The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the candidate superphylum Parcubacteria (OD1)

Rosa León-Zayas,¹,² Logan Peoples,¹ Jennifer F. Biddle,² Sheila Podell,¹ Mark Novotny,³ James Cameron,⁴ Roger S. Lasken³ and Douglas H. Bartlett¹*
¹Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, 92093, USA.
²School of Marine Science and Policy, University of Delaware, Lewes, DE, 19958, USA.
³Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA, 92121, USA.
⁴Earthship Productions, 3806 Cross Creek Rd. Suite D, Malibu, CA, 90265-4975, USA.

Summary
Candidate phyla (CP) are broad phylogenetic clusters of organisms that lack cultured representatives. Included in this fraction is the candidate Parcubacteria superphylum. Specific characteristics that have been ascribed to the Parcubacteria include reduced genome size, limited metabolic potential and exclusive reliance on fermentation for energy acquisition. The study of new environmental niches, such as the marine versus terrestrial subsurface, often expands the understanding of the genetic potential of taxonomic groups. For this reason, we analyzed 12 Parcubacteria single amplified genomes (SAGs) from sediment samples collected within the Challenger Deep of the Mariana Trench, obtained during the Deepsea Challenge (DSC) Expedition. Many of these SAGs are closely related to environmental sequences obtained from deep-sea environments based on 16S rRNA gene similarity and BLAST matches to predicted proteins. DSC SAGs encode features not previously identified in Parcubacteria obtained from other habitats. These include adaptation to oxidative stress, polysaccharide modification and genes associated with respiratory nitrate reduction. The DSC SAGs are also distinguished by relative greater abundance of genes for nucleotide and amino acid biosynthesis, repair of alkylated DNA and the synthesis of mechanosensitive ion channels. These results present an expanded view of the Parcubacteria, among members residing in an ultra-deep hadal environment.

Introduction
Ocean sediments harbour a large number of diverse microorganisms, including the presence of numerous candidate phyla (CP) within the large candidate phyla radiation of the tree of life (Whitman et al., 1998; Schauer et al., 2010; Nunoura et al., 2012; Hug et al., 2016). Underexplored areas of ocean sediment include hadal oceanic trenches, such as the Challenger Deep, the deepest ocean location on earth. It is located in the western Pacific Ocean, extending to a depth of approximately 10 920 m (Nakanishi and Hashimoto, 2011), corresponding to about 110 megapascals (MPa) of hydrostatic pressure. This environment has been shown to have high microbial activity in surface sediments based on oxygen utilization (Glud et al., 2013) but little is known about the diversity of its microbial community (Kato et al., 1998; Pathom-Aree et al., 2006; Yoshida et al., 2013; Tarn et al., 2016). During the Deepsea Challenge (DSC) Expedition a sediment core was obtained by the manned submersible Deepsea Challenger, and from it single cell-derived genomes were obtained, including a number derived from members of the candidate Parcubacteria (OD1) superphylum. These genomes provide the opportunity to examine the genome characteristics of the Parcubacteria CP in the context of an extreme habitat.

Currently, the Parcubacteria stands out among CP as one of the most studied due to its abundance in many different anoxic marine and terrestrial environments (Harris et al., 2004; Elshahed et al., 2005; Gihring et al., 2011; Peura et al., 2012). Metabolic information has been acquired from metagenomic composite genomes and by using single cell genomicos (Rinke et al., 2013;
et al. (2013). It was recovered from brackish water present at 120 m depth in Sakinaw Lake, British Columbia, Canada. Its genome is 0.6 Mb in size and estimated to be 70% complete (Rinke et al., 2013). Candidatus Paceibacter normanii (single cell AAA255-P19) was designated as a Candidatus type species for the Parcubacteria superphylum (Rinke et al., 2013). Its genome is 0.6 Mbp in size and estimated to be 70% complete (Rinke et al., 2013). Candidatus P. normanii appears to have very limited metabolic potential highlighted by the lack of genes involved in sugar and amino acid degradation, complete TCA cycle, the pentose phosphate pathway, pyruvate metabolism and the electron transport chain.

Additional Parcubacteria genomes have been recovered from metagenomes of an acetate-amended aquifer sediment, including the sublineage OD1-i, which also possess a relatively reduced genome in terms of metabolic potential, lacking the tricarboxylic acid cycle (TCA) cycle and oxidative phosphorylation components (Wrighton et al., 2012). With a mostly fermentative metabolism, this sublineage is predicted to utilize acetyl-CoA synthetase for ATP generation and to reoxidize NADH produced during glycolysis by converting pyruvate to D-lactate and acetyl-CoA to ethanol. Another nearly complete Parcubacteria genome sequence, designated RAAC4, was also recovered via genome reconstruction from a metagenome of acetate-amended sediments (Kantor et al., 2013). As with other Parcubacteria, RAAC4 lacks a TCA cycle and respiratory chain enzymes and appears to be a strictly fermentative anaerobic organism. Although it does contain genes associated with the pentose phosphate pathway and a modified Embden–Meyerhof–Parnas (EMP) pathway, and appears to be able to utilize complex organic carbon and perhaps to create biofilms, this genome sequence information has reinforced the conclusion that members of the Parcubacteria have limited numbers of metabolic genes (Kantor et al., 2013).

Most recently the Parcubacteria were described as a superphylum following the reconstruction and annotation of 429 member genomes from an acetate amended groundwater aquifer at Rifle, CO (Brown et al., 2015). One of the most significant findings derived from this massive dataset was the discovery of a great phylogenetic radiation within the Parcubacteria (14 phyla described, more undescribed). In addition, many of the genomes encode self-splicing introns within their 16S rRNA genes and lack a number of ribosomal protein genes (Brown et al., 2015). The difference in ribosomal structure suggests that Parcubacteria may have evolved alternative mechanisms for ribosome regulation and function; and as a result of such differences have been likely under sampled in environmental surveys (Brown et al., 2015). In addition to having small genomes, microscopic analyses suggest that Parcubacteria cells have extremely small cell volume size of 0.009 ± 0.002 μm³ (Luef et al., 2015). Another recent study proposed the Parcubacteria to be symbiotic organisms based on their small genomes and lack of central biosynthetic pathways, although some novel metabolic genes were also described associated with aerobic respiration, particularly genomes C7867–007 and C7867–008 (Nelson and Stegen, 2015). The potential symbiotic nature of the OD1 CP was also highlighted by the discovery of a fresh water protist, Paramoecium bursaria, in association with bacteria from the Parcubacteria superphylum (Gong et al., 2014). Most recently the description of the phylum Parcunitrobacteria suggested that the genome Candidatus Parcunitrobacter nitroensis is a member of the Parcubacteria that may have the potential to oxidize hydroxylamine (NH2OH) to nitric oxide (NO) coupled to NO reduction (Castelle et al., 2017). Castelle and colleagues reported the first genome with respiratory potential based on genes encoding a modified oxidative phosphorylation pathway. This finding shows further variability in metabolic potential exists among members of the Parcubacteria superphylum.

Here we present a comparative genomic analysis of twelve Parcubacteria single cell genomes from cells collected within Challenger Deep sediment. The results indicate that deep-sea sediment Parcubacteria belong to diverse phylogenetic groups within the superphylum and while they share many metabolic pathways with previously analyzed genomes, they also expand the known metabolic potential of Parcubacteria.

Results and discussion

Genomic properties

From the recovered sample, 3520 total cells were sorted, 704 cells were subjected to MDA, and 494 MDA reactions were positively amplified as identified by subsequent 16S rRNA gene amplification. Among the known phyla, Cyanobacteria from the genus Prochlorococcus and Alphaproteobacteria, were the most abundant lineages detected (Supporting Information Fig. S1). This trench axis is known to host greater amounts and younger and more labile organic matter than sediment located along the trench wall of the incoming plate (Glud et al., 2013). The cyanobacterial sequences are interpreted as evidence of the recent vertical transport of fresh organic matter from sunlit shallow waters (Agusti et al., 2015).

The CP with the most number of cells detected, Parcubacteria, had a total of 20 cells identified by 16S rRNA gene sequences with >85% identity to Parcubacteria representatives. This amounted to approximately 5.4% of the total single cells analyzed for the study (Supporting Information Fig. S1). Of these, 12 genomes were further analyzed (Table 1). Sequences recovered from SAgs ranged from 0.4 to 1.0 Mbp and genome completeness ranged from 25% (0.40 Mbp) to 66% (0.47 Mbp) based on
recovery of conserved, single copy marker genes including various ribosomal proteins (Parks et al., 2015). The estimated genome size for these cells averaged 1.39 Mbp, which is at the higher end of the range known for Parcubacteria (Rinke et al., 2013). Percent GC ranged from 34.1% to 45.6%, with an average of 39.2% GC, similar to that of the proposed type species Candidatus P. normanii (39%).

Open reading frame predictions ranged from 395 to 1048 genes; of these 32%–56% were of uncharacterized function (Table 1). 16S rRNA genes were recovered from the Illumina HiSeq 2500 sequencing for all but one genome and tRNA gene counts ranged from 15 to 40 per genome. Pairwise average nucleotide identity (ANI) relationships among the DSC SAGs ranged from 64% to 93%, with the scores of relatedness averaging in the low 70s and only DSC5 and DSC8 showing 93% relatedness (Supporting Information Table S1). This suggests that none of the genomes are related enough to be the same species, although DSC5 and DSC8 are near the proposed cutoff (Konstantinidis and Tiedje, 2005). The presence or absence of ribosomal proteins, including ribosomal protein L1, has been found to be a major distinguishing factor between Parcubacteria, creating OD1 and OD1-L1 lineages (Brown et al., 2015). The DSC genomes, while incomplete, are all missing ribosomal protein L30, and most lack ribosomal protein L1 and GTPase Der (except DSC5, 7 for L1 and DSC6 for Der). The loss of these genes has previously been noted in other Parcubacteria (Brown et al., 2015). Although the genomes are incomplete, DSC2 and DSC6 may be the first example of genomes that cluster within the OD1 group, but lack the ribosomal protein L1 gene. Interestingly, none of the DSC genomes recovered encoded catalytic RNAs within their 16S rRNA genes (Brown et al., 2015).

**Phylogenetic relationships**

When comparing the 16S rRNA genes to the NCBI nt database using BLASTN, the top hit percent identity for each DSC SAG 16S rRNA gene ranged from 83% to 97% similarity to environmental clone sequences, suggesting that DSC genomes are distantly related to other previously described Parcubacteria. Phylogenetic analyses were performed for the eleven genomes for which 16S rRNA genes were recovered in the sequenced genome and one of sequences from the preliminary MDA screen (Fig. 1).

Nine DSC Parcubacteria 16S rRNA gene sequences are closely related to environmental sequences associated with deep-sea environments. Homology to known 16S rRNA genes was generally low, with the only 97% identity score matching a deep-sea sediment sample obtained at a water column depth greater than 5000 m (DSC2, Schauer et al., 2010). Four of the SAG 16S rRNA genes are most closely related to sequences obtained from deep-sea sediments in the South Atlantic, Guinea Basin [DSC2 (97% similarity), 6 (89%), 7 (88%), and 9 (92%); Schauer et al., 2010; > 5000 m], two others are closely related to samples from sub-seafloor sediment of the South China Sea [DSC8 (87%) and 10 (84%); Tao et al., 2008; 3697 m below the sea surface and 0.1m below the seafloor], DSC1 is related to samples recovered from a canyon slope in the Eastern Mediterranean Sea at 3603 m (93%; Polymenakou et al., 2009) and DSC4 (93%) and 5 (96%) are closely related to sediments within the Japan Trench at depths of 7111 and 6379 m respectively (Li et al., 1999; Hori et al., 2013). No previous suggestion of deep sea relatedness across Parcubacteria has been noted, yet these new sequences from the Mariana Trench clearly show significant relationships to deep sea relatives. Although pairwise ANI analysis suggested that DSC5 and DSC8 may be within species level relatedness, their phylogeny suggests they are likely separate species.

**Table 1.** Single cell genome descriptions.

<table>
<thead>
<tr>
<th>Group</th>
<th>Study name</th>
<th>Genome size</th>
<th>% complete</th>
<th>Est. genome size</th>
<th># contigs</th>
<th>Coding base count %</th>
<th>Gene count</th>
<th>HTGs (#per genome)</th>
<th>% of Parcubacteria proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD1-L1</td>
<td>OD1DSC1</td>
<td>825670</td>
<td>58</td>
<td>1423568</td>
<td>68</td>
<td>88</td>
<td>881</td>
<td>8</td>
<td>83%</td>
</tr>
<tr>
<td>OD1DSC3</td>
<td>473888</td>
<td>66</td>
<td>718912</td>
<td>18</td>
<td>88</td>
<td>524</td>
<td>4</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC4</td>
<td>600239</td>
<td>39</td>
<td>1539074</td>
<td>48</td>
<td>87</td>
<td>601</td>
<td>8</td>
<td>83%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC8</td>
<td>380175</td>
<td>41</td>
<td>927256</td>
<td>38</td>
<td>86</td>
<td>395</td>
<td>4</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC9</td>
<td>473882</td>
<td>31</td>
<td>1528651</td>
<td>40</td>
<td>86</td>
<td>496</td>
<td>5</td>
<td>81%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC10</td>
<td>546935</td>
<td>66</td>
<td>828689</td>
<td>37</td>
<td>88</td>
<td>581</td>
<td>7</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC11</td>
<td>599599</td>
<td>39</td>
<td>1537433</td>
<td>82</td>
<td>83</td>
<td>651</td>
<td>6</td>
<td>82%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC12</td>
<td>600202</td>
<td>50</td>
<td>1381204</td>
<td>79</td>
<td>86</td>
<td>658</td>
<td>8</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>OD1-DSC</td>
<td>OD1DSC5</td>
<td>406664</td>
<td>25</td>
<td>1626656</td>
<td>47</td>
<td>88</td>
<td>397</td>
<td>10</td>
<td>67%</td>
</tr>
<tr>
<td>OD1DSC6</td>
<td>1028328</td>
<td>59</td>
<td>1742928</td>
<td>68</td>
<td>86</td>
<td>1048</td>
<td>12</td>
<td>76%</td>
<td></td>
</tr>
<tr>
<td>OD1DSC7</td>
<td>758051</td>
<td>41</td>
<td>1848904</td>
<td>69</td>
<td>86</td>
<td>770</td>
<td>11</td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td>Uhrbacteria</td>
<td>OD1DSC2</td>
<td>520375</td>
<td>32</td>
<td>1626171</td>
<td>45</td>
<td>86</td>
<td>595</td>
<td>10</td>
<td>77%</td>
</tr>
</tbody>
</table>

Statistics are based on CheckM (Parks et al., 2015).
OD1 single cell genomes from the challenger deep 2773

Fig. 1. 16S rRNA gene-based phylogenetic tree of the DSC OD1 SAGs and additional Parcubacteria superphylum members. Maximum likelihood phylogenetic tree of 16S rRNA gene sequences obtained from eleven single amplified genomes (orange), one MDA screen (DSC9), draft genomes recently described from amended subterranean aquifers as part of the Parcubacteria superphylum (rRNA genes > 1000 bp, Brown et al., 2015) and previously described single cell and composite genomes (green; Wrighton et al., 2014, Rinke et al., 2013). The OD1-DSC clade falls into a separated unnamed/unclassified lineage, a division supported by additional phylogenies (Supporting Information Figs S3–S5). Scale bar represents 0.04 changes per position. Confidence values above 50% are shown at the tree nodes. The tree was rooted using Escherichia coli K12.

Following the proposed nomenclature by Brown et al. (2015), based on 16S rRNA genes most of the DSC genomes belong to unknown or unclassified clades/phyla, with one exception, as genome DSC2 falls within the Uhrbacteria phylum (Fig. 1). The Parcubacteria superphylum is separated into two major groups, designated as OD1 and OD1-L1 (lack ribosomal protein L1). The DSC genomes fall within both major groups, with eight of them within the OD1-L1 group (DSC1, 3, 4, 8, 9, 10, 11, 12) and four of them within the OD1 group (DSC2, 5, 6, 7) (Fig. 1). OD1-L1 is a large subgroup from the Parcubacteria subphylum encompassing most of the initially described Parcubacteria genomes including those from Candidatus P. normanii and RAAC4. Genomes DSC8 and 10 are the SAGs whose 16S rRNA gene sequences are most similar to Candidatus P. normanii (Rinke et al., 2013). Of the four DSC genomes that reside within the other major subgroup of the Parcubacteria superphylum (OD1), 16S rRNA genes from DSC5, 6 and 7 form a distinct cluster associated with other deep sea environmental 16S rRNA gene sequences of unknown lineage, herein called the OD1-DSC1 clade. Phylogenetic associations of the OD1-DSC1 clade were further investigated by analysing other conserved phylogenetic gene markers, including DNA gyrase B, recombinase A and RNA polymerase B (see Supporting Information Figs S3, S4 and S5), and the results support the 16S rRNA-based relationships (Fig. 1). However, a clear separation between subclades of the OD1-DSC1 clade was not always reproduced when using the marker genes. This was due, in part, to missing marker genes in the OD1-DSC1 genomes.

To assess the relationship of the SAGs from a protein relatedness perspective within the Parcubacteria superphylum and among each other, taxonomic classification of DSC predicted proteins were recovered via DarkHorse BLASTP analysis (Podell and Gaasterland, 2007). This analysis specifically targeted proteins that have taxonomic matches to previously characterized Parcubacteria phyla using the NCBI nr database of December 2016 (Adlerbacteria, Andersenbacteria, Azambacteria, Buchanabcacteria, Brennerc bacteria, Campfellbacteria, Colwellbacteria, Falkowbacteria, Giovanninibacteria, Harri sonbacteria, Jacksonbacteria, Jorgensennbacteria, Kaiserbacteria, Kerkfeldbacteria, Komellibacteria, Kuenen bacteria, Liptonbacteria, Llyodbacteria, Magasanikbacteria, Moranbacteria, Nealonbacteria, Nyogibacteria, Nomurabacteria, Other Parcubacteria, Portnobybacteria, Ryanbacteria, Spechtbacteria, Staskawiczbacteria, Sung bacteria, Tagabacteria, Taylorbacteria, Terrybacteria, Uhrbacteria, Veblenbacteria, Vogelbacteria, Wildermuthbacteria, Wolfebacteria, Yanofskybacteria, Yonathbacteria and Zambryskibacteria). The number of predicted proteins in the DSC genomes that are associated with Parcubacteria ranges from 67% (DSC5) to 88% (DSC10) of the total predicted proteins of each genome (Table 1). Based on the percent of shared predicted proteins with Parcubacteria, DSC genomes cluster into groups that are similar to those predicted by the 16S rRNA gene phylogeny, yet are clearly distinct from each other, reinforcing the 16SrRNA gene-based data suggesting that large phylogenetic distances exist between these groups (Fig. 2A). The taxonomic distribution of protein BLAST matches varies between the different DSC groups (Fig. 2B). The analysis suggests that the DSC2 genome shares the most number of proteins with the Uhrbacteria phylum and DSC8 and 10 match the Wildermuthbacteria phylum (Fig. 2B). DSC5, 6 and 7 are more similar to Buchanabcacteria phylum, although DSC5 also shares numerous proteins with the Kerfeldbacteria phylum, clustering farther away from DSC6 and 7 (Fig. 2A). DSC4, 11 and 12 appear to belong to the Spechtbacteria (Fig. 2B). The DSC1, 3 and 9 genomes protein matches are not dominated by any particular group, but all share similar relative abundance to the Campfellbacteria, Kaiserbacteria and Nomurabacteria phyla, as well as Parcubacteria genomes C7867 analyzed by Nelson and Stegen (Nelson and Stegen, 2015). As such, the metabolic analysis of these new DSC genomes provides especially valuable new insights about these poorly understood microbial groups.

Metabolic comparisons

Considering that even when all the new DSC genomes are combined they still only yield 92.8% genome completion based on conserved marker genes, inferences about metabolism and cell structure are limited by the possibility of missing characteristics due to the incompleteness of this dataset. KEGG pathway metabolic reconstruction of the DSC genomes suggests that the cells are heterotrophic microbes with limited substrate utilization ability. As a

© 2017 Society for Applied Microbiology and John Wiley & Sons Ltd, Environmental Microbiology, 19, 2769–2784
community, the DSC genomes have the potential for glycolysis via the Embden–Meyerhof–Parnas (EMP) pathway, which is the main pathway for the conversion of glucose to pyruvate in order to generate energy. All EMP genes but one are present within the annotated proteins of the total dataset, the exception being phosphofructokinase (PFK). Bacterial PFK protein sequences retrieved from the NCBI conserved domains were searched against the DSC genomes by BLASTP and TBLASTN (data not shown), but no homologs were found in any of the SAGs. PFK is also absent from OD1-i and Candidatus P. normanii genomes, whereas RAAC4 has all of the EMP enzymes. Among the available Parcubacteria genomes in the Integrated Microbial Genomes web platform (IMG, Markowitz et al., 2014, 448 genomes at the time of the analysis), only six possess PFK. The lack of PFK in many of the Parcubacteria could reflect a primary EMP role for gluconeogenesis rather than glycolysis, using this anabolic pathway to synthesize sugar.

Fig. 2. Taxonomic classification of annotated protein within the Parcubacteria superphylum.
A. Non-metric multidimensional scaling (nMDS) of the taxonomic classification of annotated proteins within the Parcubacteria superphylum were generated from the DarkHorse analysis. DSC protein annotated as Parcubacteria per SAG range from 67% (DSC5) to 88% (DSC10) of the total of protein in the genomes. DSC genomes (in bold black) and comparison genomes (in light black; Kantor et al., 2013, Brown et al., 2015; Castelle et al., 2017) are ordinated in the 2D plot. Based on protein classifications the DSC Parcubacteria reinforces the 16S data, as they group similarly when looking the taxonomic classifications. Stress value = 0.01183975.
B. Bar plot shows the distribution of Parcubacteria taxa in SAGs. Parcubacteria taxa shown make up 5% or more than the total protein matches and include proteins identified as archaeal, viral, non-Parcubacteria bacterial matches and un-assigned proteins.

© 2017 Society for Applied Microbiology and John Wiley & Sons Ltd, Environmental Microbiology, 19, 2769–2784
molecules from pyruvate. However, many previously described Parcubacteria genomes also lack other EMP genes whose products act early in the pathway, while possessing those responsible for the second phase of the pathway (the energetic payoff phase), although it is important to keep in mind that the great majority of the genomes analyzed are incomplete. Given that the DSC genomes possess other essentially irreversible enzymes involved in the first phase in the EMP pathway such as glucokinase (DSC1, 3, 7 and 10) and pyruvate kinase (DSC2, 6 and 11), it may be that the DSC Parcubacteria accomplish fructose-6-phosphate conversion to fructose-1,6-bisphosphate using a novel, yet to be characterized PFK. With this in mind, we probed the DSC genomes for proteins from the PFK superfamily using the most diverse sequences from the conserved domains, particularly ribokinase/pfkB superfamily (cd00287). Homology search by BLASTP revealed that DSC genomes (DSC1, 5, 6, 8, 10, 11) possess genes for a sugar kinase with a pfkB pfam domain (e-values 2e-07 to 1e-103; Supporting Information Fig. S8). These genes could be candidates for a novel PFK enzyme found in the Parcubacteria, as homologs are also found annotated as a hypothetical or sugar kinase gene in other Parcubacteria genomes in IMG.

It has previously been proposed that the Parcubacteria superphylum is strictly fermentative and obligately anaerobic (Wrighton et al., 2012). The capacity for fermentative metabolism is evident in the DSC genomes, as reflected in the presence of genes for lactate and alcohol dehydrogenases, glyoxylate reductase and acetate kinase (DSC1, 5, 6, 7, 8 and 10). Genes coding for enzymes in the tricarboxylic acid (TCA) cycle are less abundant, especially in the OD1-DSC1 clade. The TCA cycle is the process by which pyruvate is converted to carbon dioxide (CO₂) to generate energy and reducing equivalents are created along with precursors for biosynthesis. The cycle enzymes encoded within the DSC genomes include 2-oxoglutarate: ferredoxin oxidoreductase (KorA and KorB; DSC5 and 11) and succinyl-CoA synthetase, citryl-CoA synthetase and citryl-CoA lyase (DSC6) (Supporting Information Table S2). These enzymes, although not present in most of the genomes, have been reported to be present in 3%-8% of the recently described Parcubacteria genomes (Brown et al., 2015). Many other enzymes that are required for a complete TCA cycle were not found in any of the DSC genomes. While all of the DSC single cell genomes are incomplete, they are unlikely to encode an entire TCA cycle, in agreement with other Parcubacteria genomes.

Subunits of NADH-dehydrogenase, the enzyme responsible for the first step in the electron transport chain, are found in DSC genomes (Supporting Information Table S2; DSC4, 6, 11 and 12). Ubiquinol-cytochrome c reductase cytochrome c1 subunit, part of complex III of the electron transport chain, is found in one genome from the OD1-DSC1 clade (DSC6) (Supporting Information Table S2). A few DSC genomes also contain genes coding for heme/copper-type cytochrome/quinol oxidases (DSC1, 10 and 11), which are involved in oxidative phosphorylation and the utilization of oxygen as a terminal electron acceptor. The DSC6 and 7 genomes also encode cytochrome c oxidase subunit II (oxygen-reducing terminal oxidase) (Supporting Information Table S2). This has also been noted in other Parcubacteria genomes, however as with the previous genomes, genes encoding the other two subunits of cytochrome C oxidase, cytochrome c oxidase subunit I (containing the catalytic domain) and cytochrome c oxidase subunit III (whose activity is not fully understood), are not present (Brown et al., 2015).

Intriguing aspects of nitrogen metabolism also exist within the OD1-DSC1 genomes. The DSC6 SAG contains...
a complete nitrate reductase operon in addition to a nitrous oxide reductase (Fig. 3, Supporting Information Table S2). Phylogenetic analyses of the active subunit of the DSC6 nitrate reductase support a respiratory rather than assimilatory function (Supporting Information Fig. S2). Most predicted proteins in the nitrate reductase operon are most closely related to genes present in Enhydrobacter aerosaccus, a member of the Gammaproteobacteria present in many environments, including marine (Staley et al., 1987; Irgens et al., 1989; Khandeparker and Anil, 2013; Leong et al., 2015; Fig. 3, Table 2, Supporting Information Fig. S2). The genes associated with nitrate reduction are present on two contigs from the DSC6 genome. Contig 1 contains the alpha subunit of the enzyme, a nitrate transporter and two transposable elements, suggesting a potential for horizontal transfer. Contig 2 contains the other nitrate reductase subunits and cofactors (Fig. 3, Table 2). The DSC2 genome encodes subunits of a nitrite reductase nirD used in the dissimilatory nitrite reduction pathway to ammonia. In addition, a number of nitrate/nitrite transporters are present, particularly in DSC5 and 6 (Supporting Information Table S2).

Genes are also present in the DSC genomes for the F type ATPase, whose main role is to catalyse the synthesis of ATP using energy generated by cellular respiration (Yoshida et al., 2001; Supporting Information Table S2; DSC1, 4, 5, 7, 10 and 11). They are also found in many of the previously described Parcubacteria (Kantar et al., 2013; Brown et al., 2015). A V type ATPase is encoded within one genome belonging to OD1-DSC1 clade (Supporting Information Table S2; DSC6). In most cases these enzymes are thought to be used for ATP hydrolysis, but some microorganisms possess V type ATPases that are able to synthesise ATP (Toei et al., 2007). This type of respiratory potential has been previously reported for Candidatus P. nitroensis (Castelle et al., 2017).

It is important to emphasize, taking all gene inventories described above into account, that only partial pathways of known respiration-linked systems exist within the DSC genomes. Nevertheless, members of the OD1-DSC1 clade (particularly DSC6) possess intriguing signs of respiration-related capacity, due to their high proportions of respiration genes in comparison to the majority of the publically available Parcubacteria genomes. In this regard they are similar to the Parcubacteria genomes C7867–007, C7867–008 and Candidatus P. nitroensis (Nelson and Stegen, 2015; Castelle et al., 2017). If these gene systems in the DSC genomes are not used for energy-yielding respiration they may serve other functions involving electron/ion/solute transport or oxidative stress adaptation.

Several lines of evidence suggest that the genes connected to respiration in DSC6 are non-contaminants. For example, E. aerosaccus is not a common contaminant found in MDA amplified genomes, no other genes with homology to E. aerosaccus were found in DSC6 or in any of the other SAGs, and the lineage probability index (LPI) scores for the genes within the nitrate reductase contigs (0.487–0.558) are in the same range of scores from other non-Parcubacteria genes found through numerous contigs. Based on kmer frequency, scaffolds that encode for the nitrate reductases are closely clustered with the majority of scaffolds in the genome (Supporting Information Fig. S6). The DSC6 V type ATPase has highest homology to that present in Clostridia-like organisms and its cytochrome C proteins have highest homology to those of Microgenomates-like organisms, perhaps reflecting additional sources of horizontal gene transfer or evolutionarily divergence from other Parcubacteria genomes.

Another important feature present in many of the DSC genomes concerns their biosynthetic capacity. All enzymes involved in the pentose phosphate pathways (PP) are represented in the OD1-L1 subgroup genomes, as is the case with up to 67% of the Parcubacteria genomes available in IMG (Supporting Information Table S2). The PP pathway generates NADPH and pentose, which are primarily utilized for anabolic purposes such as nucleotide synthesis. All of the DSC genomes share, along with 24% of other Parcubacteria, the presence of genes involved in the biosynthesis of purines and pyrimidines (Supporting Information Table S2). The coupled presence of PP and purine/pyrimidine biosynthetic pathways suggests that the DSC cells are capable of nucleotide biosynthesis, in contrast to the majority of Parcubacteria. It is interesting that DSC genomes have, among all 12 SAGs, 169 genes associated with amino acid metabolism, among those biosynthetic genes for glycine, serine, cysteine, lysine, and glutamine (Supporting Information Table S2). Most other Parcubacteria genomes are missing the capability of synthesizing amino acids (Brown et al., 2015).

For metabolic comparisons at a whole genome level, the metabolic potential (all genes) of each DSC genome was compared against 232 Parcubacteria genomes that belong to one of the described phyla available in IMG (Supporting Information Fig. S7, Supporting Information Table S3). The homology between the DSC SAGs and the genomes ascribed to the different Parcubacteria phyla varied. Metabolic potential from genomes in phyla Uhrbacteria and Falkowbacteria are more highly represented in the OD1-DSC1 clade. Clustering patterns based on metabolic relatedness do not exactly mirror the 16S rRNA gene phylogeny. For example, DSC3 and DSC4 are the most similar to each other, and to the Parcubacteria dataset from IMG, based on metabolic potential. DSC2 shares the least amount of metabolic potential with the Parcubacteria dataset, followed by the OD1-DSC1 clade. The metabolic processes shared most frequently (found to be present in 90% or more of the Parcubacteria genomes), were genes for ribosomal proteins, ATP-dependent Clp protease ATP-
Table 2. Nitrate reductase gene cluster homology information.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Start</th>
<th>End</th>
<th>Sense</th>
<th>Prokka ID</th>
<th>EC</th>
<th>Gene</th>
<th>Prokka annotation</th>
<th>NCBI BlastP annotation</th>
<th>Max score</th>
<th>Total score</th>
<th>Query cover</th>
<th>E value</th>
<th>Ident</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>108</td>
<td>1367</td>
<td>+</td>
<td>PROKKA_00706</td>
<td>1.7.99.4</td>
<td>narH</td>
<td>Respiratory nitrate reductase 1 beta chain [E. aerosaccus]</td>
<td>nitrate reductase [E. aerosaccus]</td>
<td>870</td>
<td>870</td>
<td>100%</td>
<td>0</td>
<td>99%</td>
<td>WP_050324468.1</td>
</tr>
<tr>
<td>2</td>
<td>1369</td>
<td>2289</td>
<td>+</td>
<td>PROKKA_00707</td>
<td>narW</td>
<td>PROKKA_00707</td>
<td>nitrate reductase molybdenum cofactor assembly chaperone NarW [Pausebacillus sophorae]</td>
<td>nitrate reductase [Pausebacillus sophorae]</td>
<td>574</td>
<td>574</td>
<td>91%</td>
<td>0</td>
<td>98%</td>
<td>WP_051500173.1</td>
</tr>
<tr>
<td>3</td>
<td>2286</td>
<td>2984</td>
<td>+</td>
<td>PROKKA_00708</td>
<td>1.7.99.4</td>
<td>narV</td>
<td>Respiratory nitrate reductase 2 gamma chain [Moraxella boevrei]</td>
<td>nitrate reductase [Moraxella boevrei]</td>
<td>361</td>
<td>361</td>
<td>99%</td>
<td>6.00E-123</td>
<td>77%</td>
<td>WP_019519315.1</td>
</tr>
<tr>
<td>4</td>
<td>3455</td>
<td>4639</td>
<td>+</td>
<td>PROKKA_00709</td>
<td>1.14.13.-</td>
<td>ubil</td>
<td>2-octaprenylphenol hydroxylase [M. boevrei]</td>
<td>hypothetical protein [M. boevrei]</td>
<td>543</td>
<td>543</td>
<td>98%</td>
<td>0</td>
<td>64%</td>
<td>WP_019519414.1</td>
</tr>
<tr>
<td>5</td>
<td>4950</td>
<td>6188</td>
<td>-</td>
<td>PROKKA_00710</td>
<td>narT</td>
<td>PROKKA_00710</td>
<td>putative nitrate transporter NarT [E. aerosaccus]</td>
<td>MFS transporter [E. aerosaccus]</td>
<td>823</td>
<td>823</td>
<td>100%</td>
<td>0</td>
<td>99%</td>
<td>WP_050324473.1</td>
</tr>
<tr>
<td>6</td>
<td>6558</td>
<td>8633</td>
<td>+</td>
<td>PROKKA_00711</td>
<td>2.7.13.3</td>
<td>narX</td>
<td>Nitrate/nitrite sensor protein NarX [E. aerosaccus]</td>
<td>nitrate/nitrite two-component system sensor histidine kinase [E. aerosaccus]</td>
<td>1396</td>
<td>1396</td>
<td>100%</td>
<td>0</td>
<td>98%</td>
<td>WP_050324474.1</td>
</tr>
<tr>
<td>7</td>
<td>8795</td>
<td>9454</td>
<td>+</td>
<td>PROKKA_00712</td>
<td>narL</td>
<td>PROKKA_00712</td>
<td>Nitrate/nitrite response regulator protein NarL [E. aerosaccus]</td>
<td>nitrate/nitrite response regulator [E. aerosaccus]</td>
<td>442</td>
<td>442</td>
<td>100%</td>
<td>3.00E-155</td>
<td>99%</td>
<td>WP_050324475.1</td>
</tr>
<tr>
<td>8</td>
<td>526</td>
<td>4293</td>
<td>-</td>
<td>PROKKA_00732</td>
<td>1.7.99.4</td>
<td>narG</td>
<td>Respiratory nitrate reductase 1 alpha chain [E. aerosaccus]</td>
<td>nitrate reductase [E. aerosaccus]</td>
<td>2621</td>
<td>2621</td>
<td>100%</td>
<td>0</td>
<td>99%</td>
<td>WP_050324469.1</td>
</tr>
<tr>
<td>9</td>
<td>4436</td>
<td>5542</td>
<td>-</td>
<td>PROKKA_00733</td>
<td>2.7.13.3</td>
<td>narK</td>
<td>Putative transposase DNA-binding domain protein [E. aerosaccus]</td>
<td>Putative transposase DNA-binding domain protein [E. aerosaccus]</td>
<td>704</td>
<td>704</td>
<td>99%</td>
<td>0</td>
<td>92%</td>
<td>WP_050323977.1</td>
</tr>
<tr>
<td>10</td>
<td>5682</td>
<td>6971</td>
<td>-</td>
<td>PROKKA_00734</td>
<td>NnrS</td>
<td>NnrS</td>
<td>hypothetical protein transposase [E. aerosaccus]</td>
<td>hypothetical protein transposase [E. aerosaccus]</td>
<td>840</td>
<td>840</td>
<td>100%</td>
<td>0</td>
<td>99%</td>
<td>WP_050324471.1</td>
</tr>
<tr>
<td>11</td>
<td>7236</td>
<td>7475</td>
<td>+</td>
<td>PROKKA_00735</td>
<td>2.7.13.3</td>
<td>narK</td>
<td>Nitrate/nitrite transporter NarK [E. aerosaccus]</td>
<td>Nitrate/nitrite transporter NarK [E. aerosaccus]</td>
<td>892</td>
<td>892</td>
<td>100%</td>
<td>0</td>
<td>99%</td>
<td>WP_050324472.1</td>
</tr>
</tbody>
</table>

OD1 single cell genomes from the Challenger Deep.
binding subunit ClpB, type II secretory pathway, type IV pilus assembly, cell division genes and peptidoglycan metabolism. These genes represent highly conserved metabolic processes that make up the core metabolic potential of the Parcubacteria. Some of the genes shared between <10% of all Parcubacteria genomes are abundant in these deep-sea SAGs. For example, the gene for DNA-3-methyladenine glycosylase I is shared by 8 of the DSC SAGs. This gene is involved in recognition and repair of alkylated DNA (Metz et al., 2007). The small-conductance mechanosensitive channel gene is another example, in this case shared across 6 of the SAGs. Small-conductance mechanosensitive channels are associated with protecting cells from increased turgor pressure (Lai et al., 2013), which we speculate could also facilitate cellular adaptation to elevated hydrostatic pressure (Bartlett, 2002). Other genes that are shared between five of the SAGs are predicted thiol-disulfide oxidoreductase YuxK, DCC family, and NAD(P)H-dependent FMN reductase, involved in a variety of redox reactions involving extracellular enzyme modification (Ingelman et al., 1999; Ginalska et al., 2004).

These analyses also revealed a variety of genes uniquely present in the SAGs (Table 3). Over 88% of the unique genes present in each genome encode hypothetical proteins. Among the identifiable unique genes are a number of transporters and membrane proteins, suggesting that interactions with environmental factors may determine particular differences between the SAGs.

**Cell surface characteristics and stress adaptation**

Various SAGs encode the potential for polysaccharide core modification, particularly that related to the synthesis of lipid A core found in lipopolysaccharides (Supporting Information Table S2). Among the genes present in the DSC cells are those for lipid A core-O-antigen ligase and related enzymes, 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase; EC:2.7.7.38; kdsB), 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase (KDO 8-P phosphatase; EC:3.1.3.45; kdsC), glycosyltransferases (pimB), as well as transport proteins for O-antigen and O-antigen ligase (Supporting Information Table S2). Out of 25 steps involved in the biosynthesis of the lipid A core, only the few mentioned above are present in the SAGs. Although, none of these genes are unique to the DSC SAGs, a few (e.g., 3-deoxy-manno-octulosonate cytidylyltransferase and 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase) are shared with 15% or less of all the available genomes in IMG and have not been previously documented. These results suggest that DSC OD1 cells have the potential to modify polysaccharides, perhaps facilitating membrane association.

Another feature not previously described within members of this CP, and that is shared between less than 10% of IMG available genomes, is the presence of genes associated with adaptation to oxidative stress. These features span all DSC OD1 genomes and include peroxiredoxin (DSC1, 3, 4, 6, 7, 10 and 11) and superoxide dismutase (DSC1, 2, 3, 4, 6, 7, 10 and 11) (Supporting Information Table S2). The presence of these enzymes implies some degree of tolerance to oxidizing conditions, perhaps making it possible for these cells to survive in the oxygen-exposed surficial sediment environment they were recovered from.

The DSC genomes encode both heat shock and cold shock proteins, which are also found in the majority of Parcubacteria genomes in IMG, among them the heat shock proteins involved in the sensing of misfolded proteins; DnaK, DnaJ and GrpE (Feder and Hofmann, 1999). Genes for cold shock proteins (CSP) appear to be less prevalent in previously sequenced Parcubacteria genomes (found in <50%), while CSP such as CspA, are found in the majority of DSC SAGs (DSC1, 2, 3, 4, 6, 7, 8, 10, 11 and 12; Supporting Information Table S2). CspA and its relatives are small proteins that bind to single-stranded polypyrimidine nucleic acids (Johnston et al., 2006) for the purpose of inhibiting potential deleterious mRNA secondary structures at low temperature (Jiang et al., 1997). This could be advantageous to the DSC cells growing at an in situ temperature of ~2°C–3°C.

**Horizontally transferred genes**

Each of the DSC OD1 genomes encodes some predicted proteins that have no apparent orthologs within the Parcubacteria, based on the LPI measurements for each gene using the DarkHorse program (Podell and Gaasterland, 2007). The LPI index evaluates the taxonomic similarity of BLAST matches over an entire genome, with lower values highlighting those genes matching only atypical lineages. In these cases, the genes may have been horizontally transferred, or they may represent genes that are part of the non-core pan-genome of taxonomic groups that have only one or a few database representatives.

The relative abundance of these possible HTGs varies from 0.92% (DSC3) to 3.12% (DSC2) of predicted proteins in the DSC genomes (Table 4). The most frequently observed group of potential HTGs are archaeal in origin, mostly originating in the *Methanomicrobia* and *Methanobacteria* classes within the *Euryarchaeota* (Supporting Information Table S4a). Bacterial proteins in each DSC genome with highest homology to non-Parcubacteria genomes possess LPI scores higher than 0.4 but lower than 0.7, suggestive of HGT, but less definitive. Predicted proteins with eukaryotic taxonomic relationships were less abundant and mostly consisted of hypotheticals. Three
DSC had closest matches to viral-like proteins (DSC2, 11 and 12; Supporting Information Table S4b). It is unlikely that these low LPI-scoring genes are artefacts associated with contaminating DNA that arose during single-cell isolation and DNA amplification, because most are located on contigs with high LPI scores in adjacent genes. HGT may provide varied metabolic potential to these deep-sea bacteria. It is also notable that over 51% of proposed archaeal HTGs are found in multiples per scaffold (two or more genes). GC% content of such scaffolds and individual genes also varied from the average genome GC content (Supporting Information Tables S4a and b). The GC sigma variation from the mean GC supports the possibility of horizontal transfer for many of the suggested HTGs (Supporting Information Tables S4a and b).

**Summary**

Parcubacteria reside in diverse habitats, including deep-sea sediments. Here we report on Parcubacteria within the
The genome analyses presented here reinforce the view of the Parcubacteria as organisms with small genomes that are able to metabolize organics by fermentation. While considered to have a small genomic size, they also appear capable of expanding their genome. The significance of the newly identified features to the lifestyle and environmental contributions of this enigmatic but abundant super phylum awaits further investigation.

**Experimental procedures**

**Collection and sorting**

Sediments were collected from a depth of 10,908 m and an in situ temperature of 2.5°C using a push-core apparatus controlled by a hydraulic arm attached to the manned submersible Deepsea Challenger. Sampling occurred on March 26, 2012 in the ‘East Deep’ (Fujikura et al., 2002) of the Challenger Deep at 142.59° E, 11.37° N during the Deepsea Challenge Expedition. Recovered sediment was placed in glycerol/TE buffer (Rinke et al., 2014) and first stored in liquid nitrogen and later at −80°C prior to single cell sorting at the J. Craig Venter Institute (JCVI). The sediment sample was gently vortexed and allowed to settle briefly before filtering through a 35 μm mesh (BD Biosciences, San Jose, CA, USA) to avoid larger sediment particles. Cells were stained with SYBR Green I nucleic acid stain (Invitrogen, Carlsbad, CA, USA) and sorted using a cooled FACS-Aria II flow cytometer (BD Biosciences, San Jose, CA, USA) (McLean et al., 2013). Microtiter plates of sorted single cells were stored at −80°C.

**Genome amplification and sequencing**

DNA was amplified using a custom BioCel robotic system (Agilent Technologies, Santa Clara, CA, USA) as described by McLean et al. (2013). Genomic material in the sorted microbial cells was amplified by multiple displacement amplification (MDA) in a 384-well format using GenomiPhi (GE Healthcare, Waukesha, WI, USA). 16S rRNA genes were PCR amplified from diluted MDA products using universal bacterial primers 27F and 1492R (Weisburg et al., 1991) as follows: 94°C for 3 min, 35 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 90 s, and 72°C for 10 min. PCR products were cleaned with exonuclease I and shrimp alkaline phosphatase (Thermo Fisher Scientific Inc., Waltham, MA, USA) and sent for Sanger sequencing at the Joint Technology Centre (JTC, J. Craig Venter Institute, Rockville, MD, USA). 16S rRNA gene trace files were analyzed with the CLC Workbench software program (CLC Bio, Cambridge, MA, USA). Gene sequences were evaluated for evidence of microbial DNA contamination associated with MDA reagents, based on an in-house JCVI database of common contaminants, including sequences related to the genera Escherichia, Propionibacterium, Shewanella and Pseudomonas (Supporting Information Table S5). Any single cells judged to be contaminated were removed from consideration for whole genome sequencing. These curated sequences were then compared with the NCBI nr/nt database using BLASTN (Altschul et al., 1990) for initial taxonomic screening of the sorted cells. DNA from thirteen separate Parcubacteria-classified cells was sent for Illumina HiSeq 2500 (101 bp reads) sequencing at the JTC. Raw data is available from the project number PRJEB10905.

**Assembly, annotation and genome completion**

Genomes were quality trimmed using Nesoni with mostly default parameter and q20 for quality score (www.vicbioinformatics.com/software.nesonii.shtml). Sequences were assembled using SPAdes 3 with flags for single cell genome assembly (−sc) and to reduce mismatches, short indels and error corrections (−careful) (Bankevich et al., 2012) and annotated by IMG-ER (https://img.jgi.doe.gov/cgi-bin/er/main.cgi, Markowitz et al., 2014). Annotated genomes are accessible through IMG (Genome IDs: 2547132480, 2547132497–98, 2547132500–1, 2547132503–4, 2548877123, 2548877145–6, 2548877145–8–9). Only contigs over 1000 basepairs are reported here. All metabolic comparisons were generated using the genomes that were assembled with SPAdes 3 and annotated via IMG. Later, genome DSC6 was re-assembled using SPAdes 3.6 and annotated using PROKKA (Seemann, 2014). This assembly was performed as an attempt to utilize the latest assembly algorithm to potentially generate a better assembly; this data is only presented in Table 2 and Fig. 3. Estimated genome size and completeness were
calculated using CheckM (Parks et al., 2015). Annotations for genes of interest that are discussed in detail throughout the article were manually verified via BLASTP. Functional comparisons were performed using the IMG-ER platform (Markowitz et al., 2014). Protein sequences for DSC6 nitrate reductase gene, nartG, were compared with other nitrate reductase protein sequences in order to assess their potential role in assimilation or dissimilation of nitrate (Supporting Information Fig. S2). Representative sequences from the NCBI database were aligned using muscle (Edgar 2004) and an unrooted maximum likelihood tree was generated with FastTree (Price et al., 2009).

**Taxonomy assignments**

16S rRNA gene sequences recovered from each SAG were analyzed by BLASTN against the NCBI nr/nt database (Altenschl et al., 1990). Sequences with 85% or greater similarity to Parcubacteria superphylum 16S rRNA genes were extracted and used for phylogenetic reconstruction, along with reference sequences belonging to the Parcubacteria superphylum from previous publications (Wrighton et al., 2012; Kantor et al., 2013; Rinke et al., 2013; Brown et al., 2015). Sequences were aligned with the SINA aligner (http://www.arb-silva.de/aligner; Pruesse et al., 2007) and a maximum-likelihood tree was created using FastTree (Price et al., 2009).

Genome-encoded protein predictions of DNA fragments of > 1000 bp were obtained from IMG-ER and taxonomically classified using DarkHorse software, version 2.0 (http://darkhorse.ucsd.edu/; Podell and Gaasterland, 2007), with default settings. Annotated proteins from each of the genomes were compared against the NCBI nr protein database (December 2016) with Diamond (Buchfink et al., 2015). This data was used as input to the DarkHorse program, which identifies closest database relatives for individual proteins and constructs an overall statistical profile for the query genome. This profile is used to calculate a lineage probability index (LPI) score for each individual protein, indicating whether its taxonomic match is typical or atypical for the genome being investigated.

**References**


**Supporting information**

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

Fig. S1. Phylogenetic distribution and relative abundance of Challengee Deep MDAs by 16S rRNA sequence. Fourteen phyla were identified, including known phyla (circles) and candidate phyla (triangles). The total and percent abundance distribution is 150 Cyanobacteria (40.4%; dark green), 97 Alphaproteobacteria (26%; blue), 39 Gammaproteobacteria (10.5%; grey), 20 OD1 (5.4%; red), 19 Bacteroidetes (5.1%; yellow), 14 Actinobacteria (3.8%; orange) 10 SAR406 (3.7%; green) and the remaining groups were less than 2.5% abundance (purple; Firmicutes, Gemmatimonadetes, Chlamydiae, OP11, JS1, OP3, BD1-5, TM6).

Fig. S2. Nitrate reductase alpha subunit NarG protein tree. Protein sequences were aligned using muscle and trimmed to retain similar conserved regions within the proteins. Maximum likelihood was constructed using FastTree. DSC6 nitrate reductase gene is highlighted in orange. Respiratory cluster is highlighted in red branches, while assimilatory cluster is highlighted in blue branches. Scale bar represents 0.4 changes per position, confidence values are found at the nodes.

Fig. S3. Recombine A phylogenetic distribution among the OD1-DSC genomes. Recombine A is encoded by the single copy marker gene recA and serves as a conserved phylogenetic marker. ACD genes are from Wrighton et al. 2012 and Kantor et al. 2013 and were downloaded from http://ggkbase.berkeley.edu/. GBEA genes were downloaded from https://img.jgi.doe.gov/e/1. Protein amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are sequences derived from members of the Peregines (PER) phylum, which is related to the OD1 CP.

Fig. S4. RNA Polymerase subunit beta phylogenetic distribution among the OD1-DSC genomes. RNA Polymerase subunit beta is encoded by the single copy marker gene rpoB and serves as a conserved phylogenetic marker. ACD genes are from Wrighton et al. (2014) and Kantor et al. (2013) and were downloaded from http://ggkbase.berkeley.edu/. GBEA genes were downloaded from https://img.jgi.doe.gov/e/1. Protein amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are sequences derived from members of the Peregines (PER) phylum, which is related to the OD1 CP.

Fig. S5. DNA Gyrase subunit beta subunit beta phylogenetic distribution among the OD1-DSC genomes. DNA Gyrase subunit beta encoded by the single copy marker gene gyrB serves as a conserved phylogenetic marker. The ACD sequences are from Wrighton et al. (2012) and Kantor et al. (2013) and were downloaded from http://ggkbase.berkeley.edu.
GEBA genes were downloaded from https://img.jgi.doe.gov/er. Protein amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are sequences derived from members of the Peregrines (PER) phylum, which is related to the OD1 CP.

**Fig. S6.** Kmer frequency for scaffolds of genome DSC6. Kmer frequency was calculated in IMG for scaffolds of genome DSC6. Fragment window 50, 00 bp, fragment step 500 bp, oligomer size 4, minimum variation 10. Black circle highlight the scaffolds where nitrate reductase are found.

**Fig. S7.** Whole genome comparison to available Parcubacteria genomes in IMG. All genes within the > 1000 bp contigs of all SAGs (vertical columns) were compared in IMG to all available Parcubacteria genomes (horizontal lines). Scale values range from less similar (white) to highly similar (black). Darker black in a column reflect greater similarity among a SAG to the analyzed Parcubacteria genomes. The heatmap was clustered using a Bray–Curtis dissimilarity matrix. Abundance data shown is normalized based on genome size, colour scale values range 0.0002 to 0.0016.

**Fig. S8.** Supplementary document S8 shows the BLASTP analysis generated for the pfkB protein search within each of the DSC genomes.