Title

Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 10(8)

ISSN

1751-7362

Authors

Eiler, Alexander

Mondav, Rhiannon

Sinclair, Lucas

et al.

Publication Date

2016-08-01

DOI

10.1038/ismej.2015.260

Peer reviewed

The ISME Journal (2016) 10, 1902–1914

www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria

Alexander Eiler[1], Rhiannon Mondav[1], Lucas Sinclair[1], Leyden Fernandez-Vidal[1], Douglas G Scofield[2], Patrick Schwientek[3], Manuel Martinez-Garcia[4,10], David Torrents[5,6], Katherine D McMahon[7,8], Siv GE Andersson[9], Ramunas Stepanauskas[4], Tanja Woyke[3] and Stefan Bertilsson[1]

[1]*Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden;* [2]*Department of Ecology and Genetics, Evolutionary Biology and Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden;* [3]*Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA;* [4]*Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA;* [5]*IRB-BSC Program in Computational Biology, Barcelona Supercomputing Centre, Barcelona, Spain;* [6]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain;* [7]*Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA;* [8]*Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA and* [9]*Department of Cellular and Molecular Biology, Molecular Evolution and Science for Life Laboratory, Uppsala University, Uppsala, Sweden*

Most free-living planktonic cells are streamlined and in spite of their limitations in functional flexibility, their vast populations have radiated into a wide range of aquatic habitats. Here we compared the metabolic potential of subgroups in the *Alphaproteobacteria* lineage SAR11 adapted to marine and freshwater habitats. Our results suggest that the successful leap from marine to freshwaters in SAR11 was accompanied by a loss of several carbon degradation pathways and a rewiring of the central metabolism. Examples for these are C1 and methylated compounds degradation pathways, the Entner–Doudouroff pathway, the glyoxylate shunt and anapleuretic carbon fixation being absent from the freshwater genomes. Evolutionary reconstructions further suggest that the metabolic modules making up these important freshwater metabolic traits were already present in the gene pool of ancestral marine SAR11 populations. The loss of the glyoxylate shunt had already occurred in the common ancestor of the freshwater subgroup and its closest marine relatives, suggesting that the adaptation to freshwater was a gradual process. Furthermore, our results indicate rapid evolution of TRAP transporters in the freshwater clade involved in the uptake of low molecular weight carboxylic acids. We propose that such gradual tuning of metabolic pathways and transporters toward locally available organic substrates is linked to the formation of subgroups within the SAR11 clade and that this process was critical for the freshwater clade to find and fix an adaptive phenotype.

The ISME Journal (2016) **10,** 1902–1914; doi:10.1038/ismej.2015.260; published online 19 January 2016

Introduction

Recent genome sequencing suggest that bacteria and archaea with streamlined genomes are often numerically dominant in nature, with marine and freshwater bacteria as prominent examples (Hahn *et al.*, 2012; Garcia *et al.*, 2013; Swan *et al.*, 2013; Eiler *et al.*, 2014; Giovannoni *et al.*, 2005a, b). Streamlined genomes are characterized by small size (<1.5 Mb), low guanine cytosine (GC) content (<40%) and short intergenic regions framing a restricted set of genes (Giovannoni *et al.*, 2014). The few and typically short intergenic regions, combined with a reduction in complexity of regulatory circuits, has posed the question about the capability of such streamlined cells and their populations to respond and adapt to environmental cues. Such genome minimalism has also been linked to unusual nutritional requirements, including auxotrophy for some vitamins and essential amino acids (Giovannoni *et al.*, 2005b; Tripp *et al.*, 2008, 2009; Schwalbach *et al.*, 2010; Carini *et al.*, 2013, 2014). Other apparent features are scarcity or absence of gene duplications,

Correspondence: A Eiler or S Bertilsson, Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Norbyvägen 18D, Uppsala 75236, Sweden.
E-mail: alexander.eiler@ebc.uu.se or stebe@ebc.uu.se
[10]Current address: University of Alicante, Alicante, Spain.
Received 28 July 2015; revised 24 November 2015; accepted 10 December 2015; published online 19 January 2016

extrachromosomal elements and remnants of genes and phages (Giovannoni *et al*., 2014). The proposed selective drive for this minimalism is efficient use of limiting resources (for example, C, N, P) in nutrient-poor systems (Giovannoni *et al*., 2014), which can cause genome reduction when combined with large effective population sizes (Lynch, 2007).

These features, particularly the reduced genetic repertoire of individual cells, raises the question of how such organisms can maintain sufficient metabolic flexibility to radiate or possibly cross biome boundaries. A potential explanation is provided by recent genomic analyses of Prochlorococcus where bacterial populations with streamlined genomes were shown to contain considerable sequence variation in shared genes and to harbor a high level of heterogeneity in gene content (Kashtan *et al*., 2014). A combination of point mutations and homologous recombination were suggested to be responsible for most of the allelic diversity, whereas homologous and non-homologous recombination were implicated in the loss and gain of genes. The resulting combined diversity of streamlined genomes in these large populations thus could provide the means to cross biome boundaries, such as moving from oceans to freshwaters, with dramatic differences in solutes, salinity and resources, thereby forming a strong ecological and evolutionary barrier.

Similar to larger organisms, there is little overlap in the taxonomic composition of marine and freshwater microbial assemblages (Suzuki and DeLong, 2002; Logares *et al*., 2009; Newton *et al*., 2011). Nevertheless, phylogenetic analyses have uncovered some surprisingly close relatives among marine and freshwater microbiota (Logares *et al*., 2009; Newton *et al*., 2011). Both marine and freshwater bacterioplankton communities are frequently dominated by alphaproteobacteria of the SAR11 lineage (Morris *et al*., 2002; Rusch *et al*., 2007; Eiler *et al*., 2009; Newton *et al*., 2011; Salcher *et al*., 2011; Heinrich *et al*., 2013). Members of the marine SAR11 clades are characterized as slow growing oligotrophs (Giovannoni *et al*., 2014), and they are believed to have a central role in the global cycling of both carbon, sulfur and associated elements (Malmstrom *et al*., 2004). Among their adaptations to oligotrophic marine environments are proteorhodopsin-mediated phototrophic potential (Giovannoni *et al*., 2005a), reliance on externally supplied reduced sulfur (Tripp *et al*., 2008), auxotrophy for essential amino acids (Tripp *et al*., 2009) and lineage-defined metabolic restrictions for a wide range of labile low molecular weight organic metabolites, including C1 compounds (Schwalbach *et al*., 2010; Sun *et al*., 2011; Carini *et al*., 2013). Freshwater SAR11 are phylogenetically distinct from their marine relatives and group into the uncultivated clade named 'LD12' (Zwart *et al*., 1998). LD12 is recognized as a globally distributed freshwater group (Newton *et al*., 2011) that can make up 20–40% of the total bacterioplankton in lakes (Salcher *et al*., 2011; Heinrich *et al*., 2013), and isotope tracer experiments have verified that they can grow heterotrophically (Salcher *et al*., 2011).

Here we compared the metabolic potential of 10 sequenced single amplified LD12 genomes obtained from freshwater lakes (Zaremba-Niedzwiedzka *et al*., 2013) with 18 marine SAR11 genomes, 4 single amplified genomes and 14 from isolates and originating from different coastal and oceanic habitats (Giovannoni *et al*., 2005b; Tripp *et al*., 2009; Grote *et al*., 2012; Viklund *et al*., 2013). Previous comparisons of these genomes have revealed lower recombination frequencies in the freshwater LD12 genomes (Zaremba-Niedzwiedzka *et al*., 2013) but not detailed the metabolic changes associated with the transition from marine to freshwater environments. Our analyses reveal a reduction of metabolic diversity and lower pathway redundancy in freshwater LD12 compared with marine SAR11, combined with a rewiring of the central metabolism. We propose that the process leading up to the marine–freshwater transition in SAR11 was more likely facilitated by recombination in the large and diverse populations of the marine ancestors than by gene family expansion or horizontal acquisition of metabolic modules from non-SAR11 lineages. Thus our results unveil an unknown chapter in the evolutionary history of these ubiquitous and successful bacterial groups in aquatic environments.

Methods

Genome data sets

The draft assemblies of the single amplified genomes (SAGs) and genome-sequenced SAR11 isolates were obtained from intergrated microbial genome (IMG)/ER (Markowitz *et al*., 2012). Protein sequences for the recently described pathway of dimethylsulfoniopropionate (DMSP) demethylation were retrieved for *Ruegeria pomeroyi* and HTCC1062 as identified by Reisch *et al.* (2011). These were used as queries in BLASTP searches against the 18 marine and 10 freshwater SAR11 genomes.

Genome completeness

Genome size and completeness were estimated using a conserved single-copy gene (CSCG) set of SAR11 isolates (Markowitz *et al*., 2012). The set consists of 357 CSCGs that were found to occur only once in the genomes by analysis of an abundance matrix based on hits to the protein family (Pfam) database (Punta *et al*., 2012). Hidden Markov models of the identified Pfams were used to search all single amplified genome assemblies by means of the HMMER3 software (Eddy, 2011). Resulting best hits above the trusted cutoff (TC field as provided in the HMM files from Pfam) were counted, and the completeness was estimated as the ratio of found CSCG to total CSCGs in the set. The estimated complete genome size was then calculated by dividing the estimated genome coverage by the total assembly size.

Phylogenetic reconstructions

Orthologous genes were identified by running a BLASTP similarity search using the set of all marine and freshwater genes as the query and the database. Subsequently, we applied the Markov cluster algorithm (Enright *et al.*, 2002) to the pairwise matrix of bit scores. Out of the 3598, we selected the clusters that had at least one gene pertaining to each of the seven families ('Ia.1', 'II', 'Ia.2', 'V', 'IIIa', 'Ic', 'IIIb/LD12') and no more than 30 genes in them, leaving us with 518 clusters. Genes were then aligned with muscle (Edgar, 2004), masked with gblocks (Castresana, 2000) and trees were built with RaxML (Stamatakis, 2006) using the 'PROTGAMMAJTT' model with an automatically determined number of quick bootstraps, thus resulting in one tree per cluster. At this point, we removed one cluster as it contained only gaps.

For the master ribosomal protein tree, we extracted 50 ribosomal proteins and subsequently selected only those that had at least one representative in each SAR11 subgroup ($N=40$). After alignment of randomly selected representatives from each SAR11 subgroup, we concatenated the alignments to form the so-called 'master ribosomal alignment'. A RAxML tree was constructed using the 'PROTGAMMAJTT' model with an automatically determined number of quick bootstraps. The code produced for this analysis is available under an MIT license at http://github.com/limno/ld12.

Closest hits and gene family expansions

The first hit for every freshwater gene is obviously itself (though some genes have no other hits). Searching for the first hit that is not from any of the freshwater genomes, we extracted its taxonomy. If this top-hit-non-fresh is a marine SAR11 genome, we resolve it to one of the six marine clades; if outside, we resolve it to broad taxonomic groups. Duplication events (gene family expansions) were defined by two hits associated with the same freshwater genome and ranked prior to the first non-freshwater SAR11 hit.

Results

General features and SAR11 phylogeny

The freshwater and marine SAR11 genomes were extracted from public databases (Table 1). The genomes are in varying stages of completion, with some corresponding to a single circular chromosome whereas others are incomplete and consist of 50–150 contigs. The completeness of the genomes amplified from single cells was estimated to 24–86% using a conserved single-copy gene set of SAR11 genomes (Markowitz *et al.*, 2012), with estimated genome size of the freshwater genomes ranging from 1.1 to 1.4 Mb compared with marine genomes ranging from 1.1 to 1.6 Mb. In addition to these small differences, the sampling sizes for subgroups II and IIIa are too small to determine with certainty whether differences in genome size are associated with subgroups and habitats. Unifying features are low GC (29–30%), a shared amino-acid usage and high percentage of predicted coding nucleotides (Table 1).

Phylogenetic analysis of the selected genomes using concatenated alignment of 40 ribosomal proteins revealed 7 clades (Figure 1), consistent with previous analyses (Grote *et al.*, 2012; Viklund *et al.*, 2013; Zaremba-Niedzwiedzka *et al.*, 2013; Thrash *et al.*, 2014). In this study, we have not included any other alphaproteobacterial species, which prevents a discussion about the position of subgroup V, represented by genome HIMB59, in the alphaproteobacterial tree. The disagreement about its phylogenetic placement was discussed in quite some detail in a previous publication (Viklund *et al.*, 2013), but this does not have any implications for our conclusions. Suffice it to conclude that HIMB59 is used as an outgroup to the other SAR11 genomes examined here. By using HIMB59 as an outgroup, we inferred phylogenetic trees from 518 orthologous protein clusters that had at least one representative in each of the 7 subclades. Out of these, only a minor proportion ($N=88$) of the protein trees are strictly coherent with regards to collapsing the 7 main clades (that is, all genes from a given branch are monophyletic). For a third of these ($N=33$), the branching patterns agree with the ribosomal protein tree, and the branching pattern between the freshwater LD12 (SAR11 clade IIIb) and clade IIIa is observed in 388 out of the 518 trees, albeit with different statistical significance (representing 75% of all trees as indicated in Figure 1).

Comparing this with the number of cases where the bootstrap support was $>80\%$ at this particular branch, freshwater LD12 branched closest with IIIb in 109 trees, whereas in 9 trees LD12 was closest with another SAR11 subgroup. These nine trees were inferred from individual protein alignments involved in amino-acid biosynthesis and conversion (Figures 2a–c), chemoorganotrophic energy acquisition (Figures 2d–f), carboxylic acid membrane transport (Figures 2g and h) and DNA methylation (Figure 2i). This proportion of non-matching tree topology, either 9% or 25% depending on bootstrap cutoffs (Figure 1; see a report detailing the mismatches and the Robinson–Foulds metric in Supplementary Table S1), implies alternative evolutionary histories of various gene families in SAR11.

Two trees are largely consistent with the ribosomal protein tree with the exception of the branching orders of subgroups II and III (Figures 2b and d). Another three trees have more serious inconsistencies, such that subgroups Ia and Ic are not sister clades and that the two members of subgroup IIIa do not cluster (Figures 2a, c, e, g and i). The most dramatic inconsistency is observed for the tree inferred from the tripartite ATP-independent periplasmic transporters (TRAP transporters) (Figures 2h

Table 1 Description of the genome data

Name	*IMG accession no.*	*Group*	*Biome*	*Assembled nucleotides*	*Com (fraction)*	*EGS*	*Gene count*	*Estimated gene count*	*Scaffold count*	*GC (fraction)*	*Coding bases %*	*RNA count*
HTCC1002	638341056	1a.1	Marine	1327604	0.99	1346462	1428	1448	5	0.3	95.9	35
HTCC1013	2511231042	1a.1	Marine	1302704	1.00	1302704	1411	1411	22	0.3	96.6	36
HTCC1062	637000058	1a.1	Marine	1308759	1.00	1308759	1394	1394	1	0.3	96.1	40
HTCC9565	2503283022	1a.1	Marine	1279674	0.99	1294174	1386	1402	3	0.29	96.2	34
HIMB083	2510461079	1a.2	Marine	1395997	1.00	1395997	1505	1505	1	0.29	96.8	37
HIMB140	2503754000	1a.2	Marine	1437930	1.00	1437930	1535	1535	1	0.29	97.0	31
HIMB5	2503982040	1a.2	Marine	1343202	1.00	1343202	1467	1467	1	0.29	96.8	36
HTCC7211	2503283017	1a.2	Marine	1456888	1.00	1456888	1481	1481	1	0.29	96.7	34
HTCC8051	2511231043	1a.2	Marine	1395115	0.99	1410923	1498	1515	2	0.29	96.8	36
HTCC9022	2511231041	1a.2	Marine	1362422	0.99	1381774	1464	1485	24	0.3	96.5	37
SCGCAAA288-G21	2236347014	1c	Marine	909786	0.64	1421541	1103	1723	139	0.3	94.4	29
SCGCAAA288-N07	2236347020	1c	Marine	954664	0.66	1446461	1110	1682	81	0.29	94.8	27
SCGCAAA240-E13	2236661011	1c	Marine	1401625	0.86	1629797	1621	1885	151	0.29	93.6	40
SCGCAAA288-E13	2236661012	1c	Marine	812863	0.55	1477933	948	1724	106	0.29	93.6	25
HIMB058	2265129005	2	Marine	1115049	0.99	1130887	1269	1287	59	0.3	96.9	39
HIMB114	2503283019	3a	Marine	1237371	1.00	1237371	1357	1357	1	0.3	97.0	36
IMCC9063	650716017	3a	Marine	1284727	1.00	1284727	1482	1482	1	0.32	94.9	35
SCGCAAA023-L09	2236661000	3b/LD12	Freshw	774923	0.68	1147914	921	1364	76	0.29	96.0	29
SCGCAAA024-N17	2236876027	3b/LD12	Freshw	328144	0.24	1394612	397	1687	45	0.3	95.5	14
SCGCAAA027-C06	2264265094	3b/LD12	Freshw	775384	0.72	1072915	936	1295	90	0.3	95.9	20
SCGCAAA027-J10	2236876030	3b/LD12	Freshw	792980	0.73	1080510	952	1297	82	0.3	96.0	31
SCGCAAA027-L15	2236876031	3b/LD12	Freshw	719587	0.63	1141744	840	1333	56	0.29	96.5	19
SCGCAAA028-C07	2236661008	3b/LD12	Freshw	846566	0.76	1111117	974	1278	32	0.3	96.3	19
SCGCAAA028-D10	2236347069	3b/LD12	Freshw	925141	0.78	1192329	1091	1406	57	0.3	95.6	34
SCGCAAA280-B11	2236876032	3b/LD12	Freshw	674250	0.54	1260247	815	1523	47	0.3	95.8	18
SCGCAAA280-P20	2236876029	3b/LD12	Freshw	720523	0.68	1062920	838	1236	65	0.3	95.6	24
SCGCAAA487-M09	2236347068	3b/LD12	Freshw	627365	0.61	1027382	800	1310	97	0.29	96.1	24
HIMB59	2503982041	5	Marine	1410127	1.00	1410127	1532	1532	1	0.32	96.8	39

Abbreviations: Com, completeness; EGS, estimated genome size; Freshw, freshwater. Genomes including 'SCGCAAA' in the names are single amplified genomes while others are derived from isolates.

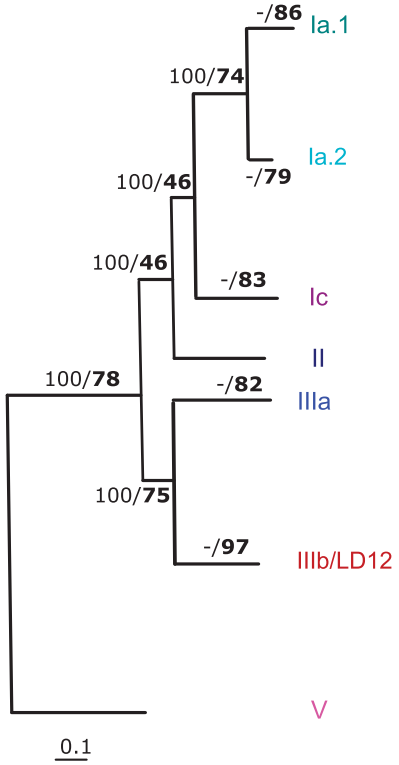

Figure 1 Evolutionary relationships between marine and freshwater SAR11 genomes. Unrooted maximum likelihood phylogenetic tree of concatenated ribosomal protein sequences from single-cell genomes and isolate genomes of the SAR11 clade. In addition to bootstrap values as inferred by maximum likelihood, the second (large and bold) number at the branching points show proportion of protein trees with corresponding branching pattern. These proportions were generated by using the Robinson–Foulds metric summarizing 518 orthologous protein trees that had at least one representative in each of the seven subclades (for details, see Supplementary Table S1). Freshwater SAR11 (subgroup IIIb/LD12) are indicated in red, whereas marine SAR11 subgroups in other colors.

and i). In these trees, all members of subgroup LD12 and two genomes of subgroup Ia are separated from all other genomes by long branches. This could be indicative of paralogous gene families or increased evolutionary rates for this gene in LD12. This is potentially interesting as TRAP transporters may be involved in the metabolite flux into the tricarboxylic acid (TCA) cycle through the import of malate (Adnan *et al.*, 2015).

Apart from the incongruence of protein trees, patterns in gene synteny (Supplementary Figure S1) and functional annotations (Supplementary Figure S2) suggest that the analyzed SAR11s group into six (plus outgroup V) distinct clades, basically confirming shared gene phylogenies (Grote *et al.*, 2012; Viklund *et al.*, 2013; Zaremba-Niedzwiedzka *et al.*, 2013; Thrash *et al.*, 2014). The distinction in gene sets between marine and freshwater genomes was shown by permutational multivariate analysis of variance ($P<0.001$) for four annotation subsystems (pfam, tigrfam, COG, KO) and also by dendrograms based on bit scores retrieved from pairwise blastn searches (Supplementary Figure S3). In addition, each of these subgroups is preferentially retrieved in a distinct habitat, indicating that they correspond to ecological coherent units (Rusch *et al.*, 2007; Andersson *et al.*, 2009; Carlsson *et al.*, 2009; Newton *et al.*, 2011; Heinrich *et al.*, 2013; Thrash *et al.*, 2014).

Energy and carbon metabolism

We identified genes encoding a complete TCA cycle (Figure 3) and a proteorhodopsin when combining the 10 partial freshwater SAR11 genomes. This led us to hypothesize that, analogously to their marine siblings, freshwater SAR11 are heterotrophs capable of supplementing heterotrophic growth on organic carbon with light-mediated ATP production (Martinez-Garcia *et al.*, 2012) when carbon starvation limits respiration (Steindler *et al.*, 2011). Still, there are conspicuous and consistent variations in gene content related to aerobic chemoorganoheterotrophy as already highlighted by previous metabolic reconstruction restricted to marine SAR11 (Tripp *et al.*, 2008, 2009; Schwalbach *et al.*, 2010; Sun *et al.*, 2011; Grote *et al.*, 2012; Carini *et al.*, 2013, 2014). Here we expand on these previous observations by showing that freshwater genomes encode all genes of the Embden–Meyerhof–Parnas (EMP) pathway, whereas most marine SAR11 representatives encode all genes of the Entner–Doudoroff (ED) pathway (Figure 3).

Another striking difference is the absence of the glyoxylate shunt (isocitrate lyase (aceA), malate synthase (malG)) and anapleurotic enzymes (phosphoenol pyruvate carboxylase (ppc); see Figure 3) in subgroups IIIa and IIIb, pointing to major differences in carbon and energy-acquisition strategies for these SAR11 clades. The glyoxylate shunt bypasses two steps in the TCA cycle where carbon dioxide is released concomitant with a production of ATP and NADH but instead enable economic use of carbon atoms at the expense of energy acquisition. It essentially allows microorganisms to use simple carbon compounds as a carbon source when complex nutrients such as glucose are not available. In the absence of available carbohydrates, the glyoxylate cycle would, for example, enable synthesis of carbohydrates needed for cell-wall assembly from lipids via acetate supplemented with inorganic carbon fixed via anapleurotic reactions (Moran and Miller, 2007).

There is evidence that marine SAR11 can oxidize several small organic carbon compounds that are potentially very abundant in marine systems, in order to satisfy cellular energy demands (Tripp *et al.*, 2009; Schwalbach *et al.*, 2010; Sun *et al.*, 2011). Several genes and pathways involved in the oxidation of C1 and methylated compounds to CO_2, previously identified in marine SAR11 genomes (Sun *et al.*, 2011), appear to be missing in the freshwater single amplified genomes (see Supplementary Material for details). At the same

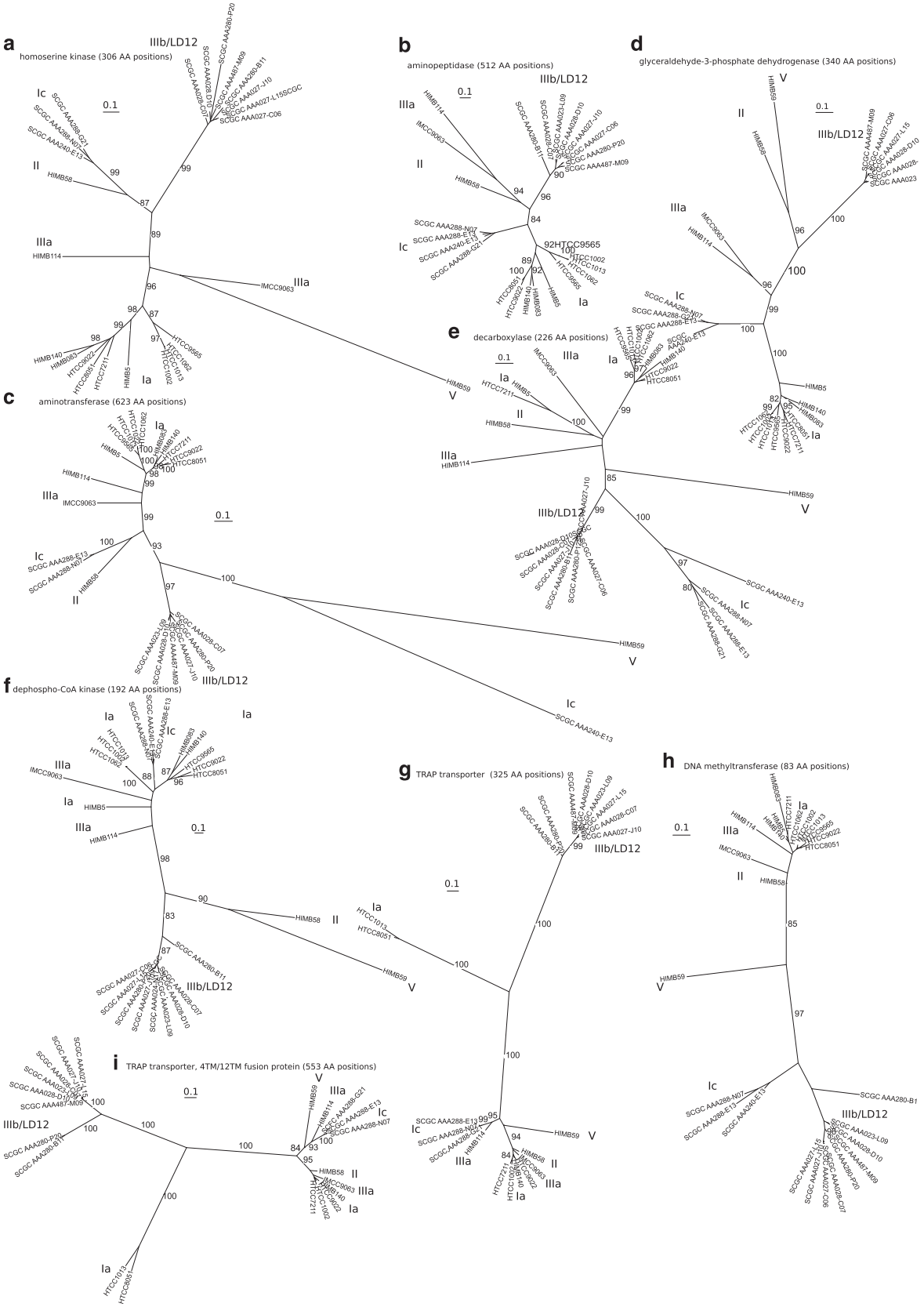

Figure 2 Nine protein trees with evolutionary inconsistency when compared with the ribosomal tree. (**a**–**i**) Unrooted phylogenetic trees of orthologous protein clusters with alternative tree topologies (that is, sister clade to IIIb/LD12 is not IIIa) when compared with the ribosomal tree (Figure 1).

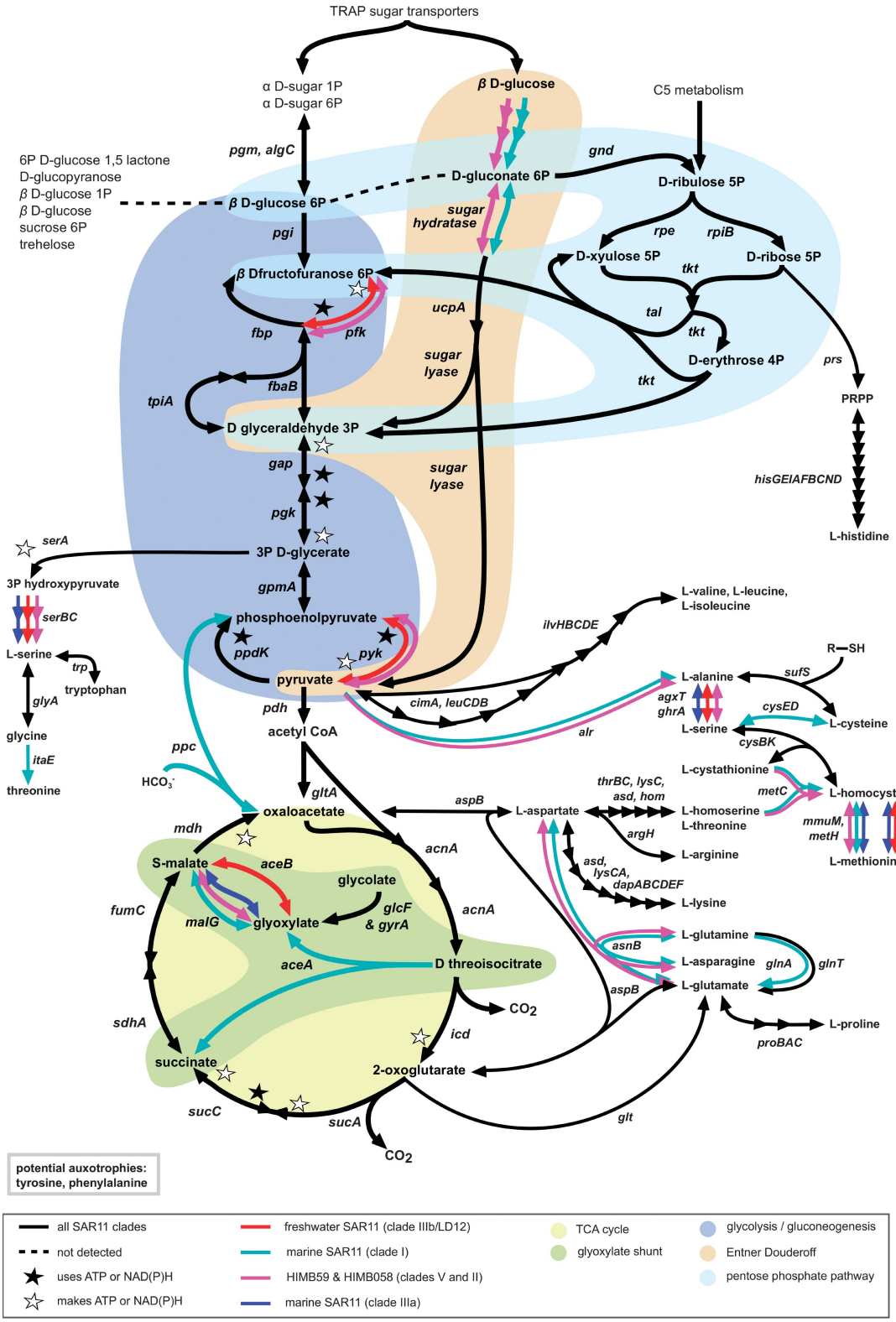

Figure 3 Central carbon metabolism and other relevant metabolic pathways identified in SAR11 genomes. Glycogenesis, TCA cycle and glyoxylate shunt are shown in the center of the plot with adjunction pathways to either side such as the pentose phosphate and ED pathway. The color of the arrows indicates genes encoded in at least one genome of each SAR11 subgroup (see legend for details). The presence and absence of genes was determined within the IMG system based on automated and manual annotations. Detailed results are in Supplementary Tables S6–S9.

time, the ability to use other common low molecular weight carbon substrates are maintained throughout the entire SAR11 lineage, one example being glycolate oxidase (glcDEF) which enable metabolic use of the abundant exudate glycolate, which is released in large amounts by nutrient-limited phytoplankton as an energy overflow mechanism (Bertilsson and Jones, 2003).

While inspecting the completeness of vitamin biosynthesis pathways among the freshwater genomes, we found no major differences when comparing our results with marine genomes (for details on the presence of vitamin biosynthesis homologs, see Supplementary Material). Confirming previous findings (Tripp *et al.*, 2008), we could not identify any sulfite reductase in marine SAR11 genomes nor could this gene be detected in any of the freshwater representatives, implying that freshwater SAR11 also rely on external supply of reduced sulfur similar to their marine siblings. This auxotrophy for reduced sulfur seems to be a common feature in abundant freshwater bacteria as revealed by genome reconstructions (Garcia *et al.*, 2015). To access reduced sulfur, marine SAR11 are capable of demethylating the abundant marine molecule DMSP. However, homologs of the DMSP pathway genes could not be identified in the 10 freshwater (LD12) or clade IIIa genomes. In addition, we could not identify any homologs to adenylylsulfate reductase, sulfate permease and sulfate exporter in the freshwater single amplified genomes, making sulfur-containing amino acids the most likely reduced sulfur source.

Osmolytes, such as DMSP, glycine betaine, proline, mannitol, taurine and spermidine, are produced in high amounts by phytoplankton and other marine microplankton, which use these compounds to maintain their intracellular osmotic balance. These organic compounds can build up intracellular concentrations as high as 100–530 mM in some phytoplankton species (Keller *et al.*, 1999; Sunda *et al.*, 2007). Analyses of marine SAR11 genomes (Giovannoni *et al.*, 2005b) have revealed numerous ABC transporters for organic osmolytes that are common in the marine environment (see Supplementary Material for detail on transporters). An ability to use glycine betaine and sulfur-containing osmolytes may also release organisms from an inability to synthesize certain amino acids and reduced (organic) sulfur compounds *de novo* (Giovannoni *et al.*, 2014). Although glycine and serine biosynthesis is enabled by 3-phosphoglycerate dehydrogenase (*serA*) in freshwater SAR11 (for more detail on amino-acid biosynthesis, see Supplementary Material), organosulfur acquisition seems to be restricted to amino acids, such as methionine and cysteine. This is not surprising considering that osmolyte diversity is restricted to non-sulfur compounds such as sucrose and trehalose in freshwater phytoplankton (Batterton and van Baalen, 1971; Joset *et al.*, 1996; Page-Sharp *et al.*, 1999). Still, no known hydrolases to use sucrose and trehalose could be identified in any of the inspected freshwater SAR11 genomes. As indicated by the presence of TRAP system, a feature shared with subgroup IIIa, freshwater SAR11 most likely use low-molecular-weight carboxylic acids.

Interestingly, two TRAP system-related orthologous gene clusters showed phylogenetic patterns highly divergent from the ribosomal protein trees (Figures 2g and h), suggesting that this transporter system has been widely exchanged within the SAR11 clade.

The marine and freshwater SAR11 pangenome

In addition to phylogenies (see Figures 1 and 2), we explored gene content variations among the SAR11 genomes. Orthologous protein clusters were formed by Markov cluster algorithm (Enright *et al.*, 2002) from pairwise BLASTP searches using 23 545 predicted SAR11 proteins. Of the 3589 total orthologous clusters identified in the 28 SAR11 genomes, 1467 were present in at least one freshwater SAG, and 368 of these were exclusive to freshwater genomes. Most of the 8332 predicted genes from the freshwater SAGs were most closely related to marine subgroup IIIa (Figure 4a). Out of the 368 orthologous clusters (including 720 genes) exclusive to freshwater SAR11, 59 clusters (including 86 genes) had no hit against bacterial and archaeal subject sequences in refseq. Altogether, 94 of the freshwater-exclusive genes (26%) were most closely matching alphaproteobacteria, whereas only 7 (2%) were most closely related to genes from marine SAR11 (Figure 4b). Although around a sixth of the predicted genes from the freshwater SAGs were most closely related to non-marine SAR11, it cannot be concluded that these genes were acquired horizontally from outside the SAR11 clade. The reason is that the origin and flux of these genes are uncertain when considering that protein diversity was highly undersampled in both the marine and freshwater branch of SAR11, highlighted by the rarefaction analysis (Figure 4c). The large diversity in protein families is further imposed by 6273 protein families in the marine and 2320 in the freshwater SAR11 pangenome as estimated by non-parametric diversity approximations (Chao, 1987).

BLASTP searches were also used to infer gene family expansions unique to the freshwater SAR11 population by applying the following criteria: two hits associated with the same genome were ranked prior to the first non-freshwater SAR11 protein. This identified at most four orthologous clusters per freshwater genome, thus likely originating from recent duplication events as they were also adjacent to each other in the genome. These orthologous clusters were exclusively annotated as hypotheticals or cell communication systems (for details, see Supplementary Table S5) and are likely located in hypervariable regions. Such hypervariable regions are apparent in a composite freshwater SAR11

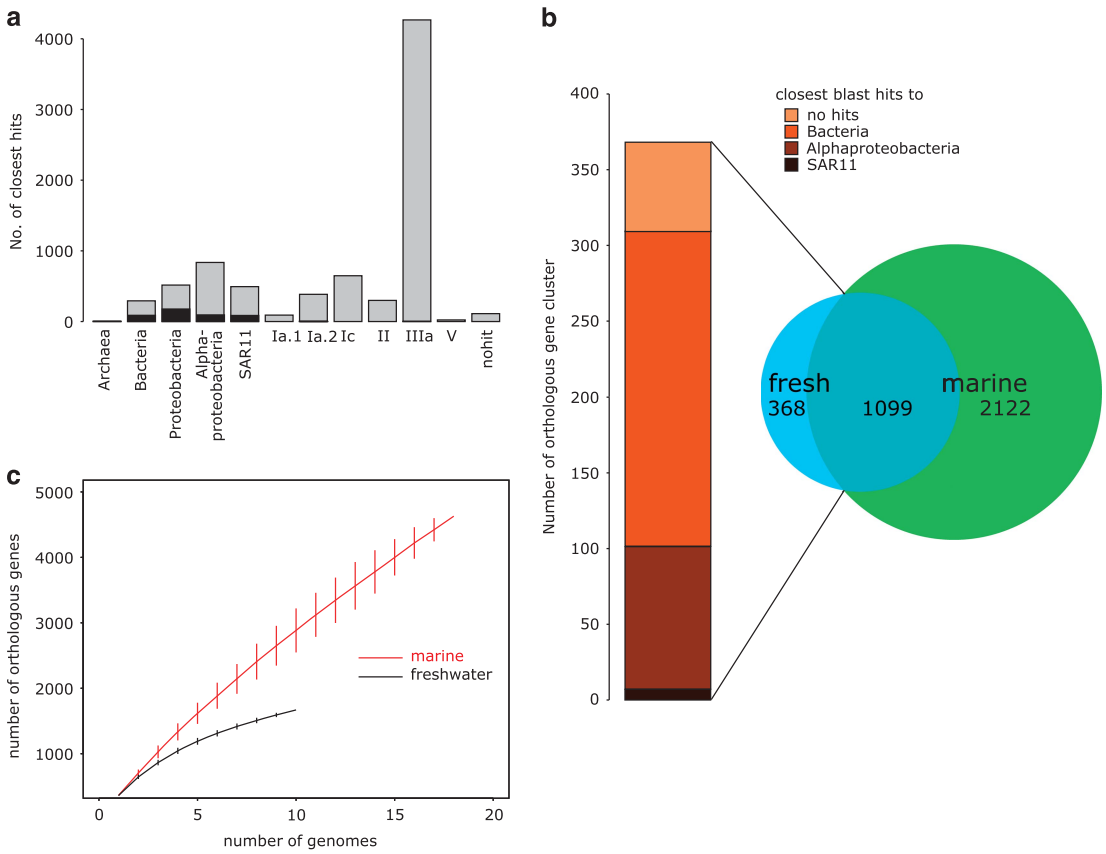

Figure 4 Closest blast hits against refseq and rarefaction curves of orthologous SAR11 protein clusters. (**a**) Bar charts revealing top hits from BLASTP against all available genomes in refseq. Hits are classified based on their taxonomic affiliation with higher taxonomic levels removed from lower taxonomic levels (that is, SAR11 hits are removed from the alphaproteobacterial bin, with alphaproteobacterial hits removed from the proteobacterial bin and so on). Black bars indicate hits of genes specific to freshwaters as inferred by orthologous clustering. (**b**) Detailed representation of freshwater-specific orthologous protein clusters and their closest hits classified on their taxonomic affiliation with higher taxonomic levels removed from lower taxonomic levels. (**c**) Results from a rarefaction analysis of orthologous protein clusters after normalization to account for the partial single-cell genomes.

genome and seem to host genes for the biosynthesis of lipopolysaccharides (Zaremba-Niedzwiedzka *et al.*, 2013). The comparative features of these 28 genomes point to the existence of large gene frequency variations in SAR11 populations.

Discussion

As shown previously (Zaremba-Niedzwiedzka *et al.*, 2013), phylogenetic reconstructions strongly suggest that the freshwater SAR11 clade originates from a population of marine ancestors with streamlined genomes considering the branching pattern leading to LD12. However, we cannot preclude with certainty that IIIa resembles a marine ancestor of LD12 or that IIIa originated from a freshwater ancestor. Nevertheless, we favor the former scenario as a freshwater origin would require two independent transitions, whereas one is sufficient for a marine SAR11 ancestor. In addition, the larger estimated marine gene set when compared with the freshwater equivalents would have required an acquisition of genes independent from the freshwater SAR11. The likelihood for such a vast gene acquisitions is challenged by large population sizes that provide low probabilities of new genes arising and being fixed in streamlined genomes. Likewise, a common streamlined ancestor is implied by freshwater and marine SAR11 clades sharing many typical features of streamlined cells, such as the small cell size (Giovannoni *et al.*, 2005b; Salcher *et al.*, 2011), auxotrophy, small genome size and low GC content. With no known exceptions, this argues against non-streamlined ancestoral populations at the main branching points.

Besides the shared reductionism, both marine and freshwater SAR11 are predicted to host a typical electron transport chain, a complete TCA cycle and a proteorhodopsin system for phototrophic energy acquisition (Giovannoni *et al.*, 2005a; Martinez-Garcia *et al.*, 2012). Still, there are conspicuous variations in gene content related to aerobic chemoorganoheterotrophy between marine and freshwater SAR11, emphasizing functional adaptations beyond osmoregulation as a necessity

to overcome the marine–freshwater boundaries. This is in agreement with recent findings from shotgun metagenomic analyses, where quantitative differences related to respiratory and glycolytic pathways were observed in the gene pools of marine and freshwater communities (Oh *et al.*, 2011; Dupont *et al.*, 2014; Eiler *et al.*, 2014). In particular, the changes in glycolysis combined with the apparent loss of the glyxoylate shunt and the capacity to oxidize a broad array of C1 compounds point to differentiation in heterotrophic substrate use between marine and freshwater SAR11, with the marine representatives more extensively relying on phytoplankton-derived osmolytes, such as proline, betaine glycine and DMSP. Instead, freshwater SAR11 feature a complete EMP, enabling efficient energy acquisition from C6 and C5 sugars combined with the TRAP system to use low-molecular-weight carboxylic acids known to be produced in abundance by photochemical degradation of humic substances in freshwaters (Bertilsson and Tranvik, 2000). Such tuning toward locally available organic substrates is likely linked to observed differences in abundance patterns in relation to environmental conditions in the respective biomes (Eiler *et al.*, 2009; Salcher *et al.*, 2011) and the formation of ecotypes within the SAR11 clade (Carlsson *et al.*, 2009; Vergin *et al.*, 2013).

The reliance on either EMP or ED can be predicted to change the energy status and the redox balance of the cell. The ED pathway (Entner and Doudoroff, 1952; Conway, 1992) has a net yield of 1 ATP, 1 NADH and 1 NADPH for each glucose molecule, whereas the EMP has a net yield of 2 ATP and 2 NADH for each processed glucose molecule. Besides the distinct yields of ATP and NAD(P)H, the EMP and ED pathways result in a range of different metabolic intermediates for each molecule of sugar consumed (Neidhardt *et al.*, 1990). It is remarkable that, while some bacterial genomes possess complete and coexisting EMP and ED pathways (suggested for HIMB59), others have one or the other, and there are also SAR11 genomes that only seem to host incomplete pathways (Grote *et al.*, 2012). A flux of genes involved in central metabolism is further implied by the incongruence in tree topology of the glyceraldehyde 3-phosphate dehydrogenase (Figure 2d), the enzyme that catalyzes the sixth step of glycolysis, a phosphorylation reaction coupled to oxidation.

In addition to the metabolic distinction between marine and freshwater SAR11, our analyses revealed that the gene pool of marine and freshwater SAR11 are highly diverse and undersampled. The notion that marine SAR11 populations appear to have rather diverse genomes was previously observed in marine metagenomes (Wilhelm *et al.*, 2007) and was also shown for *Prochlorococcus* as another abundant and widespread marine bacterium with small genomes (Kashtan *et al.*, 2014). Overall, we show that SAR11 subgroups share a core set of genes comprising approximately half of the orthologous genes in each genome, independent of whether orthologous genes are defined based on annotations or clustering using sequence similarity. As expected, the core genes encode functions defining the SAR11 clade as an organo-heterotrophic group with varying potential for photo-autotrophy, whereas the flexible genome provides functions that confer selective advantages under very different conditions. This modular structure of genes that confer the ability to take up and metabolize specific organic molecules imply metabolic flexibility and an adaptive tuning to resources.

This functional tuning within SAR11 is facilitated by high rates of recombination (Vergin *et al.*, 2007) among even distantly related marine groups fostering a global SAR11 population where multiple ecotypes with high genetic linkages coexist. A dynamic gene pool in SAR11 is implied by gene content variations, alternative tree topologies of the orthologous clusters shared among all SAR11 subgroups and, as shown previously, by high ratios of homologous recombination to mutation rate in marine SAR11 (Vergin *et al.*, 2007). Even if extremely high recombination has been suggested for marine SAR11 (Vergin *et al.*, 2007; Zaremba-Niedzwiedzka *et al.*, 2013), we argue that the exchange of genetic material seems to be mainly restricted to members within the SAR11 clade, as suggested by a common SAR11 ancestor for the orthologous genes that were present in all marine and freshwater SAR11 genomes. As the marine gene pool is undersampled, we argue that one cannot make any inferences on horizontally acquired genes from outside the SAR11 clade to facilitate the marine–freshwater transition. A likely scenario is that, as SAR11 gradually became adapted to life at low salinities, recombination events with divergent SAR11 in the former marine environment became less likely and eventually mutations overwhelmed recombination, as indicated previously (Zaremba-Niedzwiedzka *et al.*, 2013).

Resulting gene content differences that affect biosynthesis and transport functions are not restricted to specific regions but are instead embedded in different genomic contexts, as implied by the large-scale synteny variations in the SAR11 genomes. Besides these ecotype-defining and broad-scale genomic variations, hypervariable regions containing genes for the biosynthesis of lipopolysaccharides were previously identified in both marine and freshwater SAR11 clades (Grote *et al.*, 2012; Zaremba-Niedzwiedzka *et al.*, 2013). This fits the previously proposed concept of low frequency genes being implicated in evolutionary responses to local and variable biotic interactions, such as competition and phage predation (Cordero and Polz, 2014).

Adaptive selection that can lead to niche partitioning and the formation of novel ecotypes in bacterial populations is assumed to depend on a wide array of traits, including large population size (Levin and Bergstrom, 2000), a dynamic gene pool facilitated

by genetic exchange (Cohan, 2001; Popa *et al.*, 2011) and gene family expansion by duplications (Alm *et al.*, 2006; Serres *et al.*, 2009), the small size of recombined fragments (Zawadzki and Cohan, 1995) and simplicity and modularity of metabolic features (Doyle *et al.*, 2007; Lawrence, 1999). In the long run, these mechanisms can allow lineages to take major leaps into novel environments or even new biomes, even when populations are made up of individuals with streamlined genomes. For streamlined bacteria in particular, we argue that reshuffling of metabolic modules related to energy and carbon metabolism seem to facilitate such adaptive radiation, as shown in the case of metabolic tuning during the marine–freshwater transition. Following such expansion, we imply from our and previous genome reconstructions (Zaremba-Niedzwiedzka *et al.*, 2013) that gene exchange barriers arise, which fix only a small fraction of the original pangenome in the novel biome. As such, freshwater SAR11 seem to be less prone to adaptive radiation, at least if we assume that the genetic structure of global populations controls selection efficacy and ability to adapt in free-living streamlined bacteria.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for access to data storage and computing resources under project b2013274 and b2011105. This work was supported by the Swedish Research Council (Grant Numbers 2012-4592 to AE and 2012-3892 to SB) and the Communiy Sequencing Programme of the US Department of Energy Joint Genome Institute. The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. Accession codes IMG IDs for the 28 genomes are given in Table 1, with data publicly available at IMG.

Author contributions

AE drafted the manuscript; AE, SB and RM wrote the manuscript; AE, RM, LS, LFV, PS, DS, DT and TW designed and carried out bioinformatic analyses; SB, SA, MMG, KDM and RS provided sequence data and designed the freshwater SAR11 sequencing project; all authors contributed to interpretation of the data and edited the manuscript.

References

Adnan F, Weber L, Klug G. (2015). The sRNA SorY confers resistance during photooxidative stress by affecting a metabolite transporter in *Rhodobacter sphaeroides*. *RNA Biol* **12**: 569–577.

Alm E, Huang K, Arkin A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* **2**: e143.

Andersson AF, Riemann L, Bertilsson S. (2009). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME J* **4**: 171–181.

Batterton J, van Baalen C. (1971). Growth responses of blue-green algae to sodium chloride concentration. *Arch Microbiol* **76**: 151–155.

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

Bertilsson S, Tranvik L. (2000). Photochemical transformation of dissolved organic matter in Lakes. *Limnol Oceanogr* **45**: 753–762.

Carini P, Campbell EO, Morré J, Sañudo-Wilhelmy SA, Thrash JC, Bennett SE *et al.* (2014). Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. *ISME J* **8**: 1727–1738.

Carini P, Steindler L, Beszteri S, Giovannoni SJ. (2013). Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTTC1062 on a defined medium. *ISME J* **7**: 592–602.

Carlsson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. (2009). Seasonal dynamics of the SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3**: 283–295.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Chao A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783–79.

Cohan FM. (2001). Bacterial species and speciation. *Syst Biol* **50**: 513–524.

Conway T. (1992). The Entner–Doudoroff pathway: history, physiology and molecular biology. *FEMS Microbiol Rev* **9**: 1–27.

Cordero OX, Polz MF. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**: 263–273.

Doyle M, Fookes M, Ivens A, Mangan MW, Wain J, Dorman CJ. (2007). An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* **315**: 251–252.

Dupont CL, Larsson J, Yooseph S, Ininbergs K, Goll J, Asplund-Samuelsson J *et al.* (2014). Functional tradeoffs underpin salinity-driven divergence in microbial community composition. *PLoS One* **9**: e89549.

Eddy SR. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32**: 1792–1797.

Eiler A, Hayakawa DH, Church MJ, Karl DM, Rappé MS. (2009). Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. *Environ Microbiol* **11**: 2291–2300.

Eiler A, Zaremba-Niedzwiedzka K, Martínez-García M, McMahon KD, Stepanauskas R, Andersson SG *et al.* (2014). Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environ Microbiol* **16**: 2682–2698.

Enright AJ, Van Dongen S, Ouzounis CA. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acid Res* **30**: 1575–1584.

Entner N, Doudoroff M. (1952). Glucose and gluconic acid oxidation of *Pseudomonas saccharophila*. *J Biol Chem* **196**: 853–862.

Garcia SL, Buck M, McMahon KD, Grossart HP, Eiler A, Warnecke F. (2015). Auxotrophy and intra-population complementary in the 'interactome' of a cultivated freshwater model community. *Mol Ecol* **24**: 4449–4459.

Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R *et al.* (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages of freshwater bacterioplankton. *ISME J* **7**: 137–147.

Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL *et al.* (2005a). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82–85.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* **8**: 1553–1565.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005b). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ *et al.* (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *Mbio* **3**: e00252–12.

Hahn MW, 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.

Heinrich F, Eiler A, Bertilsson S. (2013). Seasonality and environmental control of freshwater SAR11 (LD12) in a temperate lake (Lake Erken, Sweden). *Aquat Microb Ecol* **70**: 33–44.

Joset F, Jeanjean R, Hagemann M. (1996). Dynamics of the response of cyanobacteria to salt-stress: deciphering the molecular events. *Physiol Plant* **96**: 738–744.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al.* (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* **344**: 416–420.

Keller H, Wirsching J, Dittmar T, Gutzwiller S, Fournier B. (1999). Cellular expression and functional analysis of Cbfa1 N-terminal isoforms. *Bone* **25**: 167.

Lawrence J. (1999). Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev* **9**: 642–648.

Levin BR, Bergstrom CT. (2000). Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proc Natl Acad Sci USA* **97**: 6981–6985.

Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. (2009). Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol* **9**: 414–422.

Lynch M. (2007). *The Origin of Genome Architecture*. Sinauer Associates, Inc.: Sunderland, MA, USA.

Malmstrom RR, Kiene RP, Cottrell MT, Kirchman DL. (2004). Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic Ocean. *Appl Environ Microbiol* **70**: 4129–4135.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y *et al.* (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acid Res* **40**: 115–122.

Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME *et al.* (2012). High throughput single cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**: 113–123.

Moran AM, Miller WL. (2007). Resourceful heterotrophs make the most of light in the coastal ocean. *Nat Rev Microbiol* **5**: 792–800.

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Neidhardt FC, Ingraham J, Schaechter M. (1990). *Physiology of the Bacterial Cell: A Molecular Approach*. Sinauer Associates: Sunderland, MA, USA.

Newton RJ, Jones SE, Eiler A, McMahon KD, 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, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R *et al.* (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* **77**: 6000–6011.

Page-Sharp M, Behm C, Smith G. (1999). Involvement of compatible solutes trehalose and sucrose in the response to salt stress of cyanobaterial Scytonema species isolated from desert soils. *Biochim Biophys Acta* **1472**: 519–528.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**: 599–609.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C *et al.* (2012). The Pfam protein families database. *Nucleic Acid Res* **40**: 290–301.

Reisch CR, Moran MA, Whitman WB. (2011). Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Front Microbiol* **2**: 172.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Scorcer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Salcher MM, Pernthaler J, Posch T. (2011). Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'That rule the waves' (LD12). *ISME J* **5**: 1242–1452.

Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ. (2010). The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ Microbiol* **12**: 490–500.

Serres MH, Kerr ARW, McCormack TJ, Riley M. (2009). Evolution by leaps: gene duplication in bacteria. *Biol Direct* **4**: 46.

Stamatakis A. (2006). RaxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ. (2011). Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. *PLoS One* **6**: e19725.

Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE *et al.* (2011). One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS One* **6**: e23973.

Sunda WG, Hardison R, Kiene RP, Bucciarelli E, Harada H. (2007). The effect of nitrogen limitation on cellular DMSP and DMS release in marine phytoplankton: climate feedback implications. *Aquat Sci* **69**: 341–351.

Suzuki MT, DeLong EF. (2002). Marine prokaryotic diversity. Staley JT, Reisenbach AL (eds). *Biodiversity of Microbial Life: Foundation of Earth's Biosphere*. John Wiley & Sons: New York, USA, pp 209–234.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF *et al.* (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J* **8**: 1440–1451.

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JW, Wilhelm LJ, Giovannoni SJ. (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**: 741–744.

Tripp HJ, Schwalbach MS, Meyer MM, Kitner JB, Breaker RR, Giovannoni SJ. (2009). Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ Microbiol* **11**: 230–238.

Vergin KL, Beszteri B, Monier A, Thrash JC, Temperton B, Treusch AH *et al.* (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* **7**: 1322–1332.

Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ. (2007). High intraspecific recombination rate in a native population of *Candidatus* Pelagibacter ubique (SAR11). *Environ Microbiol* **9**: 2430–2440.

Viklund J, Martijn J, Ettema JG, Andersson SGE. (2013). Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PLoS One* **8**: e78858.

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**: 27.

Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T *et al.* (2013). Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol* **14**: R130e.

Zawadzki P, Cohan FM. (1995). The size and continuity of DNA segments integrated in Bacillus transformation. *Genetics* **141**: 1231–1243.

Zwart G, Hiorns WD, Methé BA, van Agterveld MP, Huismans R, Nold SC *et al.* (1998). Nearly identical 16S rRNA sequences recovered from lakes in North America and Europe indicate the existence of clades of globally distributed freshwater bacteria. *Syst Appl Microbiol* **21**: 546–556.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)