Metabolic differentiation of co-occurring Accumulibacter clades revealed through genome-resolved metatranscriptomics

Elizabeth A. McDaniel1*, Francisco Moya-Flores2,3,4*, Natalie Keene Beach2, Pamela Y. Camejo2,5, Benjamin O. Oyserman6,7, Matthew Kizaric2, Eng Hoe Khor2, Daniel R. Noguera2,8, and Katherine D. McMahon1,2†

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA
2 Department of Civil and Environmental Engineering, University of Wisconsin – Madison, Madison, WI, USA
3 Genomics & Resistant Microbes (GeRM), Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Chile
4 Millenium Initiative for Collaborative Research on Bacterial Resistance (MICROB-R), Iniciativa Científica Milenio, Chile
5 Millennium Institute for Integrative Biology (iBio), Santiago, Chile
6 Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands
7 Microbial Ecology, Netherlands Institute of Ecological Research, Wageningen, The Netherlands
8 DOE Great Lakes Bioenergy Research Center, Madison, WI, USA

* equal contribution
† Corresponding Author: kdmcmahon@wisc.edu
ABSTRACT

Microbial communities in their natural habitats consist of closely related populations that may exhibit phenotypic differences and inhabit distinct niches. However, connecting genetic diversity to ecological properties remains a challenge in microbial ecology due to the lack of pure cultures across the microbial tree of life. ‘Candidatus Accumulibacter phosphatis’ is a polyphosphate-accumulating organism that contributes to the Enhanced Biological Phosphorus Removal (EBPR) biotechnological process for removing excess phosphorus from wastewater and preventing eutrophication from downstream receiving waters. Distinct Accumulibacter clades often co-exist in full-scale wastewater treatment plants and lab-scale enrichment bioreactors, and have been hypothesized to inhabit distinct ecological niches. However, since individual strains of the Accumulibacter lineage have not been isolated in pure culture to date, these predictions have been made solely on genome-based comparisons and enrichments with varying strain composition. Here, we used genome-resolved metagenomics and metatranscriptomics to explore the activity of co-existing Accumulibacter clades in an engineered bioreactor environment. We obtained four high-quality genomes of Accumulibacter strains that were present in the bioreactor ecosystem, one of which is a completely contiguous draft genome scaffolded with long reads. We identified core and accessory genes to investigate how gene expression patterns differ among the dominating strains. Using this approach, we were able to identify putative pathways and functions that may differ between Accumulibacter clades and provide key functional insights into this biotechnologically significant microbial lineage.
IMPORTANCE

‘Candidatus Accumulibacter phosphatis’ is a model polyphosphate accumulating organism that has been studied using genome-resolved metagenomics, metatranscriptomics, and metaproteomics to understand the EBPR process. Within the Accumulibacter lineage, several similar but diverging clades are defined by the polyphosphate kinase (ppk1) locus sequence identity. These clades are predicted to have key functional differences in acetate uptake rates, phage defense mechanisms, and nitrogen cycling capabilities. However, such hypotheses have largely been made based on gene-content comparisons of sequenced Accumulibacter genomes, some of which were obtained from different systems. Here, we performed time-series genome-resolved metatranscriptomics to explore gene expression patterns of co-existing Accumulibacter clades in the same bioreactor ecosystem. Our work provides an approach for elucidating ecologically relevant functions based on gene expression patterns between closely related microbial populations.
INTRODUCTION

Naturally occurring microbial assemblages are composed of closely related strains or subpopulations and harbor extensive genetic diversity. Coherent lineages are often comprised of distinct clades, ecotypes, or "backbone subpopulations" defined by high within-group sequence similarities that translate to ecologically relevant diversity (1–3).

Genetically diverse ecotypes of the marine cyanobacterium *Prochlorococcus* that are 97% identical by their 16S ribosomal RNA gene sequences exhibit markedly different light-dependent physiologies and distinct seasonal and geographical patterns (1, 2, 4, 5). Transcriptional-level differences between closely related microbial lineages have also been demonstrated, such as that of high-light and low-light *Prochlorococcus* ecotypes (6). However, dissecting the emergent ecological properties of genetic diversity observed among coherent microbial lineages is challenging given that these principles are likely not universal among the breadth of microbial diversity, and is further hindered by the lack of pure cultures (7).

Enhanced Biological Phosphorus Removal (EBPR) is an economically and environmentally significant process for removing excess phosphorus from wastewater. This process depends on polyphosphate accumulating organisms (PAOs) that polymerize inorganic phosphate into intracellular polyphosphate. ‘*Candidatus Accumulibacter phosphatis*’ (hereafter referred to as Accumulibacter) is a member of the *Betaproteobacteria* in the *Rhodocyclaceae* family and has long been a model PAO (8, 9). Accumulibacter achieves net phosphorus removal through cyclical anaerobic-aerobic phases of feast and famine. In the initial anaerobic phase, abundantly available volatile fatty acids (VFAs) such as acetate are converted to polyhydroxyalkanoates (PHAs), a
carbon storage polymer. However, PHA formation requires significant ATP input and reducing power, which are supplied through polyphosphate and glycogen degradation. In the subsequent aerobic phase in which oxygen is available as a terminal electron acceptor but soluble carbon sources are depleted, Accumulibacter uses PHA reserves for growth. Energy is recovered during aerobic PHA utilization by replenishing polyphosphate and glycogen reserves, which can be later used for future PHA formation as the anaerobic/aerobic cycle repeats. This feast-famine oscillation through sequential anaerobic-aerobic cycles operating on minute-to-hourly time scales gives rise to net phosphorus removal and the overall EBPR process (8, 10). A few ‘omics-based studies have revealed potentially regulatory modules that could coordinate Accumulibacter’s complex physiological behavior during these very dynamic environmental conditions (11–14).

Accumulibacter can be subdivided into two main types (I and II) and further subdivided into multiple clades based upon polyphosphate kinase (ppk1) sequence identity, since the 16S rRNA marker is too highly conserved among the breadth of known Accumulibacter lineages to resolve species-level differentiation (15–18). The two main types share approximately 85% nucleotide identity across the ppk1 locus (18) which mirrors genome-wide average nucleotide identity (ANI) (19). Members of different clades have been hypothesized to vary in their acetate uptake rates, nitrate reduction capacity, and phage defense mechanisms (19, 20). For example, clade IA seems to exhibit higher acetate uptake rates and phosphate release rates, whereas clade IIA can reduce nitrate and clade IA cannot (21, 22). The persistence of several phylogenetically distinct clades within the Accumulibacter lineage in both full-scale wastewater treatment plants and lab-
scale enrichment bioreactors suggests that different clades may inhabit distinct ecological
niches (23). This is largely supported by comparative genomics of numerous
Accumulibacter metagenome-assembled genomes (MAGs) recovered from different
bioreactor systems and classified using the \textit{ppk1} locus (12, 19–21, 24). However, the
ecological roles that distinct Accumulibacter clades may play is not yet clear.

Here, we used genome-resolved metagenomics and time-series
metatranscriptomics to investigate gene expression patterns of co-existing
Accumulibacter strains (representing two main clades) during a standard EBPR cycle.

We assembled high-quality genomes of the dominant and low abundant Accumulibacter
strains present in the bioreactor ecosystem. We constructed a contiguous draft genome
of an Accumulibacter clade IIC genome achieved using long Nanopore reads,
representing one of the highest-quality genomes for this clade recovered to date. Using
these assembled genomes, we performed time-series RNA-sequencing over the course
of a typical EBPR feast-famine cycle to explore expression profiles of different
Accumulibacter clades, with an emphasis on core and flexible gene content. More
generally, our work reveals putative ecological functions of co-existing taxa within a
dynamic ecosystem.
MATERIALS AND METHODS

Bioreactor Operation and Sample Collection

A laboratory-scale sequence batch reactor (SBR) was seeded with activated sludge from the Nine Springs Wastewater Treatment Plant (Madison WI, USA) during August 2015. The bioreactor was operated to simulate EBPR as described previously (24). Briefly, an SBR with a 2-L working volume was operated with the established biphasic feast-famine conditions: a 6-h cycle consisted of the anaerobic phase (sparging with N₂ gas), 110-min; aerobic (sparging with oxygen), 180-min; settle, 30-min; draw and feed with an acetate-containing synthetic wastewater feed solution, 40-min. A reactor hydraulic residence time of 12-h was maintained by withdrawing 50% of the reactor contents after the settling phase, then filling the reactor with fresh nutrient feed; a mean solids retention time of 4-d was maintained by withdrawing 25% of the mixed reactor contents each day immediately prior to a settling phase. Allythiourea was added to the synthetic feed solution to inhibit nitrification/denitrification.

Three biomass samples for metagenomic sequencing were collected in September of 2015 by centrifuging 2-mL of mixed liquor at 8,000 x g for 2 min, and DNA was extracted using a phenol:chloroform extraction protocol. Seven samples for RNA sequencing were collected throughout a single EBPR cycle (3 in the anaerobic phase, 4 in the aerobic phase) in October of 2016 following the sampling strategy conducted by Oyserman et al. (11). Samples were flash-frozen on liquid nitrogen immediately after centrifuging and discarding the supernatant. RNA was extracted using a TRIzol-based extraction method (Thermo Fisher Scientific, Waltham, MA) followed by phenol-chloroform separation and RNA precipitation. RNA was purified following an on-column
DNAse digestion using the RNase-Free DNase Set (Qiagen, Venlo, Netherlands) and cleaned up with the RNeasy Mini Kit (Qiagen, Venlo, Netherlands). Accumulibacter clade quantification was performed on biomass samples collected on the same day of the metatranscriptomics experiment using clade-specific ppk1 qPCR primers as described by Camejo et al. (23).

**Metagenomic and Metatranscriptomic Library Construction and Sequencing**

Metagenomic libraries were prepared by shearing 100 ng of DNA to 550 bp long products using the Covaris LE220 (Covaris) and size selected with SPRI beads (Beckman Coulter). The fragments were ligated with end repair, A-tailing, Illumina compatible adapters (IDT Inc) using the KAPA-Illumina library preparation kit (KAPA Biosystems). Libraries were quantified using the KAPA Biosystem next-generation sequencing library qPCR kit and ran on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were prepared for sequencing on the Illumina HiSeq platform using the v4 TruSeq paired-end cluster kit and the Illumina cBot instrument to create a clustered flow cell for sequencing. Shotgun sequencing was performed at the University of Wisconsin – Madison Biotechnology Center DNA Sequencing facility on the Illumina HiSeq 2500 platform with the TruSeq SBS sequencing kits, followed by 2x150 indexing. Raw metagenomic data consisted of 219.4 million 300-bp Illumina HiSeq reads with approximately 3.9 Gpb per sample (Table 1 – supplementary material). Nanopore sequencing was performed according to the user’s manual version SQK-LSK108.

Total RNA submitted to the University of Wisconsin-Madison Biotechnology Center was verified for purity and integrity using a NanoDrop 2000 Spectrophotometer and an Agilent 2100 BioAnalyzer, respectively. RNA-Seq paired end libraries were prepared
using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, CA). Each sample was processed for ribosomal depletion, using the Ribo-Zero™ rRNA Removal kit (Bacteria). mRNA was purified from total RNA using paramagnetic beads (Agencourt RNAClean XP Beads, Beckman Coulter Inc., Brea, CA) and fragmented by heating in the presence of a divalent cation. The fragmented RNA was then converted to cDNA with reverse transcriptase using SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, California, USA) with random hexamer priming and the resultant double stranded cDNA was purified. cDNA ends were repaired, adenylated at the 3’ ends, and then ligated to Illumina adapter sequences. Quality and quantity of the DNA were assessed using an Agilent DNA 1000 series chip assay and Thermo Fisher Qubit dsDNA HS Assay Kit. Libraries were diluted to 2nM, pooled in an equimolar ratio, and sequenced on an Illumina HiSeq 2500, using a single lane of paired-end, 100 bp sequencing, and v2 Rapid SBS chemistry.

**Metagenomic Assembly and Annotation**

We applied two different assembly and binning approaches to obtain high-quality genomes of the dominant Accumulibacter strains present in the reactors. For the first approach, Illumina unmerged reads were quality filtered and trimmed using the Sickle software v1.33 (25). Reads were merged with FLASH v1.0.3 22 (26), with a mismatch value of ≤0.25 and a minimum of ten overlapping bases from paired sequences, resulting in merged read lengths of 150 to 290 bp. FASTQ files were then converted to FASTA format using the Seqtk software v1.0 (27). Metagenomic reads from all three samples were then co-assembled using the Velvet assembler with a k-mer size of 65 bp, a minimum contig length of 200 bp, and a paired-end insert size of 300 bp (28). Metavelvet
was used to improve the assembly generated by Velvet (29). Metagenomic contigs were then binned using Maxbin (30). This approach allowed us to assemble high-quality genomes of clades IIA and IIC, named UW5 and UW6, respectively. Contigs from these genomes were scaffolded using long Nanopore reads that were generated under the standard protocol using MeDuSa (31), and manually inspected to remove short contigs and decontaminated using ProDeGe (32) and Anvi’o (33). Further scaffolding was performed on both of these genomes using Nanopore long reads using LINKS (34) and GapCloser (35).

Because this pipeline did not produce high-quality genomes of clades IA and IIF, we applied a different approach to assemble these genomes. Quality filtered reads from each metagenome were individually assembled using metaSPAdes and co-assembled together also using metaSPAdes (36). Reads from each metagenome were mapped against each of the assemblies using bbmap with a 95% sequence identity cutoff to obtain differential coverage (37), and contigs were binned using MetaBat (38). Identical clusters of bins were dereplicated across assemblies using dRep (39) to obtain the best quality genome of both clades IA and IIF, named UW4 and UW7, respectively. All genomes were quality checked using CheckM (40) and functional annotations assigned with Prokka and KofamKOALA (41, 42).

Each genome was assigned to a particular clade both by comparing the ppk1 sequence identity and ANI to previously published Accumulibacter genomes. A ppk1 database was created using sequences from previously published Accumulibacter references and clone sequences. We searched for the ppk1 gene in our draft genomes using this database and aligned the corresponding hits with MAFFT (43). A phylogenetic
tree of aligned *ppk1* gene sequences from Accumulibacter references, select clone sequences, and the outgroups *Dechloromonas aromatica RCB* and *Rhodocyclus tenuis DSM 110* with RAxML v8.1 with 100 rapid bootstraps (44). To use any given Accumulibacter reference genome in downstream comparisons, a confident *ppk1* hit had to be identified using the above methods. As a technical note, the IA-UW2 strain assembled by Flowers et al. (19) was renamed to UW3, as the original assembly contained an additional contig that was likely from a prophage. The UW3 strain assembly does not contain this contig, and therefore the UW numerical nomenclature follows in sequential order after UW3. A phylogenetic tree of *ppk1* nucleotide sequences was performed and compared to select clone sequences. Pairwise genome-wide ANI was calculated between the four Accumulibacter draft genomes and all available Accumulibacter genomes in Genbank with FastANI (45). A species tree of all Accumulibacter references and outgroup genomes was constructed using single-copy markers from the GTDBK-tk (46), aligned with MAFFT (43), built with RAxML with 100 rapid bootstraps (44), and visualized in iTOL (47). To compare genome-wide ANI scores of Accumulibacter references with pairwise nucleotide identity of the *ppk1* locus of all references, we performed pairwise BLAST for all coding *ppk1* coding regions and reported the percent identity (48).

**Identification of Core and Accessory Gene Content**

We identified core and accessory gene content of Accumulibacter clades through clustering orthologous groups of genes (COGs) using PyParanoid (49). We used the four Accumulibacter MAGs generated in this study, high-quality Accumulibacter reference genomes (above 90% completeness and less than 5% redundancy), and *Dechloromonas*
aromatica RCB and *Rhodocyclus tenuis DSM 110* as outgroups (Accumulibacter references listed in Supplementary Table 1 available at [https://figshare.com/account/projects/90614/articles/13237148](https://figshare.com/account/projects/90614/articles/13237148)). Briefly, an all-vs-all comparison of proteins for all genomes was performed using DIAMOND (50), pairwise homology scores calculated with the InParanoid algorithm (51), gene families constructed with MCL (52), and HMMs built for each gene family with HMMER (53). A phylogenetic tree was constructed for genomes used for the COGs analysis by using the coding regions of the PPK1 locus and overlaid with the presence/absence of core and flexible gene content using the ete3 python package (54). A similarity score between UW Accumulibacter genomes was calculated as the pairwise Pearson correlation coefficient between gene families using the numpy python package (55). The presence and absence of different gene families was visualized in an upset plot between UW Accumulibacter genomes using the ComplexUpset package based on the UpSetR package in R (56). All data wrangling and plotting was performed using the tidyverse suite of packages in R (57).

**Metatranscriptomics Mapping and Processing**

Metatranscriptomics reads from each of the seven samples were quality filtered using fastp (58) and rRNA removed with SortMeRNA (59). The coding regions of all four Accumulibacter genomes assembled in this study were predicted with Prodigal (60) and annotated with both Prokka and KofamKOALA (41, 42), and concatenated together to create a mapping index. Metatranscriptomic reads were competitively pseudoaligned to the reference index and counts quantified with kallisto (61). Genes with a sum of more than 10,000 counts across all 7 samples were removed, as these are likely ribosomal
RNAs that were not removed previously. Additionally, genes annotated as rRNAs by barrnap were manually removed (41). Low-count genes were removed by requiring that all genes had to have more than 10 counts mapped across 3 or more samples.

To explore genes and metabolic functions that may exhibit hallmark anaerobic-aerobic feast-famine cycling patterns, we identified sets of genes that were differentially expressed between anaerobic and aerobic conditions using DESeq2 (62). Differential expression calculations were performed for each strain separately to account for differences in the relative abundance of metatranscriptomic reads mapping back to each of the four MAGs. We analyzed expression patterns of differentially expressed genes using a threshold cutoff of $\pm 1.5 \log_2$ fold-change between anaerobic and aerobic conditions and then identified differentially expressed genes as core or accessory among clades IA (str. UW4) and IIC (str. UW6). For the core group analysis, the corresponding ortholog in each genome had to meet the minimum count threshold as described above to include in the comparison, with genes in one or both of the strains being differentially expressed. For the accessory analysis for each strain, the gene had to be differentially expressed and not contained in the other genome, but could be present in other strains or other genomes belonging to that clade.

**Data and Code Availability**

Raw sequencing files for the three metagenomes and seven metatranscriptomes have been submitted to NCBI under the BioProject accession PRJNA668760. Assembled UW5 and UW6 genomes have been deposited through IMG at accession numbers 2767802316 and 2767802455 respectively. All genomes assemblies used in this study in their current forms as well as transcriptomic count tables are available on Figshare at:
Enrichment of Co-Existing Accumulibacter Clades

We inoculated a 2-L sequencing-batch reactor with activated sludge from a full-scale wastewater treatment plant in Madison, WI, USA and maintained it for 40 months. The reactor was primarily fed with acetate and operated under cyclic anaerobic-aerobic phases with a 4-day solids retention time to simulate the EBPR process and enrich for Accumulibacter (Figure 1). During a period of stable EBPR (14 months following inoculation) in which acetate was completely diminished by the start of the aerobic phase and soluble phosphate was below 1.0 mg-L\(^{-1}\) at the end of the aerobic phase, we collected seven samples across a single cycle (three in the anaerobic phase, four in the aerobic phase) for RNA-sequencing (Figure 1A). We quantified the different Accumulibacter clades by performing qPCR of the polyphosphate kinase (ppk1) locus as described previously (23) from samples collected on the same day as the RNA-seq experiment (Figure 1B). At the time of the RNA-seq experiment, the bioreactor was primarily enriched in Accumulibacter clade IA, followed by clade IIC and clade IIA.

We assembled high-quality genomes of four Accumulibacter clades (Table 1), including the dominant IA and IIC (UW4 and UW6, respectively) clades and less abundant IIA and IIF clades (UW5 and UW7, respectively) (Figure 2). Each of the assembled genomes falls within the established Accumulibacter clade nomenclature as defined by ppk1 sequence identity and phylogenetic placement when compared to other publicly...
available Accumulibacter genomes and clone sequences (Figure 2A). Additionally, the
ppk1 hierarchical structure is also reflected by pairwise-average nucleotide identity (ANI)
boundaries between clades, with some exceptions for newly assembled Accumulibacter
genomes that fall outside of the established ‘C. Accumulibacter phosphatis’ lineage (Figure 2B). All assembled genomes are above 90% complete and contain less than 5%
redundancy as calculated by CheckM (40) (Table 1). We were able to scaffold the
assemblies of strains IIA-UW5 and IIC-UW6 using long Nanopore reads, constructing a
completely contiguous draft genome of clade IIC. To our knowledge, the assembled clade
IIC genome (UW6) is the most contiguous, highest-quality reference genome available
for this clade, providing a valuable new Accumulibacter reference genome (Figure 2C).

Distribution of Shared and Flexible Gene Content between Clades

We next characterized the functional diversity of different Accumulibacter clades
through clustering orthologous groups of genes (COGs) (Figure 3A). We performed an
updated COGs analysis from Skennerton et al. (20) and Oyserman et al. (63) to include
high-quality Accumulibacter references, and UW-generated genomes, and the outgroups
Dechloromonas aromatica and Rhodocyclus tenuis (Figure 3). The Accumulibacter
lineage exhibits a tremendous amount of functional diversity, with only approximately 25%
of COGs shared amongst all Accumulibacter clades (Figure 3A). Additionally, there is
extensive diversity within the two Accumulibacter types and within individual clades, with
only a few representatives of each clade more than approximately 75% similar to each
other by gene content (Supplementary Figure 1). Specifically, clade IIC broadly harbors
the most diversity of all sampled genomes by gene content, with representatives of this
clade not sharing more than 25% of COGs (Figure 3A). Other clades are more similar by
gene content, but this could be due to fewer genomes sampled from these clades. Most of the highest-quality, publicly available Accumulibacter genome references belong to clades IIF and IIC.

We then compared COGs between Accumulibacter clades collected from bioreactors seeded from the Nine Springs Wastewater Treatment Plant in Madison, as these genomes represent a shared origin (Figure 3B). In addition to two lab-scale bioreactors harboring distinct Accumulibacter clades (genomes UW1 and UW3, and UW4-7, respectively) (19, 24), we recently characterized clade IC (genome UW-LDO-IC) using genome-resolved metagenomics and metatranscriptomics (12). By calculating the Pearson correlation similarity of COG presence/absence, genomes in Type I and Type II cluster respectively, where genomes within Type I are overall more similar to each other, as expected (Supplementary Figure 1). Clade IIA genomes (strains UW1 and UW5) were assembled from separate enrichments derived from the same full-scale treatment plant approximately 12 years apart, but are very similar by gene families (Supplementary Figure 1). UW genomes within Type II are less similar by gene content and seem to be more diverse in this regard, though again this could be due to under-sampling of Type I (8 genomes in Type I and 14 in Type II). Additionally, clade IIF contains more overlap with Type I genomes in gene families than some Type II genomes do to each other (Supplementary Figure 1).

COG overlap and uniqueness is shown by the intersection of groups of COGs between UW Accumulibacter clades and the *Rhodocyclus* and *Dechloromonas* outgroups (Figure 3B). There are 1075 gene families shared among all UW Accumulibacter clades and outgroup genomes, with 277 separate gene families only
within the UW Accumulibacter clades (Figure 3B). Among individual clades, there are 887 groups only within clade IIC (UW6), 399 within clade IIA (UW1 and UW5), 399 within IIF (UW7), 292 within clade IA (UW3 and UW4), and 196 within IC (UW-LDO). Although clade IIC str. UW6 contains more overlap in gene content with SK-02 (Figure 3A), this intersection analysis identifies gene families that are available to individual strains within the same bioreactor ecosystem, highlighting the potential for functional differences. Below we explore these putative functional differences in shared and flexible gene content between Accumulibacter clades IA and IIC using time-series transcriptomics.

**Transcriptional Profiles of Accumulibacter Clades**

Although ppk1 locus quantification showed that the bioreactor was mostly enriched in clade IA (Figure 1B), surprisingly, more RNA-seq reads mapped to the clade IIC-UW6 genome than to the clade IA-UW4 (Figure 4A). Approximately 40 million total transcriptional reads mapped back to the clade IIC-UW6 genome across the RNA-seq experiment, whereas clades IA-UW4 and IIA-UW5 recruited roughly the same number of reads, with 14 million and 11 million, respectively (Table 2). Markedly lower reads levels mapped to the clade IIF-UW7 relative to the other three clades, and therefore we do not include clade IIF in our subsequent analyses.

The co-existing clades in the bioreactor system differ in both the presence of sets of nitrogen cycling gene sets and their expression profiles (Figure 4C). In clade IA-UW4, we detected the napAGH genes for nitrate reduction, whereas we did not detect napB, possibly due to these genes being close to the end of a contig in this genome. We also detected a nirS-like nitrite reductase, both subunits of the nitric oxide reductase norQD, and the nitrous oxide reductase nosZ in clade IA-UW4. Most of these genes in clade IA-
UW4 exhibit the highest expression patterns in the early and late anaerobic stages (Figure 4C). Conversely, clade IIC-UW6 contains the *narGHJI* machinery for nitrate reduction, a different *nirS*-like nitrite reductase than in IA-UW4, and does not contain a nitrous oxide reductase. However, clade IIC-UW6 does contain the nitrite reductase *nirBD* for reducing nitrite to ammonium, which is uniformly expressed across the cycle (Figure 4C).

We explored the top differentially expressed genes in clade IIC-UW6, comparing anaerobic and aerobic conditions (Figure 4B). Genes upregulated in the anaerobic phase belonged to pathways for central-carbon transformations, fatty acid biosynthesis, and PHA synthesis. Interestingly, all genes within the high-affinity phosphate transport system Pst were upregulated in the aerobic phase. This includes the substrate-binding component *pstS*, *phoU* regulator, and the ABC-type transporter *pstABC* complex. The higher overall transcriptional levels of clade IIC-UW6 versus clade IA-UW4 (Figure 4A) and upregulation of the high-affinity Pst system (Figure 4B) combined with the lower IIC-UW6 abundance compared to IA-UW4, suggests that clade IIC-UW6 compensates for lower abundance and competes for substrates through upregulation of transcriptional programs related to the EBPR process.

**Expression Dynamics of Genes Shared by Clades IA and IIC**

We next characterized expression dynamics of COGs that are shared between clades IA-UW4 and IIC-UW6 (Figure 5). There are very few orthologs that are differentially expressed between the anaerobic and aerobic phases in both clades that met the ±1.5-fold change threshold cutoff. These genes include a coproporphyrinogen III oxidase, NAD-reducing HoxS subunits, the *phaC* subunit of the PHA synthase, the long-
chain fatty-acid-CoA ligase FadD13, and other hypothetical proteins. The fatty-acid CoA-ligase incorporates ATP, CoA, and fatty acids of variable length to form acyl-CoA to be degraded for energy production, incorporated into complex lipids, or used in other metabolic pathways (64). The acyl-CoA ligase is upregulated in the anaerobic phase of both clades, suggesting a role in anaerobic carbon metabolism, as acyl-CoA can ultimately form acetyl-CoA through beta oxidation. The phaC subunit of the PHA synthase is differentially expressed in the anaerobic phase for both clades, as is typical of the hallmark EBPR metabolism for Accumulibacter (11, 65).

Additionally, both clades carry the bidirectional NiFe Hox system, encoding an anaerobic hydrogenase. The HoxH and HoxY beta and delta subunits form the hydrogenase moiety, and the FeS-containing HoxF and HoxU alpha and gamma subunits catalyze NAD(P)H/NAD(P)+ oxidation/reduction coupled to the hydrogenase moiety (66–68). In both clades, the FeS cluster HoxF and HoxU alpha and gamma subunits are strongly differentially expressed in the anaerobic phases, as was observed previously in a genome-resolved metatranscriptomic investigation of clade IIA-UW1 (11). However, clade IIC-UW6 is missing the HoxY delta subunit of the hydrogenase moiety, and only clade IA-UW4 exhibits differential expression of the HoxH beta subunit in the anaerobic phase. Clade IIC-UW6 does exhibit higher expression of HoxH in the late anaerobic and early aerobic phases, but does not meet the threshold cutoff for differential gene expression. The prior study focused on clade IIA-UW1 also demonstrated hydrogenase activity in the anaerobic phase, and also demonstrated hydrogen gas production after acetate addition (11). The hydrogen gas production was hypothesized to replenish NAD+ after glycogen degradation. In this ecosystem, both clades may exhibit anaerobic
hydrogenase activity and thus produce hydrogen, but due to the missing subunit in clade IIC-UW6 and lack of differential gene expression, this is uncertain.

Interestingly, several core ortholog subsets important for EBPR-related functions exhibit differential expression patterns in only one of the clades. For example, subunits of the high-affinity phosphate transporter Pst system are only differentially expressed in clade IIC-UW6 and not IA-UW4, as observed for clade IIC-UW6 above (Figure 4C). Additionally, nitrogen cycling genes that contain orthologs in both clades display different expression profiles. A nitrite reductase is only differentially expressed in the anaerobic phase in clade IA-UW4 and the denitrification protein NirQ is differentially expressed in only clade IIC-UW6. Whereas the nitrogen fixation protein subunit NifU is differentially expressed in both clades. To activate acetate to acetyl-CoA, the high-affinity acetyl-CoA synthase (ACS) or low-affinity acetate kinase/phosphotransacetylase (AckA/Pta) pathways can be employed. Although the clade IA-UW4 assembly is missing an acetate kinase, a phosphate acetyltransferase is only differentially expressed in this clade. Within clade IIC, both the acetate kinase and phosphate acetyltransferase genes exhibit relatively low levels of expression over the time-series. Both clades contain several acetyl-CoA synthases that are highly expressed in the anaerobic phase, particularly upon acetate contact, but do not exhibit stark differential expression between the two phases. Previous work based on community proteomics and enzymatic assays have suggested that although both activation pathways are present and expressed in Accumulibacter, the high-affinity pathway is preferentially used (69, 70). Although a phosphate acetyltransferase is differentially expressed in clade IA, this gene has lower
transcripts relative to the acetyl-CoA synthase, providing more evidence that the lineage (specifically both clades IA and IIC) may employ the high-affinity system.

**Differential Expression of Flexible Genes in Clade IA-UW4**

We lastly characterized expression dynamics of flexible gene content between clades IA-UW4 and IIC-UW6, focusing specifically on clade IA-UW4 (Figure 6). A majority of the differentially expressed flexible genes in clade IIC-UW6 were annotated as hypothetical proteins, and thus we focused on putative differentiating features of clade IA-UW4. Flexible gene content for IA-UW4 can be defined as any ortholog that is not present in the UW6 genome, whether that ortholog is only contained within clade IA genomes, Type I as a whole, within other Type II or clade IIC genomes, or the Dechloromonas or Rhodocyclus outgroups. Thus, the absence of a particular ortholog in the clade IIC-UW6 genome but present in the IA-UW4 genome or other genomes was inferred as a lack of functional capability of IIC in this particular ecosystem. For example, this analysis identified several groups of genes that are annotated as hypothetical genes that contain orthologs only within clade IA genomes and that are differentially expressed between the anaerobic and aerobic cycles.

Particularly striking is the differential expression profiles of nitrogen cycling genes that do not contain orthologs in the IIC-UW6 genome. As stated above, clade IA-UW4 contains the nap system whereas clade IIC-UW6 contains the nar system for reducing nitrate to nitrite (Figure 4C). In clade IA-UW4, the napDGH subunits are differentially expressed between the anaerobic and aerobic phases, and these genes are present in the outgroups, as well as both Accumulibacter types except the IIC-UW6 genome (Figure 6). Both genomes contain a nirS nitrite reductase, but the respective homologs are not
within the same COG, and thus clade IA-UW4 contains a different nirS that is upregulated in the anaerobic phase. Clade IIC-UW6 does not contain both subunits of the nitric-oxide reductase norQD, whereas clade IA-UW4 does. These subunits are not differentially expressed above the set threshold but do exhibit the highest expression upon acetate contact in the anaerobic phase (Figure 4B). Additionally, clade IIC does not contain the nitrous-oxide reductase nosZ (Figure 4B), whereas clade IA does and this is differentially expressed in the anaerobic phase (Figure 6). These results suggest that clade IA-UW4 may have the capability for full denitrification from nitrate to nitrogen gas, or at least use the reduction of nitrate/nitrite as a source of energy in the anaerobic phase in a denitrifying bioreactor. Although clade IIC-UW6 does contain the nirBD nitrite reductase complex for reducing nitrite to ammonium, these genes are uniformly expressed throughout the cycle (Figure 4B).

**DISCUSSION**

In this work, we performed time-series transcriptomics to explore gene expression profiles of two Accumulibacter strains co-existing within the same bioreactor ecosystem. By applying genome-resolved metagenomic and metatranscriptomic techniques, we were able to postulate niche-differentiating features of distinct Accumulibacter clades based on core and accessory gene content and differential expression profiles. We assembled high-quality genomes from Accumulibacter clades, using long reads to generate a single contiguous scaffold of the IIC-UW6 genome. By performing an updated clustering of orthologous groups of genes, we were able to better demonstrate the similarity and uniqueness among Accumulibacter strains from this system. Unexpectedly, although
clade IA dominates in abundance by ppk1 locus quantification, clade IIC is transcriptionally more active. By comparing the expression profiles of COGs shared between clades IA-UW4 and IIC-UW6, we found that genes involved in EBPR-related pathways may differ between the clades with respect to transcription-level regulation. Notably, clade IIC-UW6 shows a strong upregulation of the high-affinity Pit phosphorus transport system in the aerobic phase and both clades made hydrogenase transcripts in the anaerobic phase. Clade IA-UW4 contains most of the subunits for the individual key steps for denitrification, and exhibits strong anaerobic upregulation of these genes, suggesting a differentiating role for clade IA-UW4 in nitrogen cycling.

The Accumulibacter lineage harbors extensive diversity between and amongst individual clades which is exhibited not only by ppk1-based phylogenies, but also by pairwise genome-wide ANI and COG comparisons. For example, clade IIA-UW1 and clade IA-UW3 were present in the same bioreactor ecosystem and are only similar by 85% ANI (19). In the bioreactor enrichment from this study, the dominant clades IA-UW4 and IIC-UW6 are less than 80% similar by genome-wide ANI (Figure 2B). Recently, a cutoff of >95% ANI has been suggested to delineate distinct bacterial species or cohesive genetic units (45), originally based on observations of coverage gaps at 90-95% sequence identity when metagenomic reads were mapped back to assembled MAGs (71–74). These sequence cutoffs have also been benchmarked against homologous recombination rates and ratios of nonsynonymous to synonymous (dN/dS) nucleotide differences (75). However, there are some notable exceptions to the 95% ANI threshold cutoff used for delineating species, as distinct populations of the marine SAR11 clade have been observed at a 92% boundary and lower (76, 77). Among Accumulibacter
genomes adhering to the established ppk1-based phylogeny, genomes within individual clades fall within the >95% ANI threshold, whereas ANI comparisons within and between Types I and II are more variable (Figure 2B). Genomes within Type I are similar by approximately 90% ANI, whereas genomes within Type II are not as coherent.

The apparent coexistence of diverse Accumulibacter clades within an ecosystem then raises questions of ecological roles, putative interactions, and evolutionary dynamics of this lineage over time and space. Although the concept of a bacterial species is highly contested and dependent upon the operational definition of the population in question, species have been clustered based on phenotypic and metabolic characteristics, and more recently by DNA-based clustering metrics and recombination signatures (3, 7). The ecotype model of speciation suggests that cohesive, ecologically distinct ecotypes can coexist because they occupy separate niche spaces (3, 78). Because Accumulibacter has not been isolated in pure culture to date, most ecological and evolutionary inferences have been made from genome-based comparisons (20, 63) or mapping metagenomic reads back to assembled genomes (79, 80). However, the apparent diversity within the Accumulibacter lineage may require reevaluation of the phylogeny and taxonomy overall.

Although the ppk1-based clade designations coincide with genome-wide ANI based cutoffs remarkably well (Supplementary Figure 2B), there are some discrepancies. A few Accumulibacter genomes have been assembled that do not adhere to the ppk1-based population structure, but fall within the Accumulibacter lineage based on a species tree of single-copy core genes (Supplementary Figure 2A), such as the proposed novel species ‘Candidatus Accumulibacter aalborgensis’ (81). Additionally, a few genomes preliminarily classified by the GTDB as Accumulibacter branch outside of the established
nomencature, and are quite divergent based on the species and \textit{ppk1} phylogenies as well as genome-wide and \textit{ppk1} sequence similarities (Figure 2B, Supplementary Figure 2). Overall, substantial work is needed to reevaluate the phylogenetic structure of the Accumulibacter lineage as a whole and connect with signatures of homologous recombination and evolutionary trajectories within coexisting clades to understand the genetic boundaries and thus ecological interdependencies of individual clades or strains.

Although the boundaries of individual Accumulibacter clades or strains have not been entirely resolved to infer separate ecotypes, denitrification capabilities of specific clades has been repeatedly suggested to be a niche-differentiating feature (12, 21, 23). Previously, two separate bioreactors fed with either acetate or propionate as the sole carbon source were enriched with different Accumulibacter morphotypes, one with cocci morphology and the other rod morphology, which were also linked to different denitrification capabilities (22). A bioreactor highly enriched in clade IC inferred simultaneous use of oxygen, nitrite, and nitrate as electron acceptors under micro-aerobic conditions either due to metabolic flexibility or heterogeneity within the IC population (23). The resident strain IC-UW-LDO also possessed the full suite of genes for denitrification and expressed them in microaerobic conditions (12); whereas other clades either do not contain the complete genetic repertoire for denitrification or have not been shown to express them. As we observed differential expression of denitrification genes in clade IA-UW4 but not clade IIC-UW6 in this study, these expression differences could explain why the two clades can coexist and contribute to overall ecosystem functioning (e.g. phosphorus removal) by using different electron acceptors in the anaerobic/aerobic conditions. Although the synthetic feed for the bioreactors in this study inhibits
nitrification/denitrification through alythiourea addition, the expression profiles of
denitrification genes in clade IA-UW4 suggests this strain could have a differentiating role
in a denitrifying bioreactor. Recently, a denitrifying EBPR bioreactor enriched in clade IA
encoded genes for complete denitrification, whereas the co-occurring clade IC strain and
other flanking community members did not (82). Additionally, Gao et al. performed an
ancestral genome reconstruction analysis similar to Oyserman et al. 2017 (63) to
specifically understand the flux of denitrification gene families. Interestingly, the nir gene
cluster was identified as a core COG that is core to all sampled Accumulibacter genomes,
whereas the periplasmic nitrate reductase napAGH and nitrous oxide reductase nosZDFL
clusters were core to only Type I Accumulibacter genomes, and inferred to have been
lost for some Type II genomes (82). We observed differential expression of nitrate
reductase and nitrous-oxide reductase genes in clade IA-UW4, where homologs of these
genes were not present in the clade IIC-UW6 genome and other Type II genomes.

Accumulibacter's anaerobic metabolism has also been shown to differ under
varying substrate or stoichiometric conditions and possibly between the two
Accumulibacter types. From recent simulations of 'omics datasets paired with enzymatic
assays, the routes of anaerobic metabolism for Accumulibacter may differ based on
environmental conditions (83). Under high acetate conditions, polyphosphate
accumulation arises through the glyoxylate shunt, whereas under increased glycogen
concentrations, glycolysis is used in conjunction with the reductive branch of the TCA
cycle (83). Furthermore, Welles et al. found that Type II Accumulibacter is able to switch
to partial glycogen degradation more quickly than Type I, enabling Type II to fuel VFA
uptake in polyphosphate-limited conditions (84, 85). Conversely, when concentrations of
polyphosphate are high, VFA uptake rates are higher in Type I. However, these studies only identified Accumulibacter lineages based on ppk1 types, and therefore extrapolated metabolic assumptions from presumably more resolved strains to the entire type. These broad interpretations may obscure apparent functional differences within the two types and individual clades. Future experiments could integrate multi-omics approaches to investigate not only metabolic flexibility under different substrate conditions and perturbations, but how metabolisms are distributed amongst Accumulibacter clades in bioreactor ecosystems that are not completely enriched in a single clade or strain.

CONCLUSIONS

In this study, we integrated genome-resolved metagenomics and time-series metatranscriptomics to understand gene expression patterns of co-existing Accumulibacter clades within a bioreactor ecosystem. We found evidence for denitrification gene expression in the dominating clade IA, but higher overall transcriptional activity for clade IIC along with differential expression of a high-affinity phosphorus transporter. These results highlight the need for reevaluating the phylogenetic structure of the Accumulibacter lineage, and defining how potential genetic boundaries may translate to individual ecological niches allowing for the coexistence of multiple clades or strains. Additionally, we highlight an approach for exploring functions of closely related, uncultivated microbial lineages by exploring gene expression patterns of shared and accessory genes. Overall, this work underscores the need to understand the basic ecology and evolution of microorganisms underpinning biotechnological
processes to better control treatment process and resource recovery outcomes in the future (86).

ACKNOWLEDGEMENTS

We thank Dr. Jeffrey Lewis for feedback on the differential expression analyses and interpretation. This work was supported by funding from the National Science Foundation (MCB-1518130) to K.D.M. and D.R.N. F.M. and P.C. received funding from the Chilean National Commission for Scientific and Technological Research (CONICYT). E.H.K. was supported by the 2017 UW-Madison Hilldale Scholarship. E.A.M. was supported by a University of Wisconsin – Madison Department of Bacteriology Fellowship. We thank the University of Wisconsin – Madison Biotechnology Center DNA Sequencing Facility for Illumina sequencing services. This research was performed in part using the Wisconsin Energy Institute computing cluster, which is supported by the Great Lakes Bioenergy Research Center as part of the U.S. Department of Energy Office of Science.
REFERENCES


Li H. 2013. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences.


accurately reconstructing single genomes from complex microbial communities.

PeerJ 3:e1165.


47. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the


Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of


**FIGURE AND TABLE LEGENDS**

**Figure 1. Overview of Experimental Design and Enriched Accumulibacter Clades.**

1A. Overview of EBPR cycle and sampling scheme. Samples were collected for RNA sequencing over the course of a normal EBPR cycle, with reported measurements for soluble phosphorus and acetate normalized by volatile suspended solids (VSS). Samples were taken at time-points 0, 31, 70, 115, 150, 190, and 250 minutes. 1B. Abundance of different Accumulibacter clades by ppk1 copies per ng of DNA. Quantification of Accumulibacter clades was performed by qPCR of the ppk1 gene as described in Camejo et al. (23).

**Figure 2. Assembly of Four High-Quality Accumulibacter Genomes**

2A. Phylogenetic tree of ppk1 nucleotide sequences from existing Accumulibacter references, clone sequences, and the four assembled Accumulibacter genomes from this study. Tips in bold represent ppk1 sequences from metagenome-assembled genomes, whereas others are sequences from clone fragments available at the referenced accession numbers. Starred tips represent ppk1 sequences from the four assembled Accumulibacter genomes from this study. Tree was constructed using RaXML with 100 rapid bootstraps. 2B. Pairwise genome-wide ANI comparisons of all publicly available Accumulibacter references and the four new Accumulibacter genomes, denoted in bold. Genomes are grouped according to clade designation by ppk1 sequence identity and in order of the ppk1 phylogeny, except for those that do not fall in an established clade. 2C. Genome diagram of the UW6 Accumulibacter IIC MAG. Layers represent from top to
bottom: coverage over 1000 bp windows, GC content and skew across the same 1000 bp windows, coding sequences on the positive and negative strands, and positions of predicted rRNA (red) and tRNA (black) sequences.

**Figure 3. Core and Flexible Gene Content among the Accumulibacter lineage and UW genomes**

3A. Phylogenetic tree of the coding regions for the PPK1 locus among high-quality Accumulibacter references, UW genomes, and outgroup taxa *Dechloromonas aromatica RCB* and *Rhodocyclus tenuis DSM 110*. Pie chart at each node represents the fraction of genes that are shared and flexible for genomes belonging to that clade. 3B. Upset plot representing the intersections of groups of orthologous gene clusters among UW genomes and outgroups, analogous to a Venn diagram. Each bar represents the number of orthologous gene clusters, and the dot plot represents the genomes in which the groups intersect.

**Figure 4. Transcriptomic Profiles of Accumulibacter Clades**

4A. Average expression profiles of each reference genome in the anaerobic and aerobic phases. Reads were competitively mapped to the four assembled Accumulibacter clade genomes and normalized by transcripts per million (TPM). The sum of the total counts of the three samples in the anaerobic phase and four samples in the aerobic phase were averaged, respectively, and plotted on a log10 scale. 4B. Top 50 differentially-expressed genes in clade IIC-UW6. All 50 genes pass the threshold of ±1.5 log-10 fold-change across the cycle. Annotations were predicted from a combination of Prokka (41) and
KofamKOALA (42). **4C.** Denitrification gene expression among clades IIC-UW6, IA-UW4, and IIA-UW5. We summarized expression for genes involved in denitrification by highlighting parts of the cycle in which certain genes are expressed the most. This is characterized by acetate contact (pink), early anaerobic (dark purple), late anaerobic (purple), early aerobic (dark blue) and late aerobic (light blue), where uniform describes uniform expression across the entire cycle (grey). If an arrow is missing for a step, that particular clade did not contain confident annotations within the genome for those genes. For nitrate reduction, the *nar* system is represented as a black dot and the *nap* system is represented as a grey dot. Multiple arrows are shown for a single step in which particular genes were highly expressed in multiple phases.

**Figure 5. Differential Expression of Shared Genes in Clades IA and IIC**

Gene expression profiles of core COGs across the EBPR cycle that are differentially expressed between anaerobic and aerobic conditions in either IA-UW4 or IIC-UW6, or both strains. To consider a gene differentially expressed in a strain, the COG must exhibit greater or less than ±1.5 fold-change in expression. Homologs that are directly compared to each other in IA-UW4 and IIC-UW6 are orthologous genes that belong to the same cluster.

**Figure 6. Differential Expression of Accessory Genes in Clade IA**

Gene expression profiles of groups of accessory gene groups in clade IA-UW4 that are not present in the UW6-IIC genome. A gene was considered differentially expressed between the anaerobic and aerobic cycles if it exhibited greater or less than ±1.5 fold.
change in expression. Colored dots represent whether that gene contains orthologs within outgroup genomes, other Type II Accumulibacter other than the UW6-IIC genome, within Type I Accumulibacter, or only within clade IA.

**Table 1. Assembled Accumulibacter Genome Statistics**

Genome quality calculations for completeness and contaminations were made with CheckM based on the presence/absence of single copy genes (40).

**Table 2. Metatranscriptomic Reads Mapped to Accumulibacter Genomes**

Counts of raw metatranscriptomic reads mapped to each Accumulibacter genome with kallisto (61) in each sample for the three anaerobic samples and four aerobic samples, and total reads mapped for all samples.

**Supplementary Figure 1.**

Pairwise Pearson correlation coefficient of shared gene content as a measure of how similar each of the UW genomes are to each other based on COGs.

**Supplementary Figure 2.**

*S2A*) Species tree of all Accumulibacter reference genomes, UW-assembled genomes, and outgroup genomes using single-copy markers from the GTDB-tk (46). Phylogenetic tree was constructed using RAxML with 100 rapid bootstraps and visualized in iTOL. *S2B*) Pairwise BLAST percent identity scores for *ppk1* coding regions for each reference genome in Figure 2B, in order of the *ppk1* phylogeny in Figure 2A. Raw pairwise percent
identity scores are available at

https://figshare.com/articles/dataset/Pairwise_Accumulibacter_ppk1_BLAST_Percent_D_Scores/13237247.

Supplementary Table 1

Genome accessions and statistics for all Accumulibacter references used in phylogenies, ANI comparisons, and COG calculations. A reference genome was used for COG comparisons if it was a high quality genome (>95% complete and <5% redundant).

Available at https://figshare.com/account/projects/90614/articles/13237148

Supplementary File 1

Raw pairwise genome-wide ANI results for each genome calculated with FastANI.

Available at https://figshare.com/account/projects/90614/articles/13237244

Supplementary File 2

Raw pairwise BLAST percent identity (PID) results for ppk1 genes from each reference genome. Available at https://figshare.com/account/projects/90614/articles/13237247
A. Aerobic vs Anaerobic Time in Cycle (min)

Soluble Phosphorus

B. Log10 Copies/ng DNA of ppk1 by Clade

Clade
- IA
- IIC
- IIA

Analyte (mg/VSS-mg)

- 0.00
- 0.04
- 0.08
- 0.12
- 0.16

Log10 Copies/ng DNA

- 3.39x10^5
- 1.30x10^5
- 9.03x10^3
Dechloromonas aromatica RCB
Rhodocyclus tenuis DSM 110
Candidatus Accumulibacter UW3
Accumulibacter sp. CANDO1
Accumulibacter sp. BA-93
Accumulibacter sp. UBA6658
Candidatus Accumulibacter UW4
Accumulibacter sp. HKU1
Accumulibacter sp. BA-92
Candidatus Accumulibacter UW1
Candidatus Accumulibacter UW5
Candidatus Accumulibacter UW7
Accumulibacter sp. SCELSE1
Accumulibacter sp. UBA2315
Accumulibacter sp. UBA6585
Accumulibacter sp. SK-12
Accumulibacter sp. UBA2327
Accumulibacter sp. UBA5574
Accumulibacter sp. SK-01
Accumulibacter sp. BIN19
Accumulibacter sp. HKU2
Accumulibacter sp. SK-02
Candidatus Accumulibacter UW6

Type I
IC
IA
IB
IIC
Flexible gene content
Shared portion of clade
Tree scale: 0.075

Type II
IIF
IIA
Intersection Size
(Number of Orthologous Groups)
0
300
600
900
1075
887
689
399
292
277
240
196
164
146
144
121
115
110
109
101
98
80
70
68
67
50
49
47
45

A. Rhodocyclus
Dechloromonas
UW-LDO (IC)
UW4 (IA)
UW3 (IA)
UW7 (IIF)
UW1 (IIA)
UW5 (IIA)
UW6 (IIC)

Outgroup
Type I
Type II

B. (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.
The copyright holder for this preprint this version posted November 23, 2020. doi: bioRxiv preprint
A.  

<table>
<thead>
<tr>
<th>Sample</th>
<th>Anaerobic Phase</th>
<th>Aerobic Phase</th>
<th>Total Reads</th>
<th>% Total Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIC-UW6</td>
<td></td>
<td></td>
<td>41.8 million</td>
<td>61.7%</td>
</tr>
<tr>
<td>IA-UW4</td>
<td></td>
<td></td>
<td>14 million</td>
<td>20.7%</td>
</tr>
<tr>
<td>IIA-UW5</td>
<td></td>
<td></td>
<td>11.3 million</td>
<td>16.7%</td>
</tr>
<tr>
<td>IIF-UW7</td>
<td></td>
<td></td>
<td>11.3 million</td>
<td>0.85%</td>
</tr>
</tbody>
</table>

B.  

The image shows a log10 average expression graph comparing anaerobic and aerobic phases. The graph includes a heat map with fold-change values and gene expression levels. The legend includes various genes and their expression levels across different oxidation states.

C.  

The figure illustrates the oxidation state changes from NO3 to NH3, with Clades IIC, IA, and IIA shown. The oxidation states are depicted as follows: +5 (Nitrate), +3 (Nitrite), +2 (Nitric oxide), +1 (Nitrous oxide), 0 (Nitrogen), and -3 (Ammonia). The clades are color-coded for easy visualization.
<table>
<thead>
<tr>
<th>Protein</th>
<th>anaerobic Clade IA-UW4</th>
<th>anaerobic Clade IIC-UW6</th>
<th>aerobic Clade IA-UW4</th>
<th>aerobic Clade IIC-UW6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome c-552</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Putative chemotaxis protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low affinity K+ transporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein HtpX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine desulfurase IscS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine desulfurase NifS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptional regulator IscR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coproporphyrinogen III oxidase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ammonia channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methyltransferase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycerol-3-phosphate dehydrogenase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphoribulokinase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptional repressor CarH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrogen fixation protein NifU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2Fe-2S cluster assembly protein IscU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2Fe-2S ferredoxin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrogen regulatory protein P-II 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribulose bisphosphate carboxylase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flagellin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diguanylate cyclase DgcM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate acetyltransferase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD-reducing hydrogenase HoxS-B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exodeoxyribonuclease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denitrification protein NirQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferredoxin-NADP reductase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytochrome c4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP-synthase delta subunit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP-synthase subunit c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP-synthase subunit c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD-reducing hydrogenase HoxS-A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD-reducing hydrogenase HoxS-G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriohemerythrin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iron uptake protein A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive-response sensory-kinase SasA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Putative redox modulator Aix</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly(3-hydroxyalkanote) polymerase PhaC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-chain fatty-acid-CoA ligase FadD13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensor protein QseC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate transport system PhoU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate transport system PstA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptional regulatory protein QseB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly(3-hydroxyalkanote) polymerase PhaC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-chain fatty-acid-CoA ligase FadD13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensor protein QseC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrite reductase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfoxide reductase MsrB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Heatmap showing the differential expression of proteins in Clade IA-UW4 and Clade IIC-UW6 under anaerobic and aerobic conditions. The color scale represents the log2 fold-change, with red indicating upregulation and green indicating downregulation.
<table>
<thead>
<tr>
<th>Anaerobic</th>
<th>Aerobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein CbbY</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Na+/H+ Antiporter</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Na+/H+ Antiporter</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Nitrate reductase (cytochrome)</td>
<td>Nitrate reductase NapD</td>
</tr>
<tr>
<td>Nitrate reductase NapD</td>
<td>Ferredoxin-type protein NapG</td>
</tr>
<tr>
<td>Ferredoxin-type protein NapG</td>
<td>NAD-reducing hydrogenase HoxS</td>
</tr>
<tr>
<td>NAD-reducing hydrogenase HoxS</td>
<td>Nucleoid occlusion factor SlmA</td>
</tr>
<tr>
<td>Nucleoid occlusion factor SlmA</td>
<td>Cytochrome c4</td>
</tr>
<tr>
<td>Cytochrome c4</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Putative zinc metalloprotease</td>
</tr>
<tr>
<td>Putative zinc metalloprotease</td>
<td>Acetylglutamate kinase</td>
</tr>
<tr>
<td>Acetylglutamate kinase</td>
<td>Dethiobiotin synthase</td>
</tr>
<tr>
<td>Dethiobiotin synthase</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Nitrous-oxide reductase</td>
</tr>
<tr>
<td>Nitrous-oxide reductase</td>
<td>Nitrite reductase</td>
</tr>
<tr>
<td>Nitrite reductase</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Methyltransferase</td>
</tr>
<tr>
<td>Methyltransferase</td>
<td>Phospholipase</td>
</tr>
<tr>
<td>Phospholipase</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Type IV Secretion system</td>
</tr>
<tr>
<td>Type IV Secretion system</td>
<td>Transcriptional regulator NtrC</td>
</tr>
<tr>
<td>Transcriptional regulator NtrC</td>
<td>Ammonia channel</td>
</tr>
<tr>
<td>Ammonia channel</td>
<td>C4-dicarboxylate transport protein</td>
</tr>
<tr>
<td>C4-dicarboxylate transport protein</td>
<td>Glutamate racemase</td>
</tr>
<tr>
<td>Glutamate racemase</td>
<td>Putative transposase</td>
</tr>
<tr>
<td>Putative transposase</td>
<td>Conjugal transfer protein TraG</td>
</tr>
<tr>
<td>Conjugal transfer protein TraG</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</td>
</tr>
</tbody>
</table>

Presence of Orthologs:
- **Outgroup**
- **Type I**
- **Type II**
- **Clade IA only**

Log2 fold-change:
- 2
- 1
- 0
- -1
<table>
<thead>
<tr>
<th>Genome</th>
<th>Completion</th>
<th>Contamination</th>
<th>Size_Mbp</th>
<th>Contigs</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIA-UW5</td>
<td>98.99</td>
<td>5.24</td>
<td>4.88</td>
<td>68</td>
<td>64.3</td>
</tr>
<tr>
<td>IIC-UW6</td>
<td>98.57</td>
<td>2.46</td>
<td>5.18</td>
<td>1</td>
<td>61.1</td>
</tr>
<tr>
<td>IIF-UW7</td>
<td>98.97</td>
<td>3.04</td>
<td>4.87</td>
<td>102</td>
<td>66.3</td>
</tr>
<tr>
<td>IA-UW4</td>
<td>92.90</td>
<td>3.37</td>
<td>4.29</td>
<td>358</td>
<td>64.0</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IIA-UW5</td>
<td>1063676</td>
<td>1357480</td>
<td>1706229</td>
<td>2197556</td>
<td>1763359</td>
</tr>
<tr>
<td>IIC-UW6</td>
<td>5226724</td>
<td>5892548</td>
<td>6755519</td>
<td>7902549</td>
<td>5004518</td>
</tr>
<tr>
<td>IIF-UW7</td>
<td>51038</td>
<td>78249</td>
<td>84520</td>
<td>121226</td>
<td>86070</td>
</tr>
<tr>
<td>IA-UW4</td>
<td>1484341</td>
<td>2216299</td>
<td>2290441</td>
<td>3362431</td>
<td>1950259</td>
</tr>
<tr>
<td></td>
<td>UW3-IA</td>
<td>UW4-IA</td>
<td>UW-LDO-IC</td>
<td>UW5-IIA</td>
<td>UW1-IIA</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>UW3-IA</td>
<td>100.0%</td>
<td>78.3%</td>
<td>62.9%</td>
<td>40.9%</td>
<td>40.4%</td>
</tr>
<tr>
<td>UW4-IA</td>
<td>78.3%</td>
<td>100.0%</td>
<td>57.2%</td>
<td>34.0%</td>
<td>33.3%</td>
</tr>
<tr>
<td>UW-LDO-IC</td>
<td>62.9%</td>
<td>57.2%</td>
<td>100.0%</td>
<td>40.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>UW5-IIA</td>
<td>40.9%</td>
<td>34.0%</td>
<td>40.0%</td>
<td>100.0%</td>
<td>92.3%</td>
</tr>
<tr>
<td>UW1-IIA</td>
<td>40.4%</td>
<td>33.3%</td>
<td>40.0%</td>
<td>92.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>UW6-IIC</td>
<td>33.3%</td>
<td>29.2%</td>
<td>35.7%</td>
<td>35.7%</td>
<td>36.0%</td>
</tr>
<tr>
<td>UW7-IIIF</td>
<td>48.8%</td>
<td>41.5%</td>
<td>48.7%</td>
<td>52.0%</td>
<td>50.1%</td>
</tr>
</tbody>
</table>
A.

Clade IA
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. CANDO_1
- Candidatus Accumulibacter sp. UW3
- Candidatus Accumulibacter sp. HKU1
- Candidatus Accumulibacter sp. BA-92
- Candidatus Accumulibacter sp. UBA2783
- Candidatus Accumulibacter sp. UW-LDO
- Candidatus Accumulibacter sp. UBA11064

Clade IB

Clade IC

Clade IIA
- Candidatus Accumulibacter sp. UBA704
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. CANDO_1
- Candidatus Accumulibacter sp. UW3
- Candidatus Accumulibacter sp. HKU1
- Candidatus Accumulibacter sp. BA-92
- Candidatus Accumulibacter sp. UBA2783
- Candidatus Accumulibacter sp. UW-LDO
- Candidatus Accumulibacter sp. UBA11064

Clade IIB

Clade IIF

Dechloromonas aromatica RCB

Tree scale: 0.01

B.

Pairwise BLAST % Identity Score of ppk1

Clade IIA
- Candidatus Accumulibacter sp. UW4
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. UW3
- Candidatus Accumulibacter sp. BA-92
- Candidatus Accumulibacter sp. UBA2783
- Candidatus Accumulibacter sp. UWB

Clade IIB

Clades IB & IC
- Candidatus Accumulibacter sp. CANDO_1
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. UW4

Clade IIA
- Candidatus Accumulibacter sp. UBA704
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. UW3
- Candidatus Accumulibacter sp. BA-92
- Candidatus Accumulibacter sp. UBA2783
- Candidatus Accumulibacter sp. UWB

Clade IIB

Clades IB & IC
- Candidatus Accumulibacter sp. CANDO_1
- Candidatus Accumulibacter sp. UBA6658
- Candidatus Accumulibacter sp. BA-93
- Candidatus Accumulibacter sp. UW4