Photosynthesis is not a universal feature of the phylum Cyanobacteria

Rochelle M Soo, Connor T Skennerton, Yuji Sekiguchi, Michael Imelfort, Samuel J Paech, Paul G Dennis, Jason A Steen, Donovan H Parks, Gene W Tyson, Philip Hugenholtz

Molecular surveys of aphotic habitats have indicated the presence of major uncultured lineages phylogenetically classified as members of the Cyanobacteria. One of these lineages has recently been proposed as a non-photosynthetic sister phylum to the Cyanobacteria, the Melainabacteria, based on recovery of standard draft population genomes from human gut and groundwater samples. Here, we expand the phylogenomic representation of the Melainabacteria through sequencing of six diverse population genomes from gut and bioreactor samples supporting the inference that this lineage is non-photosynthetic, but not the assertion that they are strictly fermentative. We propose that the Melainabacteria is a class within the phylogenetically defined Cyanobacteria based on robust monophyly and shared ancestral traits with photosynthetic representatives. Our findings are consistent with theories that photosynthesis occurred late in the Cyanobacteria and involved extensive lateral gene transfer, and refutes the long held belief that oxygenic photosynthesis is a universal feature of this phylum.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Authors

Rochelle M. Soo[1], Connor T. Skennerton[1,2], Yuji Sekiguchi[3], Michael Imelfort[1], Samuel J. Paech[1], Paul G. Dennis[1‡], Jason A. Steen[1], Donovan H. Parks[1], Gene W. Tyson[1,2], and Philip Hugenholtz[1,4†]

Affiliations

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia

[2]Advanced Water Management Centre, The University of Queensland, St Lucia, QLD 4072, Australia

[3]Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Central 6, Ibaraki 305-8566, Japan

[4]Institute for Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia

† Correspondence to: Philip Hugenholtz, Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia, +61 7 336 53822, p.hugenholtz@uq.edu.au

‡ Current address: School of Agriculture and Food Sciences, The University of Queensland, St Lucia, QLD 4072, Australia

Introduction

Cyanobacteria are recognised primarily for oxygenic photosynthesis (Nelson & Ben-Shem 2004) a feature that is thought to be common to all members of this phylum. However, oxygenic photosynthesis is widely thought to have originated after anoxygenic photosynthesis, likely well after the primary diversification of bacterial phyla (Hohmann-Marriott & Blankenship 2011). This suggests that the Cyanobacteria as a primary bacterial line of descent predate oxygenic photosynthesis. Consistent with this inference is the relatively shallow phylogenetic depth circumscribed by photosynthetic cyanobacteria compared to other bacterial phyla based on comparative rRNA analyses (Dojka et al. 2000).

Cultured Cyanobacteria are categorized into five subsections on phenotypic grounds according to the botanical code (Rippka et al. 1979). However, over the past decade, 16S rRNA-based culture-independent molecular surveys have greatly increased our awareness of the phylogenetic breadth of the Cyanobacteria with the identification of additional major lines of descent such as YS2/4C0d-2, mle1-12, SM2F09, SM1D11 and ML635J-21 (McDonald et al. 2012; Quast et al. 2013). These deep-branching Cyanobacteria have been found in numerous environments, including drinking water (Williams et al. 2004), grassland soil (Cruz-Martinez et al. 2009), wastewater treatment plants (Liu et al. 2006), and human and animal guts (Ley et al. 2006). Many of these habitats are aphotic, which suggests that a large number of organisms phylogenetically defined as Cyanobacteria are non-photosynthetic (McDonald et al. 2012; Quast et al. 2013).

Recently, Di Rienzi *et. al.*, obtained five near complete genomes of members of the YS2/4C0d-2 lineage and confirmed the absence of photosynthetic machinery in these representatives. Additionally, comparative analysis of a 16S rRNA gene sequence (genbank acc. HM038000) obtained from freeze-dried cells of *Vampirovibrio chlorellavorus* ATCC 29753 (Gromov & Mamkaeva 1980) indicates that this bacterium is a member of the SM1D11 lineage. The original description of this organism provided no indication of photosynthetic capability further supporting absence of photosynthesis in basal cyanobacterial lineages. Di Rienzi and colleagues proposed a new phylum, Melainabacteria ("Greek nymph of dark waters"), for

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

YS2/4C0d-2 and related basal lineages, given their deeply branching position relative to photosynthetic cyanobacteria and due to their lack of photosynthetic genes (Di Rienzi et al. 2013). To further explore the inferred properties of the Melainabacteria group and to assess whether they should be excluded from the cyanobacterial phylum, we obtained six near-complete genomes representing a broader phylogenetic coverage of the Melainabacteria (representing YS2/4C0d-2, mle1-12 and SM2F09). Comparative analyses of these genomes corroborate the assertion that the Melainabacteria is a non-photosynthetic lineage, however they are robustly monophyletic with the photosynthetic cyanobacteria with which they share inferred common ancestral traits, such as cell envelope structure. We therefore suggest that the Melainabacteria represent a class within the Cyanobacteria extending the recognized metabolic capacity of this phylum.

Methods

A schematic of the workflow used in this study is presented in **Fig. S1** with details as follows:

Sample collection and DNA extraction

Faecal samples from a 12 year old make koala (*Phascolarctos cereus*) named Zagget was collected in sterile 50ml falcon tubes on 12 May, 2011 (Zag_T1), 28 July, 2011 (Zag_T2) and 24 November, 2011 (Zag_T3), at Lone Pine Koala Sanctuary, Brisbane, Australia. Ethics approval for the collection of koala faeces was obtained from the Animal Welfare Unit, The University of Queensland, under ANRFA/074/11 *A study into Koala hindgut microbiology*. Samples were snap frozen in dry ice mixed with ethanol at the time of sampling and then transferred to -80°C until further processing. Genomic DNA was extracted from faeces using a MP-BIO FASTSPIN® spin kit for soil (MP Biomedicals, Solon, OH) according to manufacturer's instructions with the exception of two extra ethanol washes (**Table S1**).

Activated sludge was sampled from two, 4 L enhanced biological phosphorus removal (EBPR) sequencing batch reactors seeded from Thornside Wastewater Treatment Plant, Queensland, on 16 February, 2011. The first reactor (EBPR1) was operated on 6h reaction cycles of 120 min anaerobic phase, 180 min aerobic phase, 20 min settling, 20 min decant (1 L volume removed from the reactor supernatant) and 14

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

min pre-feed oxygen purge. At the end of each cycle, 1 L of nutrient solution containing 800 mg/L acetate and 40 mg/L phosphate (20:1 COD/P) was added over a period of 6 min (described in detail in (Lu et al. 2006)). The second reactor (EBPR2) was operated under identical conditions with the exception that the anaerobic phase lasted 60 min and nutrient solution was added over a period of 60 min compared to 6 min for EBPR1. Mixed liquor was collected 90 min into the aerobic phase and microbial biomass was concentrated by centrifugation at 4000 rpm for 2 min. Samples were collected at six timepoints from EBPR1 (EBPR1_T1 – EBPR1_T6) and at three timepoints from EBPR2 (EBPR2_T1 – EBPR2_T3) (**Table S1**). DNA was extracted from ca. 500 mg (wet-weight) of biomass with the MP-BIO FASTSPIN® spin kit for soil according to the manufacturer's instructions (MP Biomedicals, Solon, OH).

Methanogenic sludge samples were taken from a full-scale upflow anaerobic sludge blanket (UASB) reactor treating a high-strength organic wastewater discharged from a food-processing factory (Yamada et al. 2011). Two samples of the UASB sludge (A1 and A2) were taken on different dates (A1, 25, December, 2012; A2, 16 September, 2010). The A1 sample was further subsampled into two parts (flocculant sludge [F1] and granular sludge [G1]) by gravimetric settlement. Four samples in total (A1, A2, G1, F1) were used for sequencing. DNA was extracted by a bead-beating method as described previously (Lu et al. 2006)(**Table S1**).

Community profiling of koala faeces and EBPR samples

The V6 to V8 regions of the 16S rRNA gene was amplified using fusion primers containing 454 adaptor sequences ligated to the primers 926F (5'-AAACTYAAAKGAATTGRCGG-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') (Matsuki et al. 2002). Multiplex identifiers consisting of five nucleotides were incorporated in the 1392R primer to allow for multiplexing. Fifty-microlitre PCR reactions were prepared containing 20 ng of template DNA, 5 µL of 10x buffer (Fisher Biotec, Wembley, Australia), 1 µL of 10 mM dNTP mix (Fisher Biotec) 1.5 µL BSAI (Fisher Biotech), 4 µL 25 mM $MgCl_2$ (Fisher Biotec), 1 µL of each 10 µM primer, and 1 unit of *Taq* polymerase (Fisher Biotec). Cycling conditions were 95°C for 3 min, followed by 30 cycles of 95°C for 30 s, 55°C for 30 s and 74°C for 30 s followed by a final extension of 74°C for 10 min. Following amplification, PCR products for each sample were purified using the Agencourt AMPure XP PCR

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

purification system (Beckman-Coultier) and quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). Amplicons were sequenced from the reverse primers using the Roche 454 GS-FLX Titanium platform at the Australian Centre for Ecogenomics, University of Queensland, Australia (ACE, UQ). Sequence data generated were demultiplexed and processed using a modified version of the QIIME pipeline (Caporaso et al. 2010), which uses Acacia 1.50 (app v 2.0.0) (Bragg et al. 2012) to correct homopolymer errors (modified pipeline is available at https://github.com/Ecogenomics/APP). Sequences were clustered at 97% sequence identity and the taxonomy of the representatives from each OTU was assigned using blastn v. 2.2.26 (Johnson et al. 2008) against the Greengenes database, version 12_10) (DeSantis et al. 2006).

Community profiling of UASB samples

For microbial community structure analysis of the UASB samples, the V4 regions of the 16S rRNA gene were amplified using fusion primers for the Illumina sequencing platform (Caporaso et al. 2012). Multiplex identifiers of twelve nucleotides were incorporated in the M806R primer to allow for multiplexing. PCR conditions were described elsewhere (Lu et al. 2006). Following amplification, PCR products for each sample were purified using the Agencourt AMPure XP PCR purification system (Beckman-Coultier) and quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). Amplicons were sequenced from both the forward and reverse primers using the Illumina MiSeq platform and the MiSeq 500 cycles reagent kit v2 (Illumina Inc.). Sequence data generated were demultiplexed and processed using the QIIME pipeline (Caporaso et al. 2010). Sequences were clustered at 97% sequence identity and the taxonomy of the representatives from each OTU was assigned using blastn (Camacho et al. 2009) against the Greengenes database (DeSantis et al. 2006).

Paired end sequencing

The genomic DNA from Zag_T2, EBPR1_T1, EBPR1_T3, EBPR1_T5, EBPR2_T1, EBPR2_T2, EBPR2_T3 and were sent to Aalborg University, Denmark where DNA libraries were prepared for sequencing using TruSeq DNA Sample Preparation Kits v2 (Illumina, San Diego, CA, USA) with 2 µg of DNA following the manufacturer's instructions with nebulizer fragmentation. Library DNA concentration was measured using the QuantIT kit (Molecular probes, Carlsbad, CA, USA) and paired-end

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

sequenced (2 x 150 bp with an average 250 bp fragment size) on an Illumina HiSeq2000 using the TruSeq PE Cluster Kit v3-cBot-HS and (Illumina). The Zag_T2 sample was sequenced on a whole lane of a flowcell and the EBPR1_T1, EBPR1_T3 and EBPR1_T5 were sequenced on a third of a flowcell lane each.

Zag_T1 and Zag_T3, EBPR1_T2, EBPR1_T4, EBPR1_T6, were sequenced at the Institute for Molecular Bioscience, The University of Queensland (IMB, UQ) generating paired-end 150 bp reads (with an average fragment size of 320) using the Nextera DNA Sample Prep kit (Illumina) and the Illumina HiSeq2000 platform. Each Zag sample was sequenced on a quarter of a flowcell lane each and the EBPR samples was sequenced on a third of a flowcell lane each (**Table S1**).

DNA extracts from the four UASB sludge samples (A1, A2, F1, F2) were fragmented to 250-400bp using a Covaris S2 (Covaris, Woburn, MA, USA), and were used for library preparation with a TruSeq sequencing kit (Illumina). Library DNA concentration was measured using the QuantIT kit (Molecular probes) and paired-end sequenced (2 x 250 bp with an approximate average fragment size of 300 bp) on an Illumina MiSeq system using the MiSeq 500 cycles reagent kit v2 (Illumina Inc.). Each of the four DNA libraries was sequenced in a single MiSeq run (**Table S1**).

Raw paired end 2 x 75 bp Illumina data for two male and five female gut metagenome datasets were downloaded from the public MetaHIT database. Details relating to the collection, sequencing, and analysis of the MetaHIT data are provided at http://www.metahit.eu/ (Qin et al. 2010) (**Table S1**).

Sequence assembly and population genome binning

Paired end reads for the koala faeces, MetaHIT, EBPR, and UASB samples were quality trimmed using CLC workbench v6 (CLC Bio, Taipei, Taiwan) with a quality score threshold of 0.01 (phred score 20) and minimum read lengths as follows; 100 bp for the koala faecal and EBPR and samples, 125 bp for the UASB samples, and 50 bp for the MetaHIT samples, in accordance with the read length for each dataset. No ambiguous nucleotides were accepted and Illumina sequencing adapters were trimmed if found.

Trimmed sequences for each biome were assembled using CLC's *de novo* assembly algorithm, with a kmer size of 63 for koala faecal, EBPR, and UASB samples and a kmer size of 45 for the MetaHIT data.

Population genomes were recovered from the paired-end assemblies using GroopM, version 1.0 with default settings (https://github.com/minillinim/GroopM). Briefly, reads from each sample were mapped onto their corresponding co-assemblies (scaffolds ≥ 500 bp; **Table S1**) and coverage patterns for each scaffold were calculated, transformed and projected onto a 3-dimensional plot in which scaffolds from the same population genome would cluster. Integrity of bins was initially confirmed using the GroopM visualization tool.

Population genome completeness and contamination

All contigs in each population genome bin were translated into six open reading frames and a set of 105 single copy marker genes widely conserved in Bacteria (a subset of the 111 from (Dupont et al. 2012)) were identified in the translated dataset using HMMER3 (Finn et al. 2011) with default settings and the model-specific PFAM (Punta et al. 2012) and TIGRFAM (Haft et al. 2003) thresholds was estimated as the percentage of the 105 markers identified in any given population bin, and contamination as the percentage of markers found in >1 copy in a population bin (**Table 1**). The marker gene identification and completeness/contamination calculation functions are combined in the software tool CheckM version 0.5.0 (https://github.com/Ecogenomics/CheckM).

Taxonomic assignment of population genomes

To identify putative representatives of the Melainabacteria amongst the population bins of a minimum quality threshold (>90% completeness, <10% contamination), we constructed a maximum likelihood tree based on a concatenated set of 83 marker genes with known reference genomes (see *Whole genome phylogeny* below) including previously reported standard draft Melainabacteria genomes (Di Rienzi et al. 2013).

Mate pair sequencing for Melainabacteria genome improvement

Genomic DNA extracted from Zag_T1, Zag_T2 and Zag_T3 were multiple strand-displacement amplified in triplicates using the Illustra Genomiphi V2 DNA

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

amplification kit (GE Healthcare) as per manufacturer's instructions. 1 µg of DNA was used for Mate-pair libraries using the Illumina MiSeq sequencing protocol and a gel-free protocol (2-15 kbp inserts).

EBPR1_T1, EBPR1_T6, EBPR2_T1 and EBPR2_T3 were sequenced using long-insert mate-pair sequencing according to Illumina's Nextera protocol. DNA was size selected for 3.5 kbp fragments with a standard deviation of 300bp and further sequenced at IMB, UQ with the Illumina HiSeq platform. Raw mate-pair reads were processed using Prepmate 0.2, removing read pairs where less than 30 bp remained after trimming the adaptor sequence (https://github.com/ctSkennerton/prepmate). Approximately 50% of the raw reads were retained as valid mate-pairs (reads correctly oriented in the reverse-forward direction) resulting in between 16 to 19 million read pairs per sample (**Table S1**).

For the UASB samples, one out of the four samples, A1, was sequenced at the National Institute of Advanced Industrial Science and Technology, Japan (AIST) using the Mate Pair Library Preparation Kit v2 (Illumina) and the MiSeq 500 cycles reagent kit v2 (Illumina) on an Illumina MiSeq system.

Mate-pair sequence data for each sample type (except the public MetaHIT data) was used to scaffold contigs in identified Melainabacteria population genome bins (**Table S1**) using the default settings of SSPACE v2.0 (Boetzer et al. 2011). Scaffolded assemblies were then checked from completeness and contamination using CheckM (**Table 1**).

16S rRNA gene reconstruction

16S rRNA genes are often difficult to recover via differential coverage binning due to coassembly of rRNA genes present in multiple copies, which distorts their coverage statistics. Therefore, 16S rRNA genes were independently reconstructed from the metagenomic data by extracting read pairs that matched an HMM model of the 16S rRNA gene built using HMMER v3.1b1 from 42,363 bacterial and 1,119 archaeal sequences within the 94% dereplicated Greengenes database released on October, 2012 (McDonald et al. 2012). These extracted read pairs were then mapped to the Greengenes database using BWA-MEM v0.7.5a-r405 (Li & Durbin 2010). A read

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

was considered reliably mapped if at least 85% of the read aligned to the reference sequence, and the edit distance of the alignment was at most 10% of the length of the read (e.g., less than 10 for 100 bp reads. Pairs were further filtered to remove any pair where both reads did not properly map to reference sequences within a branch length of 0.03 as measured over the Greengenes phylogeny.

The remaining pairs were clustered in a greedy manner in order to identify pairs mapping to similar Greengenes reference sequences. Reference sequences were put in ascending order according to the number of pairs mapped to them. Starting with the reference sequence with the highest number of assigned pairs, any pairs assigned to a reference sequence within a branch length of 0.03 to this reference sequence were clustered together and removed from further consideration. This process was repeated until all pairs were assigned to a cluster. Each cluster of 16S pairs was then independently assembled using CLC Workbench v6.5. Using this technique an additional 16S rRNA gene was recovered from Zag_221 (1403 bp).

16S rRNA phylogeny

16S rRNA genes from Melainabacteria population genomes were aligned to the standard Greengenes alignment with PyNAST (McDonald et al. 2012). Aligned sequences and a Greengenes reference alignment, version gg_13_5 (ftp://greengenes.microbio.me/ greengenes_release; (McDonald et al. 2012) were imported into ARB (Ludwig et al. 2004) and the Melainabacteria sequence alignments were manually corrected using the ARB EDIT tool. For constructing the alignment data of different taxon configurations, representative taxa (>1,300 nt) were selected and their alignment data were exported from ARB with the Lane mask filtering, resulting in totals of 402 and 67 taxa for two data sets. Neighbour joining trees were calculated from the masked alignments with LogDet distance estimation using PAUP*4.0 (Swofford 2003). A further analysis was run with 100 bootstrap replicates. Maximum parsimony trees were calculated using PAUP*4.0 (Swofford 2003). A heuristic search was used with a random stepwise addition sequence of 10 replicates and nearest-neighbor-interchange swapping. A further analysis was run with 100 bootstrap replicates. Maximum likelihood trees were calculated from the masked alignments using the Generalized Time-Reversible model with Gamma and I options in RAxML version 7.7.8 (Darling et al. 2014) (raxmlHPC-PTHREADS -f a -k

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

-x 12345 -p 12345 -N 100 -T 4 -m GTRGAMMAI). Bootstrap resampling data (100 replicates) were generated with SEQBOOT in the phylip package (Felsenstein 1989), and were used for 100 bootstrap resamplings. Generated trees were re-imported into ARB for visualization (**Fig. 1B**).

Whole genome phylogeny

Two sets of ubiquitous single copy marker genes were obtained and aligned from seven high-quality and four standard draft Melainabacteria population genomes (**Table 1**) and up to 434 complete bacterial and archaeal reference genomes obtained from IMG (v4.0) (Markowitz et al. 2012) using HMMER3 (Finn et al. 2011). The first set of 38 markers (Darling et al. 2014) is found in nearly all Bacteria and Archaea and the second set of 83 markers (**Table S2**) was derived from a 111 bacterial marker set (Dupont et al. 2012) based on congruent taxonomic signal as follows. Individual gene trees were constructed using FastTree v2.1.7 (Price et al. 2010) and compared to the IMG taxonomy (Markowitz et al. 2012). We used the following measure to quantify the agreement of each node in an unrooted gene tree with a specific clade *c* (e.g., Bacteria, Firmicutes) within the IMG taxonomy:

$$\text{consistency} = \max(N_L(c) / (T(c) + I_L(c)) , N_R(c) / (T(c) + I_R(c))$$

where $T(c)$ is the total number of genomes from clade *c*, the subscripts *R* and *L* indicate the subset of genomes to the 'right' and 'left' of the node under consideration, $N_x(c)$ is the number of genomes in subset *x* from clade *c*, and $I_x(c)$ is the number of genomes in subset *x* not from clade *c*. The consistency of a clade *c* was assigned the highest consistency found over all nodes. Average consistencies were then determined over all clades with at least five genomes independently at the domain, phylum, and class ranks. Gene trees where the average consistency over these three ranks was less than 0.86 were discarded as a sharp drop-off in consistency was observed beyond this threshold.

Ambiguous and uninformative alignment positions were removed from aligned sets of concatenated marker genes using gblocks (Castresana 2000) under default settings with the exception that a conserved position was not allowed to have gaps in more than half of the sequences. Phylogenetic trees were reconstructed from the two filterd marker gene alignments with outgroup configurations as detailed in **Table S3 and S4.** All tree topologies were tested for robustness using the maximum likelihood methods

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

from FastTree version 2.1.7 (JTT model, CAT approximation) (Price et al. 2010), RAxML version 7.7.8 with a JTT and Gamma models (Stamatakis 2006) (raxmlHPC-PTHREADS -f a -k -x 12345 -p 12345 -N 100 -T 8 -m PROTGAMMAJTT), and maximum parsimony method using PAUP*4.0 with heuristic search, a random stepwise addition sequence of 10 replicates, and nearest-neighbor-interchange swapping. Generated trees were imported into ARB where they were rooted, beautified and grouped for display purposes.

Melainabacteria genome annotation and metabolic reconstruction

The draft Melainabacteria genomes were submitted to IMG/ER (Markowitz et al. 2009) for automated annotation and manual analysis. KEGG maps and gene annotations were used to reconstruct the metabolism of the Melainabacteria representatives and a composite metabolic cartoon was prepared in Adobe Illustrator CS6 (**Fig. 2 and Table S5**).

Average nucleotide identity was calculated using the ANI calculator with default settings (http://enve-omics.ce.gatech.edu/ani/).

Protein family analysis

The presence of Pfams and TIGRfams for maximally differentiating cell wall types, as previously described in (Albertsen et al. 2013) and flagella assembly (Pallen & Matzke, 2006), were identified in the draft Melainabacteria genomes and 2,363 representative phyla using complete bacterial genomes obtained from IMG (v4.0) (Markowitz et al. 2012) (**Table S6**). Photosynthesis and (bacterio)chlorophyll biosynthesis genes as described in (Sousa et al. 2012) (**Table S6**) were also identified as present or absent by using the BLASTP module (Altschul et al. 1997) in IMG with an e-value of >1e-10 and amino acid identitites of ≥ 25%. Paralogs from the cobalamin pathway or later steps in the bacteriochlorophyll c pathway were removed. The colour key represents the number of species within a certain phylum that have the Pfams, TIGRfams or genes versus the total number of complete bacterial genomes obtained from IMG or the draft Melainabacteria genomes. A heat map was constructed in RStudio v0.95.265 (Racine 2012) using gplots (Warnes et al. 2013) and RColorBrewer (Neuwirth 2011) **(Fig. 3)**.

COG profiles for each genome were constructed through homology search between ORFs predicted with Prodigal v2.60 (Hyatt et al. 2010) and the 2003 COG database (Tatusov et al. 2003). Homology was assessed with BLASTP v2.2.26+ (Camacho et al. 2009) using an e-value threshold of 1e-5 and a percent identity threshold of 30%. The relative percentage of each COG was calculated in relation to the total number of ORFs predicted for each genome. All statistical plots and analyses were conducted using STAMP v2.0.1 (Parks & Beiko 2010).

Results and Discussions

During ongoing culture-independent molecular surveys, 16S rRNA phylotypes belonging to basal cyanobacterial lineages were identified in a number of habitats. These included faecal samples collected from a geriatric male koala (*Phascolarctos cinereus*; Zagget), a lab-scale sequencing batch reactor performing enhanced biological phosphorous removal (EBPR) and an upflow anaerobic sludge blanket (UASB) reactor treating a high-strength organic wastewater discharged from a food processing factory (see sampling details below). In parallel, public metagenomic datasets of human faecal samples containing members of YS2/4C0d-2 (http://www.metahit.eu/ (Qin et al. 2010)) were re-analyzed with the goal of obtaining additional genomes from this lineage.

Three distantly related (88% 16S rRNA gene identity) phylotypes belonging to the YS2/4C0d-2 lineage were detected in the koala faeces, a representative of mle1-12 was identified in the EBPR bioreactor and a representative of SM2F09 was identified in the UASB reactor (**Fig. 1**). Although Melainabacteria are typically found in low abundance, these phylotypes comprised up to 6.7%, 4.2% and 1.7% of the koala faecal, EBPR and UASB microbial communities respectively. Samples with the highest relative abundance of Melainabacteria were chosen for deep metagenomic sequencing to improve the likelihood of obtaining near-complete population genomes for comparative analysis. The relative abundance of the Melainabacteria in the MetaHIT shotgun datasets was estimated to be up to 2.6% by direct classification of 16S rRNA genes in the shotgun datasets.

Recovery of Melainabacteria population genomes

Koala faecal samples were collected from three timepoints from the same koala over a period of six months and sequenced to produce 90.7 Gbp of metagenomic data. Two EBPR reactors were sampled six times and three times respectively over a period of seven months, producing 211.6 Gbp, and the UASB reactor was sampled twice producing 31.5 Gbp of metagenomic data (**Table S1**). Human faecal metagenomes from healthy Danish individuals (two male and five female) were obtained from the public repository (http://www.ebi.ac.uk/ena/home, study *accession number ERP000108*) comprising a total of 21.6 Gbp. Multiple datasets from the same sample types were co-assembled which produced between 3,139 and 148,338 contigs with an N50 of 1.4 kbp to 4.6 kbp. Population genomes were extracted from the assemblies using differential coverage binning of each set of related metagenomes. This approach leverages differences in relative abundance of target populations between two or more related samples to identify contigs with co-varying coverage (Albertsen et al. 2013; Sharon et al. 2013). Population genomes obtained using this method were taxonomically assigned by placement in concatenated gene trees comprising all finished IMG reference genomes (Markowitz et al. 2012) (**Table S2;** see below). Six population genomes were found to form a monophyletic lineage together with the reference cyanobacterial genomes (**Fig. 1A**). These comprised one mle1-12 representative from the EBPR (EBPR_351), one SM2F09 representative from the UASB reactor (UASB_169), and four YS2/4C0d-2 representatives from koala and human faeces (Zag_1, Zag_111, Zag_221 and MH_37) consistent with 16S rRNA gene amplicon community profiling. Analysis of 16S rRNA genes recovered from four of the six population genomes also confirmed that they are members of the Melainabacteria (**Table 1 and Fig. 1B**). Further sequencing using long-insert (2-15 kb) libraries of the EBPR, UASB reactor and koala faeces were used to improve the quality of the draft population genomes (**Table S1**). The completeness and degree of contamination of the improved genomes was estimated by determining the presence and number of copies of 105 conserved single-copy bacterial marker genes (Dupont et al. 2012). All population genomes had >90% estimated completeness (>95 of the 105 conserved single-copy bacterial marker genes) and <10% contamination (multiple copies of genes expected to be present as single copies) (**Table 1**) and can be classified as high-quality draft genomes with the exception of UASB_169 which is a standard draft (Chain et al. 2009).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

An expanded phylogenetic classification of the phylum Cyanobacteria

We began analysis of the Melainabacteria genomes by constructing phylogenetic trees based on two concatenated alignments of broadly conserved single copy marker genes (see Methods; **Table S2** and (Darling et al. 2014)) The ingroup comprised 81 reference cyanobacterial genomes and 11 Melainabacteria genomes; six determined in the present study and the five most complete genomes obtained in the Di Rienzi *et. al.* (2013) study (**Fig. 1A and Table S3**). We evaluated the monophyly of the photosynthetic Cyanobacteria and Melainabacteria groups using up to 377 outgroup genomes representing 28 phyla (**Figs. S2 to S4 and Table S4**). In all cases, the evolutionary association between the two groups was reproducibly resolved with >80% bootstrap support (node 1, **Fig. 1A**). Di Rienzi *et. al.* (2013) concluded that the two groups are sister phyla rather than a single phylum based on a combined divergence slightly greater than a recommended threshold of 85% 16S rRNA gene sequence identity for distinguishing new phyla (Hugenholtz et al. 1998). However, the primary criterion for defining a new phylum is not satisfied in this instance, i.e. that the lineage is reproducibly unaffiliated with existing phyla (Hugenholtz et al. 1998) according to both 16S rRNA and genome-level phylogenies. Moreover, several inferred features beyond the evolutionarily conserved core set of genes used to construct the genome tree are consistent with a common ancestry (see below). Therefore, we propose that the phylum Melainabacteria should be reclassified as the class Melainabacteria within the phylum Cyanobacteria. Me.lai.na.bac.te.ria. Gr. n. Melaina, a nymph in Greek mythology, who presides over dark subterranean aspects; N.L. masc. n. *bacter* (from Gr. n. *baktron*), a rod; suff -*ia* ending to denote a class; N.L. fem. pl. n. *Melainabacteria* class of bacteria found in the dark

The node defining the Melainabacteria in the concatenated gene alignment tree (node 2, **Fig. 1A**) was supported in all analyses with >80% confidence consistent with the genome-based analysis of Di Rienzi *et. al.* (2013). The population genomes formed three primary lines of descent within the Melainabacteria, with the human and koala gut genomes and groundwater genome ACD20 (Di Rienzi et al. 2013) forming a monophyletic cluster. 16S rRNA-based inference provides only modest support for the monophyly of the Melainabacteria (**Figs. 1B and Fig. S5**) and indeed several of the robustly monophyletic groups therein are classified as primary cyanobacterial lines of descent (classes) in Greengenes and Silva (McDonald et al. 2012; Quast et al.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

2013). The higher resolution afforded by the genome sequences suggests that these lineages should be classified as orders within the class Melainabacteria (**Fig. 1**). We propose names for four of these orders based on habitat and analysis of the population genomes (see below), and the recognition of *Vampirovibrio chlorellavorus* in the SM1D11 lineage (**Fig. 1**). The location of the ACD20 genome could not be determined within the 16S rRNA gene tree (**Fig. 1B**) as it lacks a 16S rRNA gene sequence, but genome trees based on a refined binning of this population (Albertsen et al. 2013) (**Table 1**) indicate that it is basal and monophyletic with the order, Gastranaerophilales (YS2/4C0d2; **Fig. 1A and Figs. S2 and S3**). An additional distantly related cyanobacterial lineage ML635J-21 (McDonald et al. 2012; Quast et al. 2013) currently not represented by a sequenced genome may represent another class-level lineage within the Cyanobacteria (**Fig. S4**) highlighting the need for further genomic exploration of the cyanobacterial phylum.

Our analyses show that the photosynthetic Cyanobacteria are robustly monophyletic within the expanded context of the phylum Cyanobacteria (node 3, **Fig. 1A**). We therefore propose to reinstigate the name Oxyphotobacteria (Gibbons & Murray 1978) to describe all photosynthetic Cyanobacteria (including chloroplasts) in a single class. Gr. adj. *oxus*, acid or sour and in combined words indicating oxygen; Gr. n. *phos photos*, light; Gr. n. *baktêria*, staff, cane; suff. *-ia*, ending to denote a class; N.L. neut. pl. n. *Oxyphotobacteria*, light-requiring bacteria that produce oxygen. The name implies that the class is able to photosynthesize. We denote the order-level groupings within this class as A to G (**Fig. 1A**) in accordance with a recent genome-based analysis (Shih et al. 2013). Oxyphotobacteria are still classified primarily on morphological grounds into five subsections (Rippka et al. 1979) despite clear incongruencies between phylogenetic reconstructions and morphological complexity (Shih et al. 2013). Therefore, it is inevitable that the order- and family-level groupings within the Oxyphotobacteria will be reclassified on phylogenetic grounds with a concomitant widespread reclassification of cyanobacterial strains once this group is no longer under the jurisdiction of the Botanical Code (Komárek 2010).

Inferred metabolism of Melainabacteria genomes

Di Rienzi *et, al*. inferred metabolic properties of the class Melainabacteria based on comparative analysis of draft population genomes belonging to only one order, the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Gastranaerophilales (**Fig. 1**). We substantially increase the phylogenetic coverage of the Melainabacteria in the present study by recovery of high-quality draft population genomes spanning three of the six identified orders (**Fig. 1B**). Expanded genomic representation should provide a more balanced overview of the metabolic properties of this class including features in common with, or distinct from, the Oxyphotobacteria. We begin by proposing Candidatus species for the most complete genomes in each of the three orders obtained in this study and describe their inferred metabolic properties below.

The most complete Gastranaerophilales genome with the least number of scaffolds, Zag_221, was selected as the Candidatus species representative of the group, *Candidatus* Gastranaerophilus phascolarctosicola (**Table 1 and Fig. 1**). *Gastranaerophilus phascolarctosicola* (Gas.tra.nae.ro.phi'lus. Gr. n. *gaster* stomach; Gr. pref. *an*-, not; Gr. masc. n. *aer*, air; L. masc. adj. *philus* [from Gr. adj.*philos*], friend, loving; *Gastranaerophilus* a bacterium loving anaerobic gastric environments. 'phas.co.larc.to.si.co.la'. N.L. *Phascolarctos* the name of koala; L. suffix *-cola* inhabitant, dweller; N.L. masc. n. *phascolarctosicola* hiding in the belly of a koala).

The genome size and GC content range of the four Gastranaerophilales genomes were in accord with the Di Rienzi *et. al.* population genomes from this order (**Table 1**). Members of this group have small streamlined genomes ranging in size from 1.8 to 2.3 Mb, with the exception of ACD20, which is 2.7 Mb after binning refinement (Albertsen et al. 2013). The Gastranaerophilales genomes recovered from the koala and human faeces in the present study support the assertion (Di Rienzi et al. 2013) that this lineage comprises obligate fermenters missing the genes necessary for aerobic and anaerobic respiration, as well as the tricarboxylic acid (TCA) cycle (**Fig. 2 and Table S5**). Instead, all Gastranaerophilales genomes contain the Embden-Meyerhof-Parnas (EMP) pathway, capable of converting glucose, mannose, starch or glycogen into lactate, ethanol and/or formate (**Fig. 2**). All representative genomes have the potential to produce riboflavin, nicotinamide, biotin, dehydrofolate and pantoate as found previously (Di Rienzi et al. 2013).

Di Rienzi *et. al.* highlighted the presence of FeFe hydrogenases in their human gut population genomes speculating that these organisms are hydrogen-producing

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

anaerobes in syntrophic interactions with hydrogenotrophic methanogens or acetogens. We also identified FeFe hydrogenases in the MH_37 genome obtained from the human gut, but in contrast found Fe-only or NiFe hydrogenases in koala gut Gastranaerophilales genomes (**Fig. S6**). It is possible that the less oxygen sensitive NiFe hydrogenase would allow members of this order to colonize the jejunum as well as the more anaerobic colon (Quigley & Quera 2006).

Di Rienzi *et. al.* also reported that the Melainabacteria are flagellated based on the presence of a complete set of flagella genes in three of their draft Gastranaerophilales genomes (MEL_B1, MEL_B2 and ACD20). Of the nine Gastranaerophilales genomes available (**Table 1**), only these three had complete flagella gene sets, the remainder having only a subset of genes that would not encode a functional flagellum (**Fig. S7**). We infer that flagella were present in the ancestor of the class Melainabacteria and subsequently lost on at least two occasions based on monophyly of flagella genes common to the Gastranaerophilales and Caenarcaniphilales (**Fig. 1B and Fig. S7)**. Moreover, there appears to have been a subsequent loss of functional flagella in *'G. phascolarctocola'* and relatives (**Fig. S7**) indicating the presence of non-motile members of this order in animal gut habitats.

Based on a suggested species threshold of 95% average nucleotide identity (Goris et al. 2007), three of the Gastranaerophilus genomes can be considered to belong to the same species (**Table 1 and Fig. S8**). Two of these genomes were recovered from humans (MH37 and MEL_C1) and the third from a koala (Zag_1) despite the structural and physiological differences between the human and koala gut. The majority of genes that differ between these genomes are hypothetical proteins or phage associated (**Fig. S8**) typical of differences seen between strains belonging to the same species.

Candidatus Obscuribacter phosphatis (Ob.scur.i.bac.ter. L. adj. *obscurus* dark; N.L. masc. n. *bacter* (from Gr. n. *baktron*), a rod; *Obscuribacter* a bacterium found in the dark. 'phos.pha.tis'. N.L. n. *phosphatis*, phosphate; N.L. *phosphatis* accumulating phosphate) EBPR_351 (*'O. phosphatis'*) representing the order Obscuribacterales (**Fig. 1 and Table 1**) is conspicuous among the Melainabacteria genomes because of its larger size (5 Mb) and associated metabolic versatility. '*O. phosphatis'* contains

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

the genes necessary for polyphosphate metabolism, including a low affinity inorganic phosphate transporter (PiT), polyphosphate kinase 1 (used to synthesize or degrade polyP while consuming or generating respectively ATP directly), polyphosphate kinase 2 (degrades polyP producing GTP from GDP), exopolyphosphatase (degrades polyP in a non-reversible reaction that does not generate ATP directly), polyphosphate:AMP phosphotransferase and adenylate kinase (Seviour & Nielsen 2010). *'O. phosphatis'* has the capacity for aerobic and anaerobic respiration, and fermentation, allowing it to function during both the oxic and anoxic phases of EBPR (Blackall et al. 2002). It contains genes encoding a complete respiratory chain including Complexes I, II and IV and an F-Type ATPase (**Fig. 2**). Di Rienzi *et. al.* concluded that the Melainabacteria lack electron transport chains and are therefore incapable of respiration. This highlights the dangers of inferring phylum- or class-level functionality based on limited phylogenetic sampling of the lineage.

Like the Gastranaerophilales, *'O. phosphatis'* has the capability to metabolize a wide range of simple carbohydrates via the EMP pathway, and also fatty acids via the beta-oxidation pathway (**Fig. 2**). Under oxic conditions, we predict that *'O. phosphatis'* will fully oxidize one or more of these substrates via the TCA cycle, feeding NADH into the electron transport chain with a cbb3-type cytochrome as the terminal oxidase. This family of cytochromes is typically used in microaerophilic conditions (Kulajta et al. 2006) suggesting that *'O. phosphatis'* may be found within flocs where oxygen concentrations are lower (Albertsen et al. 2013). Under anoxic conditions, we predict that it performs either respiration with nitrate as the terminal electron acceptor or, in the absence of nitrate, mixed-acid fermentation with the potential to produce ethanol, lactate, formate, succinate, CO_2 and H_2 (**Fig. 2**). The presence of these metabolic pathways suggests that '*O. phosphatis*' has adapted to more dynamic environments (requiring greater metabolic plasticity) with 'feast-famine' nutrient cycles such as those artificially imposed on EBPR bioreactors.

Candidatus Caenarcanum bioreactoricola ('Caen.arc.an.um' L. neut. n. *caenum* mud, sludge; L. neut. n. *arcanum* secret, hidden; N.L. neut. n. *Caenarcanum* a bacterium hidden in sludge. 'bio.re.ac.to.ri.co.la'. L. suffix *-cola* inhabitant, dweller; N.L. masc. n. *bioreactericola* living in a bioreactor) UASB_169 (*'C. bioreactoricola'*) representing the order Caenarcaniphilales, has an estimated genome size of ~2 Mb

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

and a remarkably low GC content of 27.7%, the lowest GC content reported for Cyanobacteria. Similar to the Gastranaerophilales genomes, '*C. bioreactoricola*' lacks the genes necessary for aerobic and anaerobic respiration, as well as the TCA cycle, suggesting that *'C. bioreactoricola'* has a streamlined metabolism only producing energy via fermentation with ethanol and lactate as the main fermentation products. *'C. bioreactoricola'* contains the subunits for Fe-only hydrogenase and the potential to produce hydrogen as a by-product from the fermentation process. Like the Gastranaerophilales, *'C. bioreactoricola'* may also be a hydrogen producer living in syntrophy with methanogens or acetogens in the bioreactor, as microcolonies of syntrophic bacteria are often observed in the granules from UASB systems and electron transfer in these microcolonies is thought to mostly occur through interspecies hydrogen transfer (Schmidt & Ahring 1996).

Emergence of photosynthesis in the Cyanobacteria

Melainabacteria resemble Oxyphotobacteria in their cell envelope gene complement comprising genes indicative of a Gram-negative (diderm) cell wall (**Fig. 3**). This includes genes for the biosynthesis of Lipid A for the production of lipopolysaccharide (LPS) as previously reported for members of the Gastranaerophilales (Di Rienzi et al. 2013). Oxyphotobacteria also have unusual cell envelope components for Gram-negative bacteria including porins (*somA, somB*), which are thought to help anchor the outer membrane to the peptidoglycan layer (Hansel et al. 1998; Hoiczyk & Hansel 2000). All Melainabacteria have closely related homologs to the oxyphotobacterial *somA* and *somB* genes suggesting that their cell envelopes comprise similar porins.

Di Rienzi *et. al*. highlighted the presence of putative circadian rhythm (*rpaA* and *rpaB*) and light response (*nblS*) regulators in the Gastranaerophilales, which are diagnostic of Cyanobacteria. We identified orthologs of these genes in all three orders of the Melainabacteria suggesting that these are uniquely ancestral features of the phylum. Together with the unambiguous phylogenetic placement of the Melainabacteria within the cyanobacterial radiation based on comparative analysis of highly conserved marker genes (**Fig. 1**), the conservation of features characteristic of Oxyphotobacteria in the Melainabacteria further support a common ancestry and the proposal for a single phylum.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

The most conspicuous difference between the Melainabacteria and Oxyphotobacteria is the absence of chlorophyll biosynthesis genes in the former (**Fig. 3**) consistent with previous findings (Di Rienzi et al. 2013). All subunits for photosystems I and II, and the electron transport chain were absent from the Melainabacteria genomes sequenced in this study. Only the ATPase subunits (all genomes) and cytochrome b6/f subunits (*petB* and *petC* in *'O. phosphatis'*) are present, both of which play roles independent of photosynthesis (Mulkidjanian et al. 2007). Another trait characteristic of photosynthetic cyanobacteria, carbon fixation, is similarly absent in the Melainabacteria (**Fig. 2**) indicating that these organisms do not engage in a photoautotrophic lifestyle. Instead, members of this lineage appear to be chemoheterotrophs with diverse functionality.

The idea of non-photosynthetic Cyanobacteria is contrary to the prevailing dogma that all members of this phylum are photosynthetic (Shih et al. 2013). However, this should not be a controversial conclusion given that photosynthesis is only found in sub-lineages of several other phyla such as the Proteobacteria, Firmicutes, Acidobacteria and Chloroflexi (Bryant & Frigaard 2006). The complete absence of photosynthetic apparatus in the Melainabacteria suggests that the Oxyphotobacteria acquired photosystems after diverging from the common ancestor of the Melainabacteria (**Fig. 1**). This is consistent with the inferences that photosynthesis genes have an extensive history of lateral transfer (Hohmann-Marriott & Blankenship 2011) and that photosynthesis developed late in the Cyanobacteria (Xiong et al. 2000).

The acquisition of oxygenic photosynthesis in the Oxyphotobacteria had profound impact not only on the biosphere (Hohmann-Marriott & Blankenship 2011) but left imprints in their genomes that are now apparent by contrasting with their newly sequenced non-photosynthetic relatives. For example, there was a great expansion of ATP-driven transport systems in the Oxyphotobacteria likely for acquiring bicarbonate (COG0600/0715/1116) and iron (COG0609/1629/0735) necessary for photosynthesis and respiration (**Fig. S9**). The additional energy available to Oxyphotobacteria via oxygenic photosynthesis may also explain the widespread

acquisition of energy-intensive biosynthetic pathways such as secondary metabolite synthesis (Shih et al. 2013) (**Fig. S10**).

675

Conclusion

Our findings refute the long-standing dogma that all Cyanobacteria are photosynthetic. We infer that the cyanobacterial ancestor was a non-photosynthetic chemoheterotroph and that photosystems were acquired after divergence of the classes
680 Melainabacteria and Oxyphotobacteria. We suggest that the acquisition of oxygenic photosynthesis resulted in an increase in genome complexity within the Oxyphotobacteria (followed by a subsequent reduction and streamlining in the *Prochlorococcus* lineage; (Partensky & Garczarek 2010)) while the Melainabacteria mostly retained a simpler ancestral metabolism. Consistent with the phylogenetic
685 depth of the Melainabacteria, members of this class occupy a wide range of environmental niches with varied metabolic properties, mostly centred around fermentative lifestyles, that nonetheless extend the known metabolic diversity of the Cyanobacteria. These include respiratory nitrate reduction (*'O. phosphatis'*) and flagella-based motility (in some Gastranaerophilales (Di Rienzi et al. 2013)). If the
690 inclusion of *Vampirovibrio chlorellavorus* (Coder & Goff 1986) in the Melainabacteria is confirmed by genomic sequencing, then parasitism can also be added to the known phenotypes of Cyanobacteria. The availability of eleven mostly high quality draft genomes representing multiple orders within the Melainabacteria (**Table 1**) provides a sound basis for further investigations into this fascinating group,
695 for example, via spatial visualization (Moter & Göbel 2000) and genome-directed isolation (Pope et al. 2011; Tyson et al. 2005).

Acknowledgements

We thank Norman Pace, Tal Dagan and Michael Galperin for providing valuable
700 perspective on the study and Karen Nilsson and Jacqui Brumm at Lone Pine Koala Sanctuary for facilitating collection of koala faecal samples, Adam Skarshewski for implementing the concatenation of single copy marker genes, Serene Low and Margaret Butler for preparing samples for Illumina sequencing, Fiona May for 454 pyrotags, Queensland Centre for Medical Genomics, UQ and Aalborg University for
705 Illumina paired end and mate pair sequencing. We also thank Akiko Ohashi and Satoko Matsukura at AIST for Illumina shotgun and 16S amplicon sequencing. We

thank Tim Lilburn for information on the ATCC strain of *Vampirovibrio chlorellavorus*. We thank Satoshi Hanada and Aharon Oren for etymological advice and clarification of nomenclature versus classification issues (we are classifying here). This work is supported by the Australian Research Council (ARC) through project DP120103498, strategic funds from the Australian Centre for Ecogenomics; G.W.T. is supported by an ARC Queen Elizabeth II fellowship [DP1093175]; R.M.S and C.T.S is supported by an ARC Postgraduate Award (APA).

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, and Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotech* 31:533-538.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

Blackall LL, Crocetti GR, Saunders AM, and Bond PL. 2002. A review and update of the microbiology of enhanced biological phosphorus removal in wastewater treatment plants. *Antonie Van Leeuwenhoek* 81:681-691.

Boetzer M, Henkel CV, Jansen HJ, Butler D, and Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578-579.

Bragg L, Stone G, Imelfort M, Hugenholtz P, and Tyson GW. 2012. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat Meth* 9:425-426.

Bryant DA, and Frigaard NU. 2006. Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* 14:488-496.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden T. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. . 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 7:335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M et al. . 2012. Ultra-high-throughput microbial

community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6:1621-1624.

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

Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C et al. . 2009. Genome Project Standards in a New Era of Sequencing. *Science* 326:236-237.

Coder DM, and Goff LJ. 1986. The host range of the Chlorellavorus bacterium ("Vampirovibrio chlorellavorus"). *Journal of Phycology* 22:543-546.

Cruz-Martinez K, Suttle KB, Brodie EL, Power ME, Andersen GL, and Banfield JF. 2009. Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J* 3:738-744.

Darling AE, Jospin G, Lowe E, Matsen FAIV, Bik HM, and Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, and Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069-5072.

Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, Goodrich JK, Bell JT, Spector TD, Banfield JF et al. . 2013. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife* 2.

Dojka MA, Harris JK, and Pace NR. 2000. Expanding the Known Diversity and Environmental Distribution of an Uncultured Phylogenetic Division of Bacteria. *Applied and Environmental Microbiology* 66:1617-1621.

Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Richter RA, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH et al. . 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6:1186-1199.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164-166.

Finn RD, Clements J, and Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29-37.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Gibbons NE, and Murray RGE. 1978. Validation of Cyanobacteriales Stanier in Gibbons and Murray 1978 as a New Order of the Kingdom Procaryotae Murray 1968, and of the Use of Neuter Plural Endings for Photobacteria and Scotobacteria classes nov. Gibbons and Murray 1978: Request for an Opinion. *International Journal of Systematic Bacteriology* 28:332-333.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, and Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57:81-91.

Gromov B, and Mamkaeva K. 1980. Proposal of a new genus Vampirovibrio for chlorellavorus bacteria previously assigned to Bdellovibrio. *Mikrobiologia* 49:165-167.

Haft DH, Selengut JD, and White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res* 31:371-373.

Hansel A, Pattus F, Jurgens UJ, and Tadros MH. 1998. Cloning and characterization of the genes coding for two porins in the unicellular cyanobacterium Synechococcus PCC 6301. *Biochim Biophys Acta* 1399:31-39.

Hohmann-Marriott MF, and Blankenship RE. 2011. Evolution of photosynthesis. *Annu Rev Plant Biol* 62:515-548.

Hoiczyk E, and Hansel A. 2000. Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope. *J Bacteriol* 182:1191-1199.

Hugenholtz P, Goebel BM, and Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180:4765-4774.

Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, and Hauser L. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, and Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5-9.

Komárek J. 2010. Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular background with phenotype and ecological consequences (genus and species concept). *Hydrobiologia* 639:245-259.

Kulajta C, Thumfart JO, Haid S, Daldal F, and Koch HG. 2006. Multi-step assembly pathway of the cbb3-type cytochrome c oxidase complex. *J Mol Biol* 355:989-1004.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Ley RE, Turnbaugh PJ, Klein S, and Gordon JI. 2006. Microbial ecology: Human gut microbes associated with obesity. *Nature* 444:1022-1023.

Li H, and Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589-595.

Liu B, Zhang F, Feng X, Liu Y, Yan X, Zhang X, Wang L, and Zhao L. 2006. Thauera and Azoarcus as functionally important genera in a denitrifying quinoline-removal bioreactor as revealed by microbial community structure comparison. *FEMS Microbiol Ecol* 55:274-286.

Lu H, Oehmen A, Virdis B, Keller J, and Yuan Z. 2006. Obtaining highly enriched cultures of Candidatus Accumulibacter phosphates through alternating carbon sources. *Water Res* 40:3838-3848.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G et al. . 2004. ARB: a software environment for sequence data. *Nucleic Acids Res* 32:1363-1371.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P et al. . 2012. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* 40:D115-D122.

Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, and Kyrpides NC. 2009. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25:2271-2278.

Matsuki T, Watanabe K, Fujimoto J, Miyamoto Y, Takada T, Matsumoto K, Oyaizu H, and Tanaka R. 2002. Development of 16S rRNA-Gene-Targeted Group-Specific Primers for the Detection and Identification of Predominant Bacteria in Human Feces. *Applied and Environmental Microbiology* 68:5445-5451.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, and Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6:610-618.

Moter A, and Göbel UB. 2000. Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. *Journal of Microbiological Methods* 41:85-112.

Mulkidjanian AY, Makarova KS, Galperin MY, and Koonin EV. 2007. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol* 5:892-899.

Nelson N, and Ben-Shem A. 2004. The complex architecture of oxygenic photosynthesis. *Nat Rev Mol Cell Biol* 5:971-982.

Neuwirth E. 2011. RColorBrewer: ColorBrewer palettes.

Pallen MJ, and Matzke NJ. 2006. From The Origin of Species to the origin of bacterial flagella. *Nat Rev Micro* 4:784-790.

Parks DH, and Beiko RG. 2010. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26:715-721.

Partensky F, and Garczarek L. 2010. Prochlorococcus: advantages and limits of minimalism. *Ann Rev Mar Sci* 2:305-331.

Pope PB, Smith W, Denman SE, Tringe SG, Barry K, Hugenholtz P, McSweeney CS, McHardy AC, and Morrison M. 2011. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science* 333:646-648.

Price MN, Dehal PS, and Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5:e9490.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al. . 2012. The Pfam protein families database. *Nucleic Acids Res* 40:D290-D301.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al. . 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-65.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, and Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590-596.

Quigley EM, and Quera R. 2006. Small intestinal bacterial overgrowth: roles of antibiotics, prebiotics, and probiotics. *Gastroenterology* 130:S78-90.

Racine JS. 2012. RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics* 27:167-172.

Rippka R, Deruelles J, Waterbury JB, Herdman M, and Stanier RY. 1979. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *Journal of General Microbiology* 111:1-61.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Schmidt JE, and Ahring BK. 1996. Granular sludge formation in upflow anaerobic sludge blanket (UASB) reactors. *Biotechnol Bioeng* 49:229-246.

Seviour RJ, and Nielsen PH. 2010. *Microbial Ecology of Activated Sludge*: IWA Publishing.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, and Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23:111-120.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R et al. . 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A* 110:1053-1058.

Sousa FL, Shavit-Greivink L, Allen JF, and Martin WF. 2012. Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. *Genome Biology and Evolution*.

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

Swofford D. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN et al. . 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, and Banfield JF. 2005. Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. *Appl Environ Microbiol* 71:6319-6324.

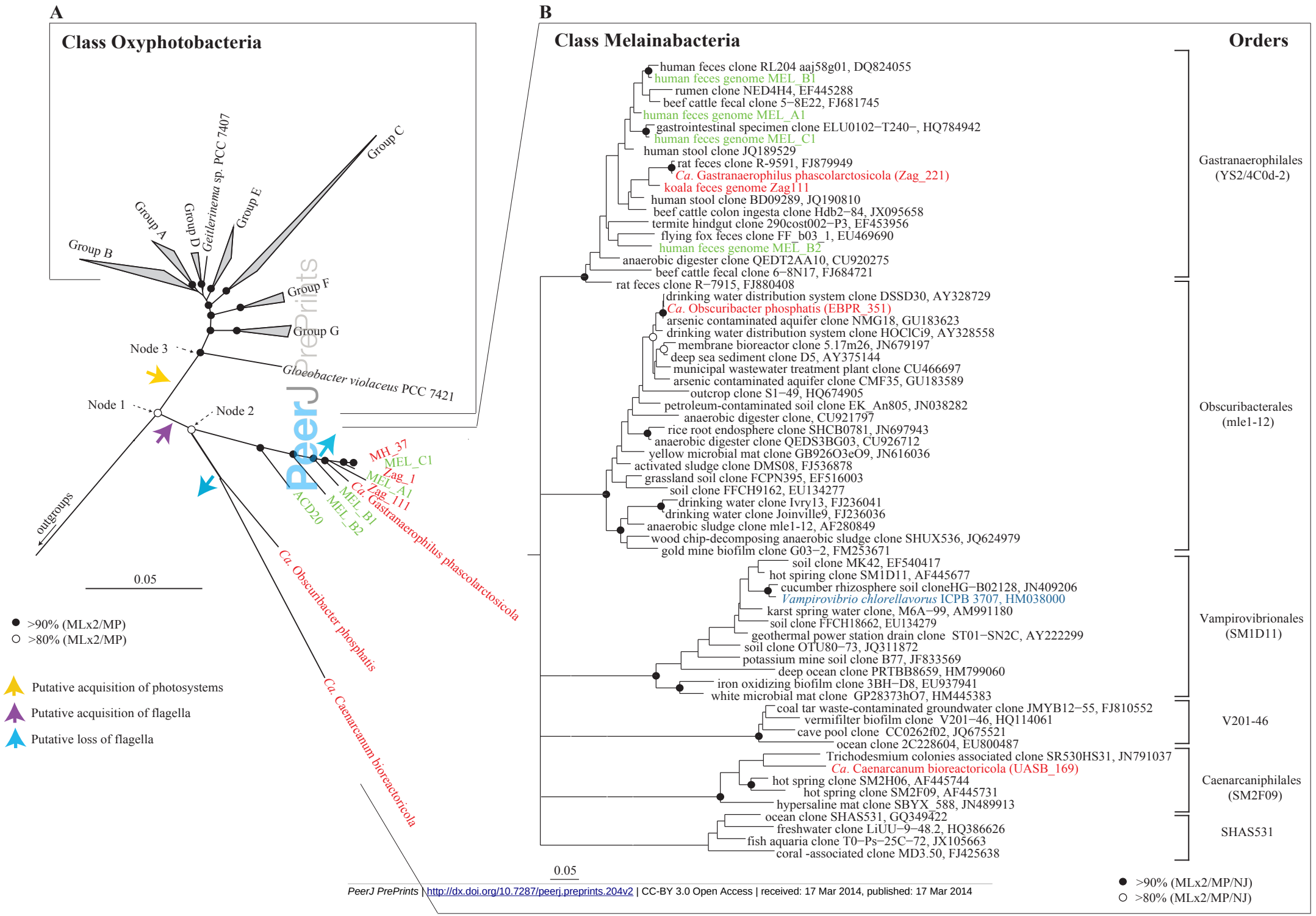
Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M et al. . 2013. gplots: Various R programming tools for plotting data.

Williams MM, Domingo JW, Meckes MC, Kelty CA, and Rochon HS. 2004. Phylogenetic diversity of drinking water bacteria in a distribution system simulator. *J Appl Microbiol* 96:954-964.

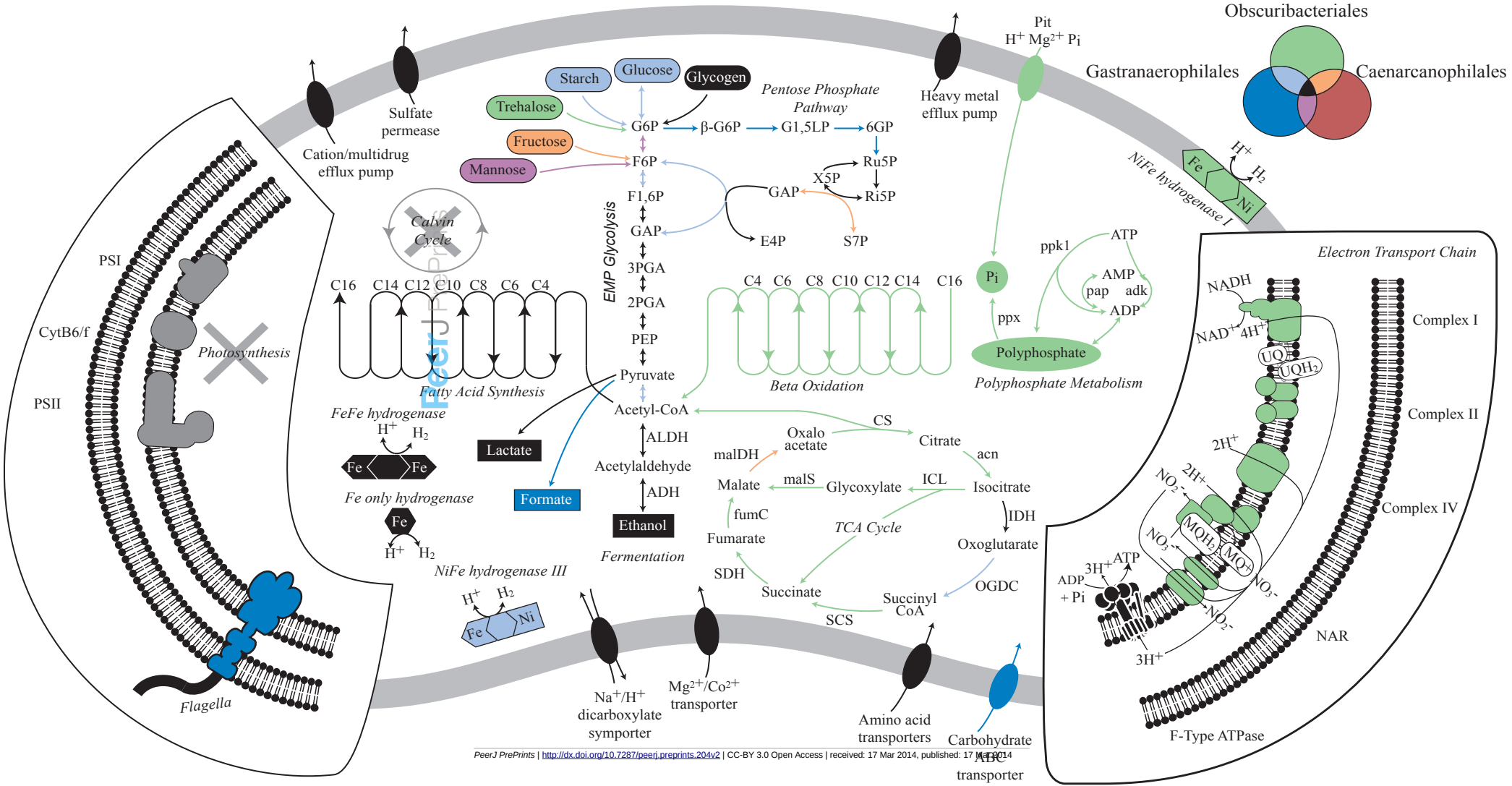
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Xiong J, Fischer WM, Inoue K, Nakahara M, and Bauer CE. 2000. Molecular Evidence for the Early Evolution of Photosynthesis. *Science* 289:1724-1730.

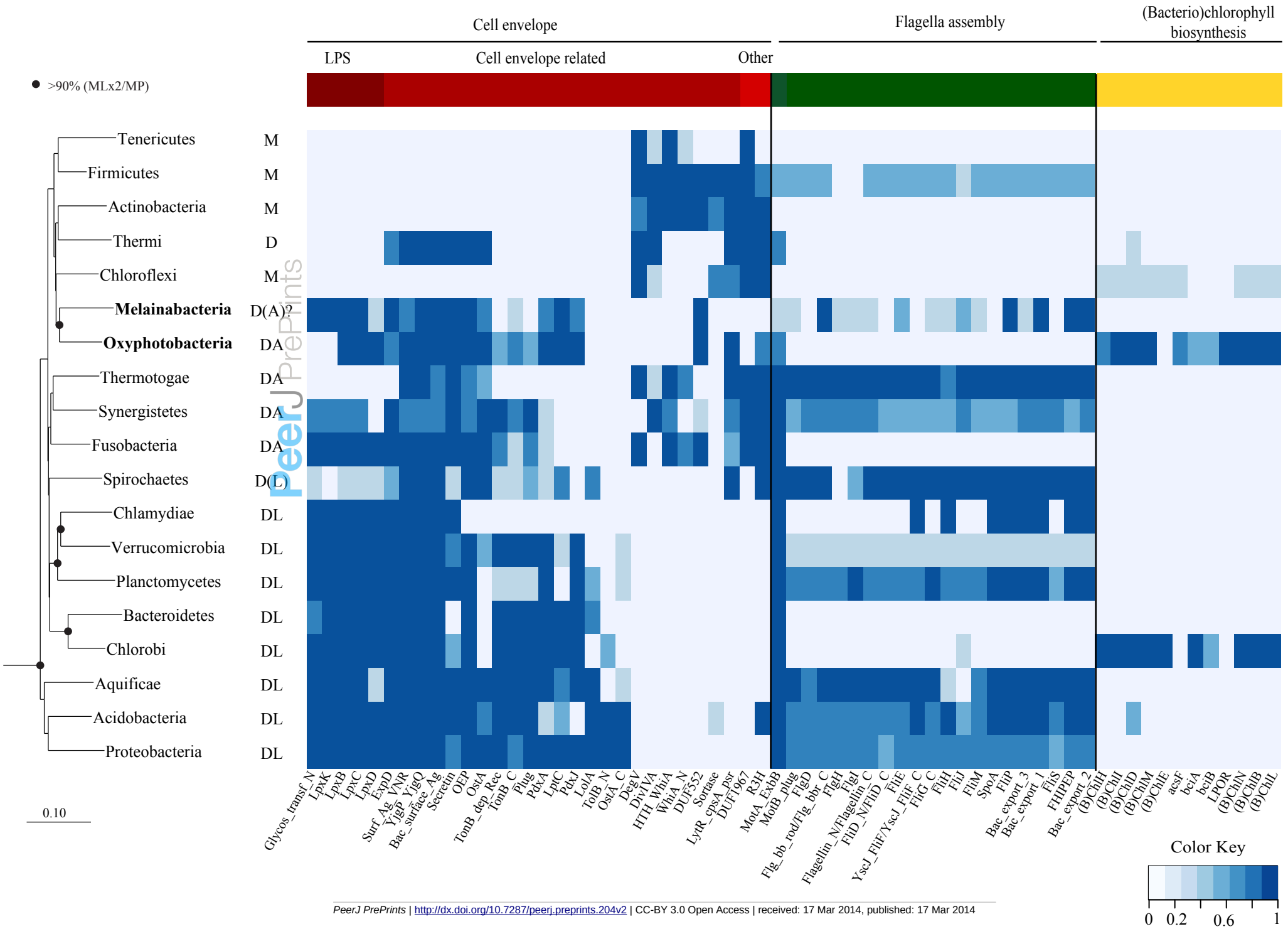
Yamada T, Kikuchi K, Yamauchi T, Shiraishi K, Ito T, Okabe S, Hiraishi A, Ohashi A, Harada H, Kamagata Y et al. . 2011. Ecophysiology of uncultured filamentous anaerobes belonging to the phylum KSB3 that cause bulking in methanogenic granular sludge. *Appl Environ Microbiol* 77:2081-2087.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.204v2 | CC-BY 3.0 Open Access | received: 17 Mar 2014, published: 17 Mar 2014

Population genome	# of scaffolds	Estimated genome size (Mbp)	%GC	# of genes[1]	rRNAs[2]	Estimated completeness (%)[3]	Estimated contamination (%)[3]	Proposed candidatus name	Study
Zag_221	14	1.8	38.5	1838 (1799)	16S	100.0	1.0	Gastranaerophilus phascolarctosicola	Present study
Zag_1	322	2.0	34.9	2194 (2160)	-	94.3	1.9		Present study
Zag_111	65	2.2	36.7	2313 (2257)	5S, 16S 23S	98.1	5.7		Present study
MH_37	157	2.2	34.1	2402 (2360)	-	100.0	1.0		Present study
MEL_A1	1	1.9	33.0	1879 (1832)	5S, 16S, 23S	100.0	2.9		Di Rienzi et al., 2013
MEL_B1	21	2.3	35.4	2269 (2219)	5S, 16S, 23S	100.0	1.0		Di Rienzi et al., 2013
MEL_B2	26	2.3	36.3	2262 (2215)	5S, 16S, 23S	100.0	1.9		Di Rienzi et al., 2013
MEL_C1	4	2.1	34.1	2162 (2120)	5S, 16S, 23S	100.0	1.9		Di Rienzi et al., 2013
ACD20[4]	185	2.7	33.5	2455 (2325)	5S, 23S	100.0	2.9		Di Rienzi et al., 2013
EBPR_351	8	5.5	49.4	4392 (4342)	5S, 16S, 23S	99.1	7.6	Obscuribacter phosphatis	Present study
UASB_169	67	1.8	27.5	1917 (1870)	16S, 23S	94.3	0.0	Caenarcanum bioreactoricola	Present study

[1]Numbers in brackets for # of genes is the number of protein coding genes.
[2]16S rRNA lengths are >1000bp.
[3]Estimated completeness and estimated contamination is based on 105 single copy marker genes (a subset of the 111 single copy marker set from Dupont et al., 2012).
[4]ACD20 is the corrected genome from Albertsen et al., 2013 as the original completeness for ACD20 was 100.0% and original contamination was 107.6%.