costs: single atoms can make an evolutionary difference. *Trends Genet* **25:** 5–8.

Brinkhoff, T., Giebel, H.A., and Simon, M. (2008) Diversity, ecology, and genomics of the Roseobacterclade: a short overview. *Arch Microbiol* **189:** 531–539.

Broberg, O., and Persson, G. (1988) Particulate and dissolved phosphorus forms in freshwater: composition and analysis. *Hydrobiologia* **170:** 61–90.

Carrigg, C., Rice, O., Kavanagh, S., Collins, G., and O.'Flaherty, V. (2007) DNA extraction method affects microbial community profiles from soils and sediments. *Appl Microbiol Biotechnol* **77:** 955–964.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283–1287.

Debroas, D., Humbert, J.F., Enault, F., Bronner, G., Faubladier, M., and Cornillot, E. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget – France). *Environ Microbiol* **11:** 2412–2424.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Downing, J.A., Prairie, Y.T., Cole, J.J., Duarte, C.M., Tranvik, L.J., Striegl, R.G., *et al.* (2006) The global abundance and size distribution of lakes, ponds, and impoundments. *Limnol Oceanogr* **51:** 2388–2397.

Duarte, C.M., and Cebrián, J. (1996) The fate of marine autotrophic production. *Limnol Oceanogr* **41:** 1758–1766.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Eiler, A., and Bertilsson, S. (2004) Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environ Microbiol* **6:** 1228–1243.

Eiler, A., and Bertilsson, S. (2007) Flavobacteria blooms in four eutrophic lakes: linking population dynamics of freshwater bacterioplankton to resource availability. *Appl Environ Microbiol* **73:** 3511–3518.

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

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., *et al.* (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of Bacteria, Archaea, Fungi, and viruses in soil. *Appl Environ Microbiol* **73:** 7059–7066.

Fine, P.V.A., and Kembel, S.W. (2011) Phylogenetic community structure and phylogenetic turnover across space and edaphic gradients in western Amazonian tree communities. *Ecography* **34:** 552–565.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41** (Database issue): D808–D815. doi:10.1093/nar/gks1094.

Garcia, S.L., McMahon, K.D., Martinez-Garcia, M., Sricastava, A., Sczyrba, A., Stepanauskas, R., *et al.* (2012) Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J* **7:** 137–147.

Gianoulis, T.A., Raes, J., Patel, P.V., Bjornson, R., Korbel, J.O., Letunic, I., *et al.* (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* **106:** 1374–1379.

Gilbert, J.A., Meyer, F., Schriml, L., Joint, I.R., Muhling, M., and Field, D. (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Stand Genomic Sci* **3:** 183–193.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., *et al.* (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242–1245.

Glöckner, F.O., Fuchs, B., and Amann, R. (1999) Bacterioplankton composition of lakes and oceans: a first comparson based on fluorescence in situ hybridization. *Appl Environ Microbiol* **65:** 3721–3726.

Goslee, S.C., and Urban, D.L. (2007) The ecodist package for dissimilarity-based analysis of ecological data. *J Stat Softw* **22:** i07.

Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., and Rappé, M.S. (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *Mbio* **3:** e00252-12. doi:10.1128/mBio.00252-12.

Hahn, M.W., Scheuerl, T., Jezberová, J., Koll, U., Jezbera, J., Šimek, K., *et al.* (2012) The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-living Polynucleobacter population. *PLoS ONE* **7:** e32772.

Ilikchyan, I.N., McKay, R.M.L., Zehr, J.P., Dyhrman, S.T., and Bullerjahn, G.S. (2009) Detection and expression of the phosphonate transporter gene phnD in marine and freshwater picocyanobacteria. *Environ Microbiol* **11:** 1314–1324.

King, G.M., and Weber, C.F. (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Microbiol* **5:** 107–118.

Kolmonen, E., Haukka, K., Rantala-Ylinen, A., Rajaniemi-Wacklin, P., Lepistö, L., and Sivonen, K. (2011) Bacterioplankton community composition in 67 Finnish lakes differs according to trophic status. *Aquat Microb Ecol* **62:** 241–250.

Kunin, V., Raes, J., Harris, J.K., Spear, J.R., Walker, J.J., Ivanova, N., *et al.* (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4:** 198.

Letunic, I., and Bork, P. (2011) Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39:** 475–478.

Lindström, E.S. (2000) Bacterioplankton community composition in five lakes differing in trophic status and humic content. *Microb Ecol* **40:** 104–113.

Lindström, E.S., Kamst-Van Agterveld, M.P., and Zwart, G. (2005) Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. *Appl Environ Microbiol* **71:** 8201–8206.

Logares, R., Bråte, J., Bertilsson, S., Clasen, J.L., Shalchian-Tabrizi, K., and Rengefors, K. (2009) Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol* **17:** 414–422.

Lozupone, C.A., and Knight, R. (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104:** 11436–11440.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32:** 1363–1371.

Martinez-Garcia, M., Swan, B.K., Poulton, N.J., Gomez, M.L., Masland, D., Sieracki, M.E., and Stepanauskas, R. (2012) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6:** 113–123.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112.

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., and Bertilsson, S. (2011) A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75:** 14–49.

Oh, S., Caro-Quintero, A., Tsementsi, D., Deleon-Rodriguez, N., Luo, C., Poretsky, R., and Konstantinidis, K. (2011) Metagenomic insights into the evolution, function and complexity iof the planktonic microbial community of Lake Lanier, a temperate freshwater system. *Appl Environ Microbiol* **77:** 6000–6011.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., *et al.* (2008) Vegan: Community Ecology Package. http://cran.r-project.org/web/packages/vegan/vegan.pdf

Peura, S., Eiler, A., Bertilsson, S., Nykänen, H., Tiirola, M., and Jones, R.I. (2012) Distinct and diverse bacterioplankton communities in the hypolimnion of boreal lakes are dominated by candidate division OD1. *ISME J* **6:** 1640–1652.

Quaiser, A., Zivanovic, Y., Moreira, D., and López-Garcia, P. (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5:** 285–304.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** D590–D594.

Raes, J., Koerbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8:** R10.

Raes, J., Letunic, I., Yamada, T., Jensen, L.J., and Bork, P. (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* **7:** 437.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., *et al.* (2011) Detecting novel associations in large data sets. *Science* **334:** 1518–1524.

Riesenfeld, C.R., Schloss, P.D., and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38:** 525–552.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al.* (2007) The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical pacific. *PLoS Biol* **5:** 398–431.

Schindler, D.W. (1978) Factors regulating phytoplankton production and standing crop in the world's lakes. *Limnol Oceanogr* **23:** 478–486.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007) Camera: a community resource for metagenomics. *PLoS Biol* **5:** e75.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML Web Servers. *Syst Biol* **57:** 758–771.

Storey, J.D. (2002) A direct approach to false discovery rates. *J R Stat Soc Ser B* **64:** 479–498.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., *et al.* (2003) The COG database: an update version includes eukaryotes. *BMC Bioinformatics* **4:** 41. doi:10.1186/1471-2105-4-41.

Tranvik, L., Downing, J., Cotner, J., Loiselle, S., *et al.* (2009) Lakes and reservoirs as regulators of carbon cycling and climate. *Limnol Oceanogr* **54:** 2298–2314.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261–5267.

Webb, C.O., Ackerly, D.D., McPeek, M.A., and Donoghue, M.J. (2002) Phylogenies and community ecology. *Annu Rev Ecol Syst* **33:** 475–505.

Webb, C.O., Ackerly, D.D., and Kembel, S. (2011). Phylocom: Software for the analysis of phylogenetic community structure and character evolution (with phylomatic and ecoevolve). User's manual, version 4.2 [WWW document]. URL http://www.phylodiversity.net/phylocom/

Weizhong, L. (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* **10:** 359. doi:10.1186/1471-2105-10-359.

Yannarell, A.C., and Triplett, E.W. (2005) Geographic and environmental sources of variation in lake bacterial community composition. *Appl Environ Microbiol* **71:** 227–239.

Zwart, G., Crump, B.C., Agterveld, M.P.K.V., Hagen, F., and Han, S.K. (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat Microb Ecol* **28:** 141–155.

Supporting information

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

Fig. S1. Boxplots depicting the GC % of reads from each metagenome.

Fig. S2. MEGAN classification into Bacteria, Archaea, Eukaryota and viruses.

Fig. S3. Non-metric multidimensional scaling plot comparing marine and freshwater metagenomes (stress-value = 0.10). This plot is based on Horn–Morisita distances from COGs abundance lists of 25 marine and freshwater metagenomes.

Fig. S4. Examples for phylogenetic trees of metagenomic sequences representing homologues of the nirK (A), pstA/B (B) and the nifH/bchL/chlL family (C). Trees were constructed by using the quick parsimony option (in ARB) to add aligned metagenomic sequences to RAxML master trees. For the 16S rRNA gene, the SILVA106 reference tree was used as the master tree. Blue indicates sequences obtained from marine, whereas red indicates samples from freshwater systems.

Table S1. Characteristics of marine metagenomes used for comparative analyses.

Table S2. Information about the number of raw reads and the removal of reads during the preprocessing steps. The last columns represent the number of reads and basepairs used for subsequent analyses.
Table S3. List of single copy core COGs) used for normalization (Ciccarelli *et al*., 2006; Raes *et al*., 2007).
Table S4. COGs that were significantly over- or under-represented in the freshwater metagenomes when compared with the marine metagenomes. Results from Wilcoxon test.
Table S5. COGs that were significantly related with total phosphorus concentrations in the freshwater metagenomes. Results from MINE (Reshef *et al*., 2011).