Open

The ISME Journal (2012) 6, 2067–2077

www.nature.com/ismej

ORIGINAL ARTICLE

Targeted recovery of novel phylogenetic diversity from next-generation sequence data

Michael DJ Lynch, Andrea K Bartram and Josh D Neufeld
Department of Biology, University of Waterloo, Waterloo, ON, Canada

Next-generation sequencing technologies have led to recognition of a so-called 'rare biosphere'. These microbial operational taxonomic units (OTUs) are defined by low relative abundance and may be specifically adapted to maintaining low population sizes. We hypothesized that mining of low-abundance next-generation 16S ribosomal RNA (rRNA) gene data would lead to the discovery of novel phylogenetic diversity, reflecting microorganisms not yet discovered by previous sampling efforts. Here, we test this hypothesis by combining molecular and bioinformatic approaches for targeted retrieval of phylogenetic novelty within rare biosphere OTUs. We combined BLASTN network analysis, phylogenetics and targeted primer design to amplify 16S rRNA gene sequences from unique potential bacterial lineages, comprising part of the rare biosphere from a multi-million sequence data set from an Arctic tundra soil sample. Demonstrating the feasibility of the protocol developed here, three of seven recovered phylogenetic lineages represented extremely divergent taxonomic entities. These divergent target sequences correspond to (a) a previously unknown lineage within the BRC1 candidate phylum, (b) a sister group to the early diverging and currently recognized monospecific Cyanobacteria *Gloeobacter*, a genus containing multiple plesiomorphic traits and (c) a highly divergent lineage phylogenetically resolved within mitochondria. A comparison to twelve next-generation data sets from additional soils suggested persistent low-abundance distributions of these novel 16S rRNA genes. The results demonstrate this sequence analysis and retrieval pipeline as applicable for exploring underrepresented phylogenetic novelty and recovering taxa that may represent significant steps in bacterial evolution.
The ISME Journal (2012) **6,** 2067–2077; doi:10.1038/ismej.2012.50; published online 12 July 2012
Subject Category: integrated genomics and post–genomics approaches in microbial ecology
Keywords: rare biosphere; bioprospecting; molecular ecology; organellar evolution; next-generation sequencing; 16S rRNA

Introduction

Historically, the study of complex microbial communities has been limited by the biases of laboratory cultivation. Even with culture-independent methods, such as brute force Sanger sequencing (Venter *et al.*, 2004; Rusch *et al.*, 2007), a superficial depth of sampling over the past two decades has generated an incomplete picture of complex microbial ecosystems, such as those found in soils and sediments. Recent developments in sequencing technologies, such as the 454 Life Sciences and Illumina platforms, have provided unprecedented access to the genetic diversity of microbial assemblages (Costello *et al.*, 2009; Gloor *et al.*, 2010; Youssef *et al.*, 2010; Bartram *et al.*, 2011; Campbell *et al.*, 2011). As a result of these next-generation sequencing approaches, researchers now collect sequence data even from previously unobserved populations from within microbial communities. Not surprisingly, these investigations have uncovered a large number of unclassified sequences (Bartram *et al.*, 2011; Lecroq *et al.*, 2011), which increase in relative proportion with decreasing relative abundance. However, because of the short-read lengths currently available on these platforms, detailed phylogenetic characterization of these data has not been possible. Furthermore, differentiating genuinely novel diversity from background 'noise', for example, error introduced by PCR and sequencing, has been problematic.

The set of unclassified sequences within an environment has substantial overlap with the so-called 'rare biosphere' (Sogin *et al.*, 2006; Bartram *et al.*, 2011) and the ecology and metabolic roles of these low-abundance organisms are poorly understood. Rare biosphere members may act as keystone species with important contributions to community metabolism, demonstrate increased abundance and activity under changed biogeochemical conditions and evade viral predation by virtue of low relative abundance (Pedrós-Alió, 2007). Further investigating these organisms could provide insight

Correspondence: JD Neufeld, Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada.
E-mail: jneufeld@uwaterloo.ca
Received 21 February 2012; revised 11 April 2012; accepted 21 April 2012; published online 12 July 2012

into adaptations for lifestyles at low abundance, yield novel enzymes for industrial or biomedical applications and, importantly, improve our understanding of the phylogenetic history and evolution of microorganisms. In order to better access the nucleic acids of rare organisms, early attempts to increase sequencing throughput above that of traditional PCR-based clone libraries, serial analysis of ribosomal sequence tags observed higher than expected microbial diversity in Arctic tundra with a high proportion of rare organisms (Neufeld *et al.*, 2004; Neufeld and Mohn, 2005). Initial attempts at characterizing near full-length 16S ribosomal RNA (rRNA) gene sequences of these organisms mostly recovered sequences corresponding to previously observed uncultured clones or were only somewhat distantly related to existing cultivated species (Neufeld *et al.*, 2008). However, the data set analyzed in this previous study was only based on ~2000 sequences, which is not nearly adequate sequence coverage for characterizing α-diversity of soils.

In addition to the challenges of inadequate sequence coverage, accurately exploring rare biosphere members is hindered by the presence of sequencing errors and artifacts within high-throughput sequencing data. Confounders such as incorrect base calling, PCR errors, chimeras and pseudogenes from organisms can manifest as species richness, interfering with inferences of rare biosphere dynamics such as α-diversity. These concerns have been presented elsewhere (Quince *et al.*, 2009; Reeder *et al.*, 2009; Galand *et al.*, 2009; Dickie, 2010; Huse and Welch, 2010; Kunin and Engelbrektson, 2010; Tedersoo and Nilsson, 2010), and together, these studies suggest that naïve sequence surveys of the rare biosphere are faced with substantial challenges. Here, we demonstrate that targeted analysis and retrieval of rare biosphere sequences can largely circumvent these concerns, providing access to rare and uncharacterized lineages directly.

In this study, we leveraged an existing data set of ~6.5 million assembled paired-end Illumina reads from the 16S rRNA gene that was derived from an Arctic tundra sample (Bartram *et al.*, 2011). This research used a combination of a targeted bioinformatics, PCR amplification and DNA sequencing to retrieve low-abundance operational taxonomic units (OTUs) with only weak similarity to known organisms. The identified short-read sequences enabled the design of oligonucleotide primers for the targeted acquisition of nearly full-length 16S rRNA gene sequences from highly divergent phylogenetic lineages. This combined bioinformatics and molecular pipeline was assessed for its ability to specifically amplify targeted sequences that represent unique phylogenetic diversity. The subsequent phylogenetic analysis demonstrated that this approach was effective and can now be applied to mine the diversity of additional terrestrial, aquatic and host-associated environments.

Materials and methods

Bioinformatic analysis of Illumina library

In a previous study, Bartram *et al.,* (2011) evaluated a 125-nucleotide paired-end Illumina sequence data set from tundra soil collected at Alert, Nunavut, Canada (82°30′N 62°19′W). In that study, a large proportion of 97% sequence identity clusters were unclassified at most taxonomic ranks. The most abundant sequence from within each cluster was selected and clusters representing fewer than 10 sequences were excluded to reduce the influence of singletons and sequencing artifacts. To filter for potentially novel taxonomic lineages, representative sequences for each cluster were searched against the nearly 2 000 000 curated sequences within SILVA SSU-Parc release 106 (Pruesse *et al.*, 2007) using BLASTN v.2.2.23+ (Altschul *et al.*, 1997), recording hits with ⩾90% sequence identity. Furthermore, as partial sequences are included in the SILVA SSU-Parc release, only hits with matches over ⩾80% of sequence length were maintained. A network representation of the BLASTN relationships between the Alert library and SILVA SSU-Parc named (top five hits) and unnamed (top hit) sequences using Cytoscape v.2.8.1 (Smoot *et al.*, 2011) was used to characterize sequence novelty. Unconnected nodes, corresponding to uncharacterized sequence clusters within the Alert library, were selected for closer inspection.

Sequence clusters uncharacterized in BLASTN network analysis were evaluated for phylogenetic novelty using Maximum Likelihood reconstruction. Representative sequences were combined with the Ribosomal Database Project (RDP) classifier training set v.6 (Wang *et al.*, 2007) and aligned to the consensus bacterial secondary structure model using ssu-align v.0.1 (Nawrocki and Eddy, 2010). A Maximum Likelihood phylogeny was constructed from the alignment using default parameters in RAxML v.7.2.8 (Stamatakis, 2006) with the GTRGAMMA model of sequence evolution. The resulting phylogeny was manually inspected for monophyletic lineages of Alert clusters that were phylogenetically distinct from sequences in the RDP training set. Forward unique lineage (UL) primers specific to each of these clades were designed from the highly variable 3′ end of the V3 region. As a positive control to validate the protocol, clade-specific custom primers were designed against an Acidobacteria and an Alphaproteobacteria sequence, both common operational taxonomic units in the sequence library (Bartram *et al.*, 2011). Designed primers were subjected to BLASTN against the non-redundant sequence set to ensure specificity of the primer to the clade.

Amplification and sequencing of ULs

Template genomic DNA was extracted from the same Arctic soil (Alert Nunavut, Canada) examined earlier (Neufeld and Mohn, 2005; Neufeld *et al.*, 2008;

Bartram *et al.*, 2011) using the FastDNA kit for soil (MP Biomedicals, Solon, OH, USA). Amplification of 16S rRNA gene fragments was performed with forward primers designed from rare and abundant arctic tundra V3 sequences (as described above) and 1492R (Lane, 1991) was used as the reverse primer, unless otherwise stated (see Supplementary Table S1). All PCR amplifications were carried out in 25-μl volumes containing 0.4 μM of each forward and reverse primer, 200 μM of each deoxynucleoside triphosphate (deoxyribonucleotide triphosphate, New England Biolabs, Ipswich, MA, USA), 2 mM $MgSO_4$, 15 μg bovine serum albumin (Sigma-Aldrich, St Louis, MO, USA) and 0.5 U of *Taq* DNA polymerase (New England Biolabs). Some optimization was required to obtain product for four of the primer sets (UL5.1/1492R, UL9.2/1492R, UL13.1/1512uR and UL13.1/907R), where primer concentration was increased to 0.8 μM. PCR conditions consisted of a 5 min denaturation step at 95°C, followed by 30 cycles of 30 s at 95°C, 30 s at the primer specific annealing temperature (see Supplementary Table S1) and 72°C for 30 s, with a final extension step for 7 min at 72°C. The annealing temperatures were determined experimentally by running PCR using a temperature gradient (DNA Engine; BioRad, Hercules, CA, USA). Selection of an appropriate annealing temperature was based on the highest temperature that still gave a detectable amplification product seen on an agarose gel (1% agarose pre-cast with 1 × gel red nucleic acid stain; Biotium, Hayward, CA, USA). The resulting PCR products were cloned using the TA TOPO cloning kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Colony PCR was performed on white colonies using the M13F and M13R primer pair. PCR amplifications were carried out in 30-μL volumes with the same PCR reagent concentrations and temperature conditions described above, except without bovine serum albumin and the denaturing, annealing and extension times were extended to 1 min. PCR products were sequenced bidirectionally using the Sanger method by Beckman Coulter Genomics (Danvers, MA, USA).

Phylogenetic and taxonomic analysis of novel 16S rRNA gene sequences

Sequences were manually assembled from paired-end reads and evaluated for conservation of the V3 region downstream of the primer, which was used as a measure of amplification fidelity. Additionally, divergence from existing sequences was characterized by BLASTN (Altschul *et al.*, 1997) analysis against GenBank's non-redundant (nr) database. A heatmap of sequence identities of top hits against nr was generated using the gplots library within R (R Development Core Team, 2011). Putative taxonomy for each sequence was assigned by the naïve Bayesian classification of the RDP Classifier v.2.1 (Wang *et al.*, 2007).

Sequences were screened for chimeras by comparison to the GreenGenes curated sequence set (DeSantis *et al.*, 2006) using UCHIME (Edgar *et al.*, 2011), and putative chimeric sequences were removed. In order to provide seed sequences of known phylogeny and taxonomy, near full-length sequences were combined with bacterial sequences from the Living Tree Project release 106 (Yarza *et al.*, 2008, 2010). Additionally, sequences observed in a previous analogous study (Neufeld *et al.*, 2008) were added to contrast methodological improvements. Experimental (Alert) and seed sequences were aligned to a consensus bacterial secondary structure model using ssu-align (Nawrocki and Eddy, 2010). To reduce computational effort, sequence redundancy was reduced to 90% within the known Living Tree Project seed sequences using CD-HIT (Li and Godzik, 2006). A maximum likelihood phylogeny was constructed with RAxML v.7.2.8 (Stamatakis, 2006) using the GTRGAMMA model of sequence evolution, maintaining the best scoring tree of 100 iterations. Bootstrap support values were derived from 1000 iterations of maximum likelihood bootstrap.

In order to investigate the potential that sequences of organelle origin were contributing to the UL diversity, UL9 and UL13 were further evaluated to establish phylogenetic position within Bacteria. Two seed sequence data sets were obtained from SILVA SSU-Parc release 108 (Pruesse *et al.*, 2007) corresponding to chloroplast organellar and Cyanobacteria, as well as mitochondria 16S rRNA gene sequences. Outgroup sequences from *Escherichia coli* (GenBank: AB035920) and *Vibrio vulnificus* (GenBank: BA000037) were used to root the mitochondrial phylogeny. Redundancy within each sequence set was reduced to 90% using CD-HIT (Li and Godzik, 2006). Relevant experimental sequences were added to the corresponding sequence set and alignments were constructed using ssu-align (Nawrocki and Eddy, 2010). Maximum likelihood phylogenies were inferred using RAxML v.7.28 (Stamatakis, 2006) and two schemes for models of sequence evolution; (1) GTRGAMMA for all sites and (2) GTRGAMMA for non-paired characters and the RNA structural model S16 for paired sites inferred from the consensus secondary structure generated by ssu-align. Node support values were summarized from 100 maximum likelihood bootstrap replicates for each evolutionary model scheme, as well as local support values from the Shimodaira Hasegawa test implemented in FastTree v.2.1.4 (Price *et al.*, 2010) using default parameters.

All phylogenies were viewed using FigTree v.1.3.1 (Rambaut). Near full-length 16S rRNA gene sequences were deposited in GenBank (Accession no.: JQ307004–JQ307092).

Probing metagenomic libraries with candidate sequences

Little is known regarding the distribution of sequences occupying low-abundance ranks and

sequences constituting the experimental subset observed in the Alert sequence library likely exist in other, similar environments. The Canadian MetaMicrobiome Library (CM²BL; http://www.cm2bl.org; Neufeld *et al.*, 2011) contains 12 V3 SSU metagenomic libraries from soil samples throughout Canada, including agricultural, tundra, boreal, oil sand, compost and wetland locations. Representative V3 sequences from each UL were queried against a non-redundant BLAST database of these libraries at 97% sequence identity using BLASTN v.2.2.23+ (Altschul *et al.*, 1997).

Results

Network analysis and phylogenetic novelty

Naïve assembly and CD-HIT clustering of approximately 12 million raw paired-end sequences derived from an Arctic tundra soil library (Bartram *et al.*, 2011) generated close to 6.5 million assembled sequences for comparison with sequence databases. Most assembled V3-region sequence clusters had BLASTN hits within the 'known' threshold of ⩾90% sequence identity and ⩾80% length against SILVA SSU-Parc release 106, represented as connected nodes (Figure 1, Supplementary S1). Sequence clusters representing 97% sequence identity groups that were abundant or of known taxonomy tended to occur in highly connected subtrees (for example, Figure 1a). Unconnected nodes (for example, Figure 1b), corresponding to V3 sequence clusters that lacked BLASTN association with SILVA 16S rRNA gene sequences, had the highest potential for phylogenetic novelty and were analyzed further. A total of 558 nodes were unconnected, 512 of which successfully aligned to the bacterial 16S rRNA gene model representing 28 203 sequences (0.44% of the full library). In phylogenetic screening of unconnected nodes, representative sequences tended to be distributed throughout well-defined clades with known taxonomy and were thus less likely to represent novel phylogenetic entities (Supplementary Figure S2). Eight clades consisting of multiple Alert OTU clusters that were notable or phylogenetically distinct from known seed sequences were selected and oligonucleotide primers specific to each clade were designed, primarily against the highly variable 3′ end of the V3 region (Supplementary Table S1 primers).

Directed PCR amplification of novel lineages

Using a temperature gradient of annealing temperatures to identify stringent PCR conditions, near full-length SSU sequences were amplified and sequenced from seven of the eight experimental clades using custom UL primers (Supplementary Table S1). Only UL14 primers designed against a sister group to the *Clostridium* genus (Supplementary Figure S2) did not successfully amplify full-length 16S rRNA genes. UL primer design and PCR amplification were also

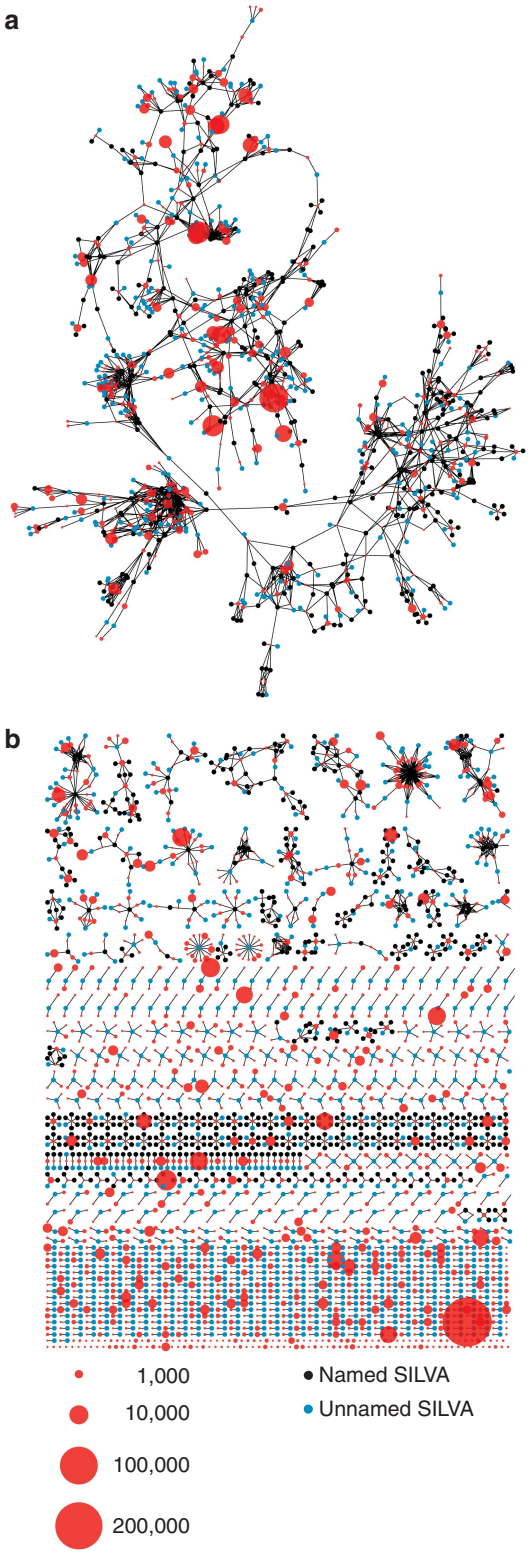

Figure 1 Network analysis of Alert library sequences against SILVA SSU-Parc release 106 (Pruesse *et al.*, 2007). Red nodes represent 97% sequence identity clusters with diameters corresponding to cluster abundance. SILVA sequences are represented by black (named) and blue (unnamed) nodes. Edges represent a BLASTN result of ⩾90% identity across ⩾80% of the V3 region. (**a**) A highly connected subgroup corresponding to the Bacteroidetes/Chlorobi group (**b**) low degree or unconnected nodes representing sequence clusters of potential phylogenetic novelty.

conducted on two relatively abundant Alert library sequences, representing Acidobacteria and Alphaproteobacteria sequences, to serve as positive controls (Supplementary Table S2). Demonstrating the specificity of the targeted PCR, nearly all retrieved 16S rRNA gene sequences were associated with the anticipated V3 region because the Sanger sequence data directly adjacent to the PCR primers was consistent with the original V3-region sequence. However, five sequences from UL13 were associated with the Eukaryota (Figure 2) despite amplification with the prokaryote-specific 1512uR (Weisburg *et al.*, 1991) reverse primer (Bartram *et al.*, 2011). Subsequent investigations with RDP Probematch (Cole *et al.*, 2007; Cole *et al.*, 2009) revealed a surprisingly high identity for this primer against archaeal sequences, fully matching 75% of Archaea. The alternative reverse primer we used, 907R, matched only 0.39% of Archaea. RDP Probematch does not compare against Eukaryota sequences, but as Eukaryota 18S rRNA gene sequences are closer to Archaea than Bacteria it implies some identity of the primer with Eukaryota sequences. Two near full-length sequences were putative chimeras as determined by UCHIME and were excluded from analyses. In total, 84 clones were successfully sequenced for near full-length bacterial 16S rRNA genes.

Typically, sequence divergence noted in the V3 region of targeted OTUs was maintained in the near full-length 16S rRNA genes, as compared with the two retrieved positive control sequences (Figure 2). The BLASTN results against GenBank's non-redundant database indicated very high novelty, especially when restricted to named isolates (Figure 2a). In particular, sequences from UL4, UL5 and UL13 demonstrated sequence identities $<85\%$ and as low as 75% (Figures 2a and b). Sequences amplified with the positive control primers had BLASTN hits within the Alphaproteobacteria (Sphingomonadales) and Acidobacteria (Gp4), consistent with the taxonomy and phylogeny inferred from the V3 region (Supplementary Figure S2).

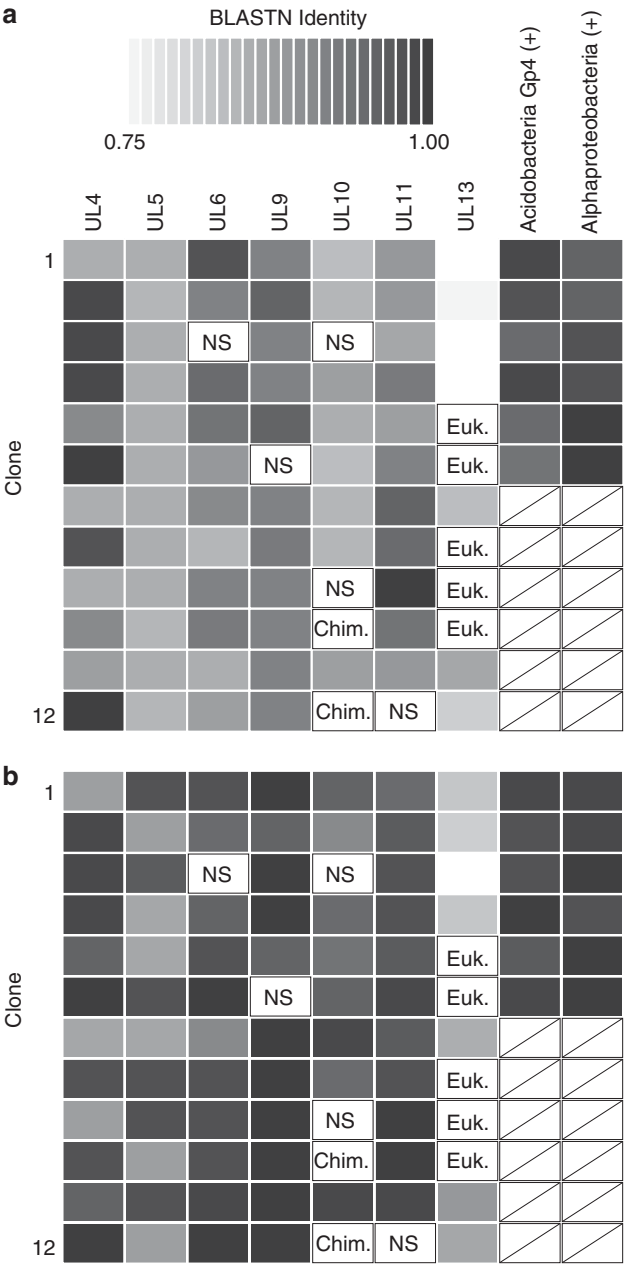

Figure 2 Heatmap representing identities of highest BLASTN hits for each amplified clone from putative novel lineages. Sequences amplified from Alert, NU; (**a**) excluding uncultured or environmental sequences and (**b**) unfiltered non-redundant NCBI (National Center for Biotechnology Information) database. Euk. = eukaryota, Chim. = chimeric sequence, NS = no sequences successfully obtained. Only six sequences were attempted for positive control (+) groups.

Phylogenetic distribution of novel lineages

Of the seven UL sequence sets amplified in this study, three were highly divergent from known bacterial lineages, representing significant, novel phylogenetic entities within the Cyanobacteria (UL9) and the BRC1 (UL5) candidate phylum, as well as a divergent group within the mitochondrial clade (UL13). Large species-rich bacterial phyla, such as the Firmicutes and various Proteobacteria, typically did not contain UL sequences (Figure 3). One exception was sequences retrieved with primers corresponding to the highly divergent UL13 clade, which resolved within the Alphaproteobacteria. Sequences from each targeted UL tended to be monophyletic (Figure 3), with the primary exception being sequences from UL11, which were broadly distributed throughout the phylogeny. One custom primer, UL14, did not successfully amplify product from the Alert soils and was therefore not represented in Sanger sequencing. This primer was designed against a clade of V3 sequences that resolved as sister to Clostridium taxa (Supplementary Figure S2).

Sequences from UL4 grouped in several clades within the Bacteroidetes and all classified to the Sphingobacteriales. Half of the 12 sequences formed

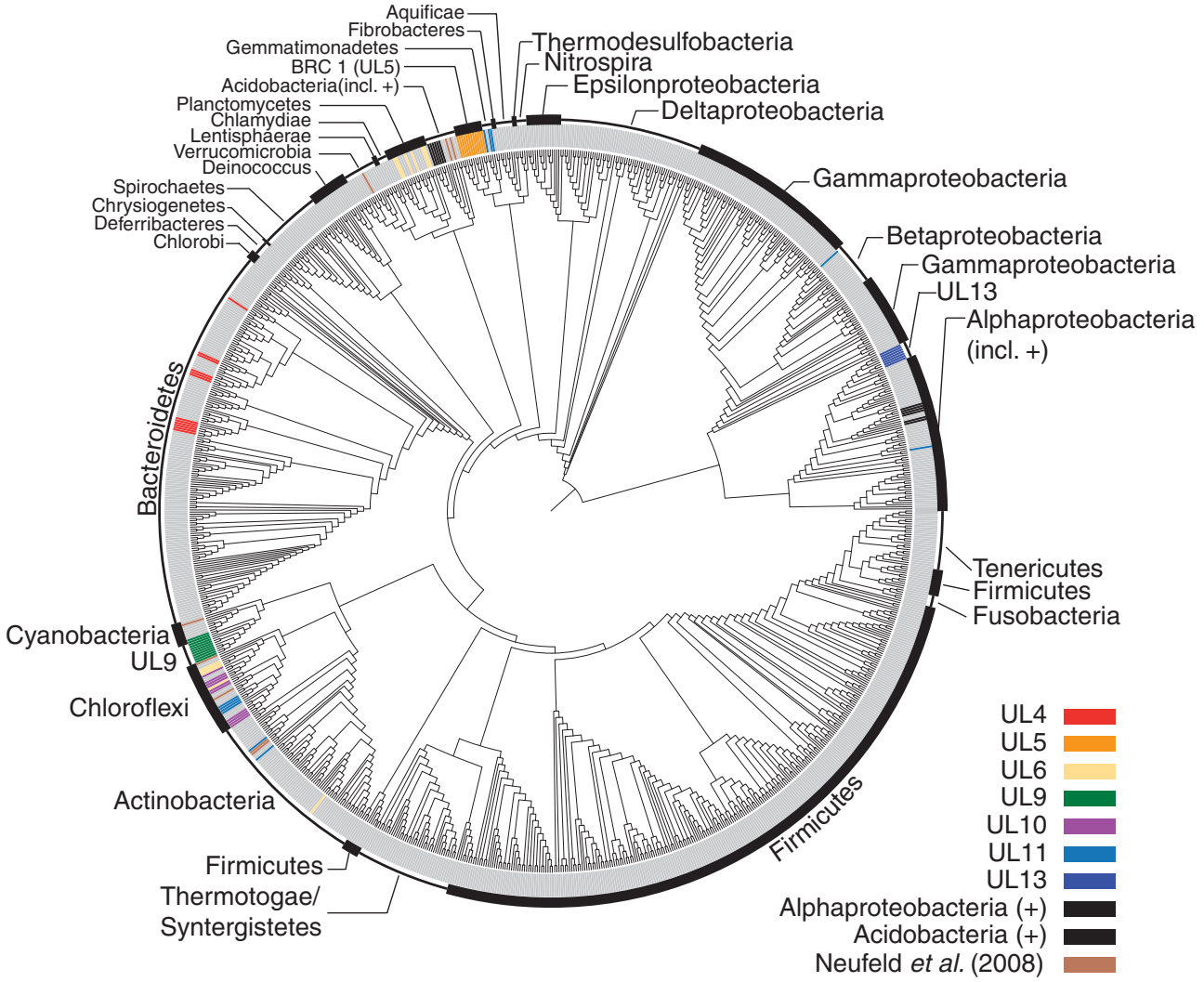

Figure 3 Maximum Likelihood phylogeny constructed from nearly full-length 16S rRNA gene sequences. Sequences amplified from putative novel V3 sequences combined with seed sequences of known taxonomy derived from the Living Tree Project.

a sister clade relative to *Nubsella zeaxanthinifaciens* (Figure 4a) and showed highest identity with *Pedobacter* sp. in BLASTN analysis. Three of the remaining sequences appeared to be phylogenetically novel, sister to, but divergent from aquatic bacterial species *Microscilla marina* and *Flexibacter elegans* (Figure 4b). This topology was not well supported by bootstrap analysis, although the clade's position within the larger group was well supported. Similar to UL4, sequences from UL6 predominantly grouped in three areas of the tree; however, six of 11 sequences grouped diffusely throughout the Planctomycetales (Figure 4c). Four of the remaining sequences grouped strongly within the Chloroflexi, but could not be assigned to more specific taxonomic ranks (Figure 4d).

Sequences from UL10 resolved within the Chloroflexi, distributed in three separate clades roughly corresponding to the Sphaerobacterales, Herpetosiphonales and Anaerolineales, each with strong bootstrap support (Figure 4d). Only sequences resolved within the Anaerolineales were classified as such, consistent with the long branches in the phylogeny leading to sequences sister to *Roseiflexus* and *Chloroflexus* (Herpetosiphonales), as well as clades of UL10 sequences without taxonomic seeds (Spaerobacterales-like).

The primers for the amplification of UL11 were the least specific in this study, with sequences distributed throughout the bacterial phylogeny. Three sequences strongly supported as monophyletic with *Gemmatimonas* showed the most significant novelty among UL11 sequences (Figure 4e). Additionally, three UL11 sequences were strongly supported as sister to *Thermomicrobium* (Figure 4d).

Importantly, the remaining UL sequence sets, UL5, UL9 and UL13, were each strongly supported as monophyletic and demonstrated the most phylogenetic novelty among the experimental lineages in this study. Sequences from UL5 occurred in two fully supported sister groups that did not resolve with known seed sequences (Figure 4e), consistent

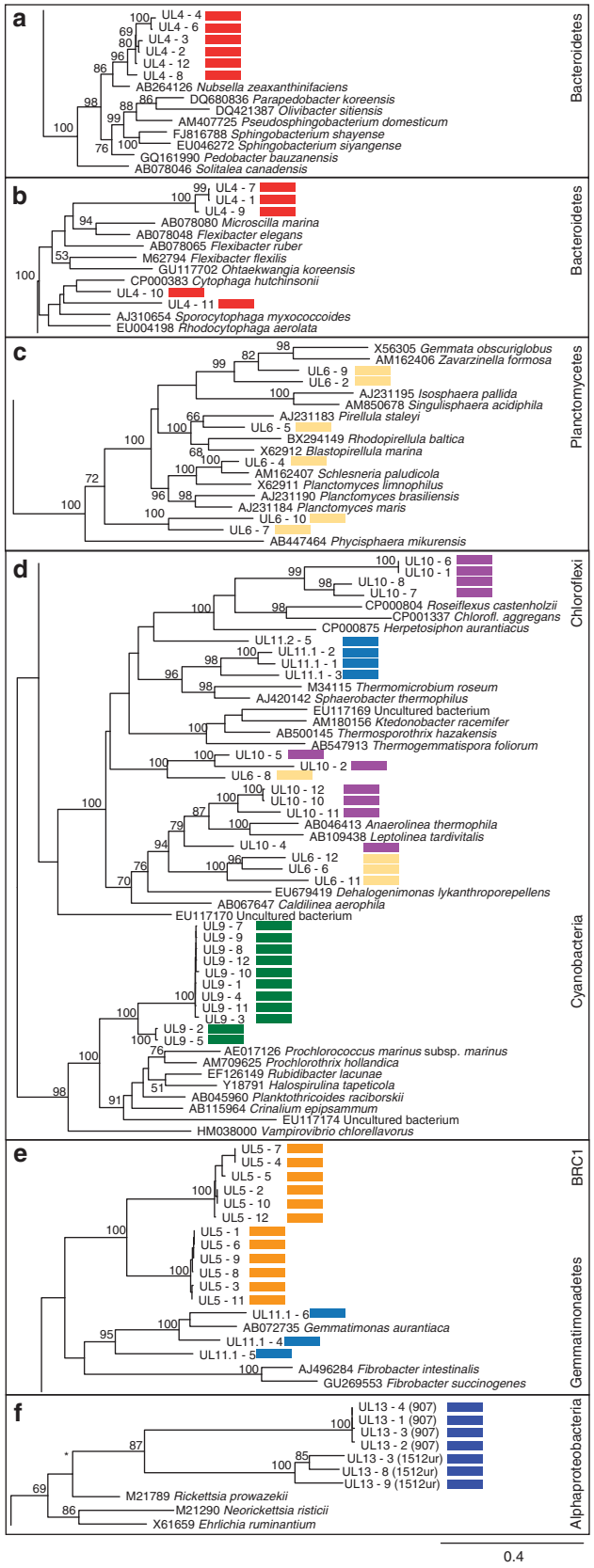

Figure 4 Subtrees (**a**–**f**) of experimental UL sequences from Figure 3 demonstrating local phylogenetic patterns. Node support values correspond to maximum likelihood parametric bootstrap scores $\geqslant 50\%$. * = all subsequent branch lengths are 50% to scale.

with BLASTN results (Figure 2). The clade of UL5 sequences had ambiguous phylogenetic placement in our results, although one group was identified by the RDP classifier as BRC1 (Derakshani *et al.*, 2001), a candidate phylum. The sister group had no clear classification subordinate to domain.

One set of putatively novel sequences, UL9, was phylogenetically and taxonomically resolved as Cyanobacteria. These sequences were fully supported as sister to all Cyanobacteria in phylogenetic analysis (Figure 4d) and formed two distinct, fully supported groups, each exhibiting high internal sequence identity. The smaller of these groups, with two sequences, showed high sequence identity (95%) to *Gloeobacter violaceus* (GenBank: FR798924), an early diverging monospecific cyanobacterium not included in the Living Tree Project. The remaining nine sequences were all nearly identical to a single uncultured, unpublished clone derived from moss pillars in east Antarctica (GenBank: AB630682), but were approximately 10% divergent from the next closest sequence within GenBank.

Bacterial 16S rRNA gene sequences amplified by UL13 were monophyletic and highly divergent while weakly resolving within the Rickettsiales, an order containing mitochondrial sequences (Thrash *et al.*, 2011) and other obligate intracellular aerobic species. These sequences formed two clades, each quite divergent from the other and known seed sequences in the phylogeny (Figure 4f). Similarly, BLASTN and RDP classification further supported the novelty of these sequences, with very low sequence identity with top hits in GenBank (Figure 2) and no strong classification at taxonomic ranks subordinate to Bacteria.

As the phylogenetic backbone used here did not include organelle sequences, we performed additional phylogenetic analyses by including 16S rRNA gene sequences from chloroplasts and mitochondria to explore the origin of UL9 and UL13 sequences. Informative GenBank matches, most notably *Gloeobacter*, were also included. The cyanobacterial phylogenetic placement of UL9 sequences was consistent with previous phylogenies (Figures 3 and 4) even after chloroplast 16S rRNA sequences were included (Figure 5a). Experimental sequences were monophyletic with *Gloeobacter* and sister to the remaining Cyanobacteria and chloroplast sequences. Sequences in UL13 resolved as two monophyletic groups, both supported within clades corresponding to mitochondrial sequences from bikont ('two flagella') organisms (Figure 5b). One clade was moderately supported as sister to mitochondrial sequences from *Acanthamoeba* sp., while the other group was strongly supported as monophyletic with uncultured organisms and grouped with organelles from predominantly algal lineages including Rhodophyta (red algae) and Chromalveolata. The 16S rRNA genes from mitochondria are poorly represented in current sequence databases

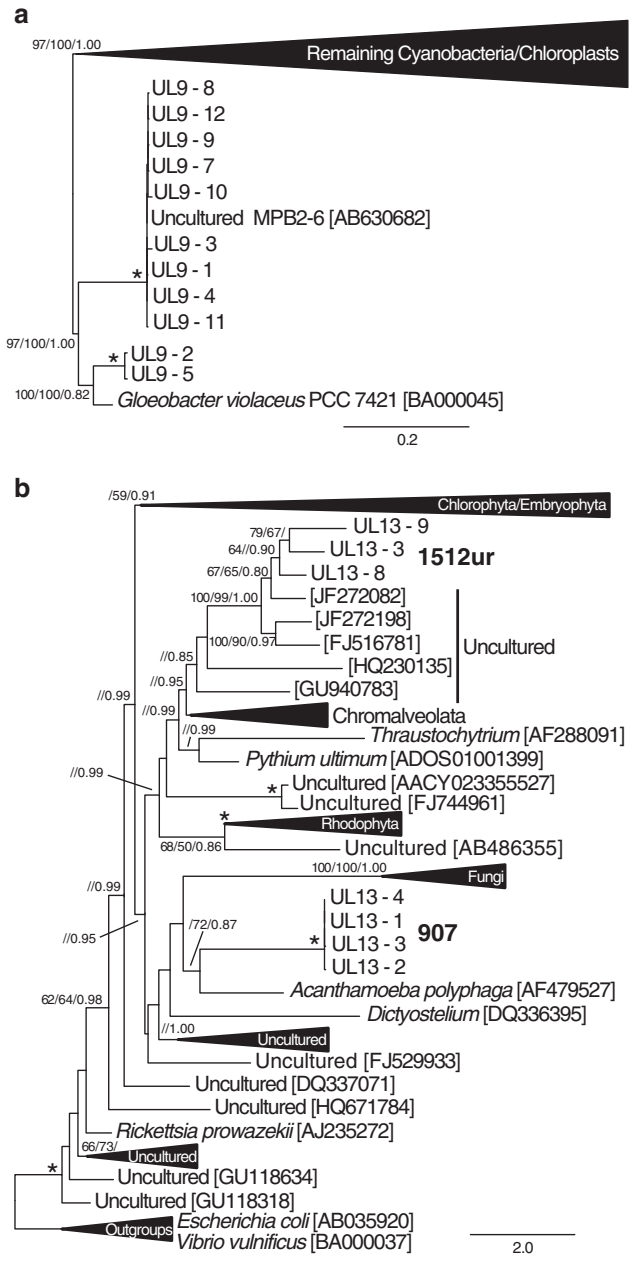

Figure 5 Maximum likelihood phylogenies resolving UL sequences with proximity to organelle 16S rRNA gene sequences. (**a**) UL9 sequences with cyanobacterial and chloroplast 16S rRNA gene sequence data and (**b**) UL13 sequences with mitochondrial 16S rRNA gene sequence data. Node support values correspond to ML bootstrap (GTRGAMMA)/ML bootstrap (S16)/SH-like test (GTR + Γ). Note: size of collapsed wedges does not correspond to the number of taxa. * = full phylogenetic support (100/100/1.00). Support values <50% or 0.5 are not shown.

and, in general, phylogenies inferred from mitochondrial sequences tended to be poorly supported here (Figure 5b).

Searching for novel sequences in other data sets

The V3-region sequences identified as novel in this study were low-abundance in our Arctic tundra sample and poorly represented in existing databases. We hypothesized that these sequences would be associated with the other low relative abundance sequences in additional soils. To test this hypothesis, we compared the V3 regions targeted in this study to equivalent 16S rRNA gene data sets generated for 12 soils collected as part of a soil metagenomics resource (Canadian MetaMicroBiome Initiative; CM²BL; Neufeld *et al*., 2011). The V3 sequences within putative novel clades were not well represented across CM²BL sequence libraries, with only UL5, UL6 and UL13 lineages observed in amounts comparable to the Alert library (Table 1). Representation was highest in UL5, represented in 8 of 12 libraries. Highest overall counts were observed in the Arctic Tundra 2 (2ATN) library, the library at a latitude most similar to the Alert, NU site. Only two environments, compost (13CO) and Northern Peatlands (8NP), did not have any BLASTN matches to the Alert V3 sequence from putative novel lineages.

Table 1 Sequence counts of BLASTN hits of unknown lineage sequences against the Canadian MetaMicroBiome Library (CM²BL) (http://www.cm2bl.org)

CM²BL	*No. of sequences*	*UL4*	*UL5*	*UL6*	*UL9*	*UL10*	*UL11*	*UL13*
Northern peatlands (8NP)	1 704 862	0	0	0	0	0	0	0
Oil sand 1 (4TS)	1 664 455	0	2	2	2	0	0	0
Arctic tundra 1 (1AT)	2 095 381	0	0	0	0	0	0	18
Temperate rain forest (7TR)	1 800 755	0	1	2	0	0	0	7
Agricultural soil, wheat (11AW)	1 375 901	0	111	15	0	0	0	0
Arctic tundra 2 (2ATN)	1 882 676	0	219	61	0	0	0	0
Boreal coniferous forest (5BF)	1 663 561	0	66	16	0	0	0	0
Temperate deciduous forest (6TD)	1 910 118	0	0	0	0	0	0	21
Compost (13CO)	1 583 559	0	0	0	0	0	0	0
Wetland soil (9WLM)	1 852 938	0	12	16	0	0	0	2
Agricultural soil, soy (10AS)	1 624 052	0	73	4	0	0	0	0
UW community garden (20CG)	1 552 389	0	27	6	0	0	0	1

Abbreviations: CM²BL, Canadian MetaMicroBiome Library; UL, unique lineage.

Discussion

Many computational pipelines exist for analyzing the taxonomy and phylogeny of 16S rRNA gene sequence data generated by pyrosequencing and

Illumina platforms (Schloss *et al.*, 2009; Caporaso *et al.*, 2010; Giongo *et al.*, 2010). The approach outlined in this study represents a major methodological improvement for characterizing the phylogenetic distribution of unclassified microbial diversity analyzed by short-read, high-throughput sequencing studies, which is a fraction often overlapping with the rare biosphere. In particular, we report a high success rate (7 of 8 primers) for the specific amplification of putatively novel lineages that contribute less than 1.0×10^{-4}% of all sequences in an environmental DNA extraction. Additionally, the recovery of highly novel clades, particularly UL5, UL9 and UL13, suggests that directed targeting of phylogenetic novelty from high-throughput sequencing projects is feasible. This protocol was further validated by the exclusive amplification and recovery of positive control sequences. Investigations specifically targeting novel phylogenetic lineages offer the potential of not only increasing the breadth of taxonomic knowledge, but offer an additional tool for investigating deep branching lineages of life in general (Sogin *et al.*, 2006; Wu *et al.*, 2011; Youssef *et al.*, 2012).

The lineages UL5, UL9 and UL13 each represented significant and repeatable highly novel phylogenetic groups and demonstrated the value of this approach. Each group was monophyletic and either completely unique to this study or significantly increased knowledge of uncultured sequence data within GenBank. One of the two internally consistent clades of UL5 sequences was classified as BRC1 using the RDP classifier. However, BLASTN analysis was ambiguous, only showing 92% identity with the BRC1 clade (Figure 2). Regardless, as sequence identity of UL5 against BRC1 sequences is within the observed range of existing BRC1 lineages (Derakshani *et al.*, 2001) and the two UL5 clades are fully supported as monophyletic, they likely represent two additional species within phylotype-defined BRC1, significantly adding to its known diversity.

Extreme environments tend to contain unique cyanobacterial populations. Specifically, polar environments harbor species with high tolerance to UV (Quesada *et al.*, 1999; George *et al.*, 2001) and temperature extremes (Tang *et al.*, 1997). UL9 primers amplified 16S rRNA genes from two distinct Cyanobacteria species with strong bootstrap support for the sister relationship between the isolates, and full resolution as sister to all Cyanobacteria (Figures 3, 4d and 5a). Based on the phylogenetic resolution and BLASTN results, this clade appears to be a novel sister group to *Gloeobacter violaceus*, significantly adding to our understanding of the early evolution of the Cyanobacteria. *Gloeobacter* is a monospecific lineage representing an early radiation of Cyanobacteria and contains several features highly divergent from other cyanobacterial species, including the absence of thylakoids (Nakamura and Kaneko, 2003; and references therein). At minimum, UL9 sequences indicate the presence of a second clade of Cyanobacteria, along with *Gloeobacter*, that diverged very early in cyanobacterial evolution. Furthermore, the single observation of a near-identical sequence in Antarctica provides an interesting case of microbial dispersal. The GenBank sequence nearly identical to the larger UL9 clade was isolated from a moss pillar within a freshwater lake in eastern Antarctica. This sequence is potentially a lichen photobiont based on the association of moss and lichen in Antarctica (Victoria *et al.*, 2006). As lichens are a primary colonizer for tundra, one potential source of this sequence is as a previously unobserved cyanobacterial photobiont in lichen. Such an association would help explain the bi-polar distribution of this sequence as lichen species can be easily distributed and are tolerant to environmental stress such as desiccation. The near complete absence of this lineage within CM²BL libraries is notable (Table 1), suggesting that it is not broadly distributed, arguing for animal (for example, bird) or anthropogenic dispersal.

The set of sequences with the highest taxonomic novelty recovered in this study, UL13 (Figures 4f and 5b), were so divergent that inferences about ecology are difficult, although intriguing. Despite UL13 sequences resolving as sister to the obligate intracellular parasite *Rickettsia* (Figure 4f), the phylogenetic support tended to be weak, likely due to the magnitude of sequence divergence. The closest BLASTN matches with mitochondrial sequences are not surprising given the phylogenetic placement near *Rickettsia* and within algal mitochondria (Figure 5b). Related intracellular parasites and the SAR11 clade likely had a role in the evolution of mitochondria (Thrash *et al.*, 2011; Rodríguez-Ezpeleta and Embley, 2012). There are relatively few mitochondrial 16S rRNA gene sequences available in public sequence databases, resulting in large gaps in our knowledge of bacterial 16S rRNA gene sequence data. An analysis relying exclusively on sequence divergence and non-phylogenetic classification schemes or poor taxon sampling (for example, Figure 4f) would have incorrectly inferred bacterial novelty within the Rickettsiales instead of within unknown mitochondrial diversity. This uncharacterized 16S rRNA gene sequence diversity for mitochondria should be addressed, as microbial diversity studies tend not to correctly account for sequences of organellar origin.

It is possible that the rare and highly divergent 16S rRNA gene sequences amplified in this study did not represent bacterial species active in the ecosystem, and instead correspond to pseudogenes, dormant organisms or other such components. The fact that these sequences successfully aligned to valid 16S rRNA gene structures suggests these are *bona fide* 16S rRNA genes. Unfortunately, activity or metabolism cannot be inferred from single-gene DNA-based data. The directed amplification of these

sequences did not necessarily recover members of potential unique clades in the proportions present in the environment or observed in high-throughput sequencing. This is not surprising, although it does indicate this approach should not be used to explore diversity relationships within the rare biosphere, but rather to further explore the evolutionary history and phylogenetic novelty of species constituting rare or uncharacterized groups.

The majority of ULs amplified here occurred at low relative abundance in Alert soils. Their relatively low abundance suggests that their constituent genes would not readily contribute to metagenomic libraries constructed from this and similar soil sites. Due to the magnitude of phylogenetic novelty observed, these organisms likely also represent highly divergent genomes that would be valuable to target further by cell sorting and inclusion with The Microbial Earth Project (http://genome.jgi.doe.gov/programs/bacteria-archaea/MEP/index.jsf). With near full-length 16S rRNA gene sequences available, existing soil libraries can be further probed for these organisms, in attempts to isolate and amplify genomic material. This technique would therefore have applications in bioprospecting, specifically targeting phylogenetically ULs. The number of highly divergent lineages observed here, combined with the high proportion of sequences with unknown taxonomy from Alert soils, suggests that polar environments should be further explored for microbial diversity. This will not only improve our understanding of the ecology of these systems as they face an uncertain future, but also increase our knowledge of microbial diversity and organellar evolution.

Acknowledgements

This research was supported by Discovery and Strategic Project Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and an Early Researcher Award from the Government of Ontario.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl Environ Microbiol* **77**: 3846–3852.

Campbell BJ, Yu L, Heidelberg John F, Kirchman DL. (2011). Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci USA* **108**: 12776–12781.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* **7**: 335–336.

Cole J, Chai B, Farris R. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* **37**: D141–D145.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Derakshani M, Lukow T, Liesack W. (2001). Novel bacterial lineages at the (sub) division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. *Appl Environ Microbiol* **67**: 623–631.

Dickie I. (2010). Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytol* **4**: 916–918.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106**: 22427–22432.

George AL, Murray AW, Montiel PO. (2001). Tolerance of Antarctic cyanobacterial mats to enhanced UV radiation. *FEMS Microbiol Ecol* **37**: 91–101.

Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, Gano KA *et al.* (2010). PANGEA: pipeline for analysis of next generation amplicons. *ISME J* **4**: 852–861.

Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R *et al.* (2010). Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* **5**: e15406.

Huse S, Welch D. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.

Kembel SW, Eisen Jonathan A, Pollard KS, Green JL. (2011). The phylogenetic diversity of metagenomes libraries. *PLoS ONE* **6**: e23214.

Kunin V, Engelbrektson A. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Lane D. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. John Wiley and Sons: New York, pp 115–175.

Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, Baerlocher L *et al.* (2011). Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc Natl Acad Sci USA* **108**: 13177–13182.

Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Nakamura Y, Kaneko T. (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**: 137–145.

Nawrocki EP, Eddy SR. (2010). ssu-align: a tool for structural alignment of SSU rRNA sequences. Available at: http://selab.janelia.org/software.html.

Neufeld JD, Engel K, Cheng J, Moreno-Hagelsieb G, Rose DR, Charles TC. (2011). Open resource metagenomics: a model for sharing metagenomic libraries. *Stand Genomic Sci* **5**: 203–210.

Neufeld JD, Yu Z, Lam W, Mohn WW. (2004). Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities. *Environ Microbiol* **6**: 131–144.

Neufeld JD, Li J, Mohn WW. (2008). Scratching the surface of the rare biosphere with ribosomal sequence tag primers. *FEMS Microbiol Lett* **283**: 146–153.

Neufeld JD, Mohn WW. (2005). Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. *Appl Environ Microbiol* **71**: 5710–5718.

Pedrós-Alió C. (2007). Dipping into the rare biosphere. *Science* **315**: 192.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS ONE* **5**: e9490.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Quesada A, Vincent WF, Lean DRS. (1999). Community and pigment structure of Arctic cyanobacterial assemblages: the occurrence and distribution of UV-absorbing compounds. *FEMS Microbiol Ecol* **28**: 315–323.

Quince C, Lanzén A, Curtis T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Rambaut A. FigTree: Tree Figure Drawing Tool. http://tree.bio.ed.ac.uk.

R Development Core Team (2011). *R: A Language And Environment For Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Reeder J, Knight R. (2009). The 'rare biosphere': a reality check. *Nat Methods* **6**: 636–637.

Rodríguez-Ezpeleta N, Embley TM. (2012). The SAR11 group of Alpha-Proteobacteria is not related to the origin of mitochondria. *PLoS ONE* **7**: e30520.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Schadt CW, Martin AP, Lipson DA, Schmidt SK. (2003). Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science* **301**: 1359–1361.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**: 431–432.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored ''rare biosphere''. *Proc Natl Acad Sci USA* **103**: 12115–12120.

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

Tang EPY, Tremblay R, Vincent WF. (1997). Cyanobacterial dominance of polar freshwater ecosystems: are high-latitude mat-formers adapted to low temperature? *J Phycol* **33**: 171–181.

Tedersoo L, Nilsson R. (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol* **188**: 291–301.

The Microbial Earth Project. http://genome.jgi.doe.gov/programs/bacteria-archaea/MEP/index.jsf.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ *et al.* (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci Rep* **1**: 13.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Victoria FC, Albuquerque MP, Pereira AB. (2006). Lichen-moss association in plant comunnities of the Southwest Admiralty Bay, King George Island, Antarctica. *Neotrop Biol Conserv* **1**: 84–89.

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

Weisburg WG, Barns SM, Pelletier DA, Lane DJ. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* **173**: 697–703.

Wu D, Wu M, Halpern A, Rusch DB. (2011). Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PloS ONE* **6**: e18011.

Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO *et al.* (2010). Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. *Syst Appl Microbiol* **33**: 291–299.

Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH *et al.* (2008). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol* **31**: 241–250.

Youssef NH, Couger M, Elshahed MS. (2010). Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, USA): the role of the rare biosphere. *PloS ONE* **5**: e12414.

Youssef N, Steidley BL, Elshahed MS. (2012). Novel high-rank phylogenetic lineages within a sulfur spring (Zodletone spring, Oklahoma, USA) revealed using a combined pyrosequencing/Sanger approach. *Appl Environ Microbiol* **78**: 2677–2688.

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