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

Elizabeth A. McDaniel¹*, Francisco Moya-Flores²,³,⁴*, Natalie Keene Beach²,⁶, Pamela Y. Camejo²,⁵, Ben O. Oyserman⁷,⁸, Matthew Kizaric², Eng Hoe Khor², Daniel R. Noguera²,⁹, 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 Geroscience Center for Brain Health and Metabolism (GERO), Santiago, 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 Carollo Engineers, Inc., Broomfield, CO, USA
⁷ Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands
⁸ Microbial Ecology, Netherlands Institute of Ecological Research, Wageningen, The Netherlands
⁹ DOE Great Lakes Bioenergy Research Center, Madison, WI, USA

* equal contribution
† Corresponding Authors: elizabethmcd93@gmail.com and kdmcmahon@wisc.edu
Natural microbial communities consist of closely related taxa 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 strains 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 Nanopore reads. We identified core and accessory genes to investigate how gene expression patterns differed among the dominating strains. Using this approach, we were able to identify putative pathways and functions that may confer distinct functions to Accumulibacter strains 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 shared sequence identity of the polyphosphate kinase (ppk1) locus. 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). For example, 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 when there is a lack of pure cultures (7–9).

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 (10, 11). 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 is 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. Accumulibacter recovers energy during aerobic PHA degradation 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 (10, 12). A few ‘omics-based studies have revealed potential regulatory modules that could coordinate Accumulibacter’s complex physiological behavior during these very dynamic environmental conditions (13–16), but few have attempted to assign unique gene expression dynamics to specific clades (17).

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 (18–21). The two main types share approximately 85% nucleotide identity across the ppk1 locus (21) which mirrors genome-wide average nucleotide identity (ANI) (22). Members of different clades have been hypothesized to vary in their acetate uptake rates, nitrate reduction capacity, and phage defense mechanisms (22, 23). For example, clade IA seems to exhibit higher acetate uptake rates and phosphate release rates, whereas evidence exists that clade IIA can reduce nitrate but clade IA cannot (24, 25). Additionally, the Accumulibacter
lineage exhibits specific trait variation within individual \textit{ppk1}-defined clades for specific traits, such as denitrification capabilities. For example, differences among clade IC strains with respect to anoxic nitrate reduction coupled to phosphorus uptake suggests that the ability to carry out each sequential step in denitrification could be strain-specific (14, 17, 26–28). 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 (17, 29). This is largely supported by comparative genomics of Accumulibacter metagenome-assembled genomes (MAGs) recovered from different bioreactor systems and classified using the \textit{ppk1} locus (14, 17, 22–24, 28, 30). However, the ecological roles that distinct Accumulibacter clades may play are 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 four high-quality genomes of the dominant (clades IA and IIC) and low abundant (clades IIA and IIF) Accumulibacter strains present in the bioreactor ecosystem. Additionally, we constructed a contiguous draft genome of an Accumulibacter clade IIC genome achieved using long Oxford Nanopore Technology (ONT) reads, representing one of the highest-quality genomes for this clade recovered to date. Using these assembled genomes, we performed time-series RNA-sequencing throughout a typical EBPR feast-famine cycle to study expression profiles of different Accumulibacter strains, 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 (30). 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. Allylthiourea was added to the synthetic feed solution to inhibit ammonia oxidation and indirectly 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. (13). 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. (29).

**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 Gbp per sample. 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. The 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 (31). Reads were merged with FLASH v1.0.3 22 (32), 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 (33). 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 (34). Metavelvet
was used to improve the assembly generated by Velvet (35). Metagenomic contigs were
then binned using Maxbin (36). This approach allowed us to assemble high-quality
genomes of clades IIA and IIC, named UW5 and UW6, respectively. Long Nanopore
reads were generated under the standard protocol and used to scaffold assembled
contigs with MeDuSa (37), and manually inspected to remove short contigs and
decontaminated using ProDeGe (38) and Anvi’o (39). Further scaffolding was performed
on both of these genomes using Nanopore long reads using LINKS (40) and GapCloser
(41).

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 and co-assembled assembled using metaSPAdes
(42). Reads from each metagenome were mapped against each of the assemblies using
bbmap with a 95% sequence identity cutoff to obtain differential coverage (43), and
contigs were binned using MetaBat (44). Identical clusters of bins were dereplicated
across assemblies using dRep (45) to obtain the best quality genome of both clades IA
and IIF, named UW4 and UW7, respectively. All genomes were quality checked using
CheckM (46) and functional annotations assigned with Prokka and KofamKOALA (47,
48).

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 (49). A phylogenetic
tree of alignedppk1gene sequences from Accumulibacter references, select clone
sequences, and the outgroupsDechloromonas aromatica RCB andRhodocyclus tenuiss
DSM 110 with RAxML v8.1 with 100 rapid bootstraps (50). To use any given
Accumulibacter reference genome in downstream comparisons, a confidentppk1hit had
to be identified using the above methods. As a technical note, the IA-UW2 strain
assembled by Flowers et al. (22) was renamed to UW3, as the original assembly
contained an additional contig that was likely from a prophage. The prophage contig was
removed from the IA-UW2 strain and renamed to UW3, and therefore the UW numerical
omenclature follows in sequential order after UW3. A phylogenetic tree ofppk1
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 (51). A species tree of all
Accumulibacter references and outgroup genomes was constructed using single-copy
markers from the GTDBK-tk (52), aligned with MAFFT (49), built with RAxML with 100
rapid bootstraps (50), and visualized in iTOL (53). To compare genome-wide ANI scores
of Accumulibacter references with pairwise nucleotide identity of theppk1locus of all
references, we performed pairwise BLAST for allppk1coding regions and reported the
percent identity (54).

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 (55). We used the four
Accumulibacter MAGs generated in this study (Table 1) and high-quality Accumulibacter
reference genomes (above 90% completeness and less than 5% redundancy) for this analysis, with *Dechloromonas aromatica* RCB and *Rhodocyclus tenuis* DSM 110 as outgroups (Accumulibacter references listed in Supplementary Table 1 available at https://figshare.com/articles/dataset/Accumulibacter_References_Metadata/13237148).

Briefly, an all-vs-all comparison of proteins for all genomes was performed using DIAMOND (56), pairwise homology scores calculated with the InParanoid algorithm (57), gene families constructed with MCL (58), and HMMs built for each gene family with HMMER (59). 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 (60). A similarity score between UW Accumulibacter genomes was calculated as the pairwise Pearson correlation coefficient between gene families using the numpy python package (61). The presence and absence of different gene families were visualized in an upset plot between UW Accumulibacter genomes using the ComplexUpset package based on the UpSetR package in R (62). All data wrangling and plotting were performed using the tidyverse suite of packages in R (63).

**Metatranscriptomics Mapping and Processing**

Metatranscriptomics reads from each of the seven samples were quality filtered using fastp (64) and rRNA removed with SortMeRNA (65). The coding regions of all four Accumulibacter genomes assembled in this study (Table 1) were predicted with Prodigal (66) and annotated with both Prokka and KofamKOALA (47, 48), and concatenated together to create a mapping index. Metatranscriptomic reads were competitively pseudoaligned to the reference index and counts quantified with kallisto (67). 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 (47). 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 among the dominant IA-UW4 and IIC-UW6 strains, we identified sets of genes that were differentially expressed between anaerobic and aerob conditions using DESeq2 (68). 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
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: https://figshare.com/projects/Metabolic_Plasticity_of_Accumulibacter_Clades/90614. All code is available at https://github.com/elizabethmcd/R3R4.

RESULTS

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 operation 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 performed qPCR of the \(ppk1\) as described previously (29) to quantify Accumulibacter clades present in the reactor 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.

From metagenome samples collected from this bioreactor earlier, we assembled high-quality genomes of belonging to four different Accumulibacter clades (Table 1), including the dominant IA and IIC (strains IA-UW4 and IIC-UW6, respectively) clades and
the less abundant IIA and IIF clades (strains IIA-UW5 and IIF-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 (46) (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 the IIC-UW6 strain. To our knowledge, the assembled IIC-UW6 genome 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 characterized the functional diversity of different Accumulibacter clades through clustering orthologous groups of genes (COGs) (Figure 3A). We repeated a COGs analysis from Skennerton et al. (23) and Oyserman et al. (69), but included an updated set of high-quality Accumulibacter references, the 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 (strains UW1 and UW3, and UW4-7, respectively) (22, 30), we recently characterized clade IC (strain UW-LDO-IC) using genome-resolved metagenomics and metatranscriptomics (14). Overall, 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 COG presence/absence (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 in gene content 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 COG groups 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), the intersection analysis (Figure 3B) identified gene families that are available to individual strains within the same bioreactor ecosystem, highlighting the potential for functional differentiation.

**Transcriptional Profiles of Accumulibacter Strains**

Although *ppk1* locus quantification showed that the bioreactor was mostly enriched in clade IA (see 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 strain IIC-UW6 genome across the full time-series experiment, whereas strains IA-UW4 and IIA-UW5 recruited roughly the same number of reads, with 14 million and 11 million, respectively (Table 2). Markedly lower read levels mapped to the IIF-UW7 strain relative to the other three clades, and therefore we do not include the IIF-UW7 strain in our subsequent analyses. We explored the top differentially expressed genes in strain 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 (Figure 4B).

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 strain 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 strain IA-UW4. Most of these genes in strain IA-UW4 exhibit the highest expression patterns in the early and late anaerobic stages (Figure 4C). Conversely, strain IIC-UW6 contains the \(narGHIJ\) machinery for nitrate reduction, a different \(nirS\)-like nitrite reductase than in IA-UW4, and does not contain a nitrous oxide reductase. However, strain IIC-UW6 does contain the nitrite reductase \(nirBD\) for reducing nitrite to ammonium, which is uniformly expressed across the cycle (Figure 4C).

**Expression Dynamics of Genes Shared by Strains IA-UW4 and IIC-UW6**

When characterizing expression dynamics of COGs shared between strains IA-UW4 and IIC-UW6, we found very few orthologs that exceeded the 1.5-fold change cutoff for differential expression in both strains (Figure 5). 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 several 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 (70). 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 (13, 71).

Additionally, both the IA-UW4 and IIC-UW6 strains carry most of the subunits for 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 (72–74). 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 strain IIA-UW1 (13). However, the IIC-UW6 genome is missing the HoxY delta subunit of the hydrogenase moiety, and only strain IA-UW4 exhibits differential expression of the HoxH beta subunit in the anaerobic phase. Strain 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. A prior study focused on strain IIA-UW1 also demonstrated hydrogenase activity in the anaerobic phase, and then demonstrated hydrogen gas production after acetate addition (13). The hydrogen gas production was hypothesized to replenish NAD+ after glycogen degradation. Subsequent comparative genomics analysis suggested that the Hox system is an ancestral state of Accumulibacter (69). Furthermore, recent metabolic models which integrated hydrogen production under periods of excess acetate or glycogen were better able to capture the diversity of observed EBPR stoichiometry (75). In this ecosystem,
both strains may exhibit anaerobic hydrogenase activity and thus produce hydrogen, but
due to the missing subunit in the IIC-UW6 genome and lack of differential gene
expression, this is uncertain.

Interestingly, only one strain exhibited differential expression patterns of several
core ortholog subsets important for EBPR-related functions exhibit. For example, subunits
of the high-affinity phosphate transporter Pst system are only differentially expressed in
strain IIC-UW6 and not IA-UW4, as observed for strain IIC-UW6 above (see Figure 4B).
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 strain IA-UW4 and the denitrification protein NirQ is differentially expressed in
only strain IIC-UW6. Whereas the nitrogen fixation protein subunit NifU is differentially
expressed in both strains. 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. The strain IA-UW4 assembly is missing an acetate kinase,
and phosphate acetyltransferase is only differentially expressed in this strain but not in
IIC-UW6. Within strain IIC-UW6, both the acetate kinase and phosphate acetyltransferase
genes exhibit relatively low levels of expression over the time-series. Both strains contain
several acetyl-CoA synthases that are highly expressed in the anaerobic phase,
particularly upon acetate contact, but do not exhibit notable differences in 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 (76, 77).
Although a phosphate acetyltransferase is differentially expressed in strain IA-UW4, 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 Strain IA-UW4**

We lastly characterized expression dynamics of flexible gene content in strain IA-UW4 (Figure 6). Since a majority of the differentially expressed flexible genes in strain IIC-UW6 were annotated as hypothetical proteins, 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 IIC-UW6 genome, whether or not 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 strain IIC-UW6 genome but present in the IA-UW4 genome or any other genomes was inferred as a lack of functional capability of IIC-UW6 in this particular ecosystem.

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 strain IIC-UW6 contains the *nar* system for reducing nitrate to nitrite (Figure 4C). In strain 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 strain IA-UW4 contains a different *nirS* that is upregulated in the anaerobic phase. Strain IIC-UW6 does not contain both subunits of the nitric-oxide reductase *norQD*, whereas strain 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 (see Figure 4B). Additionally, strain IIC-UW6 does not contain the nitrous-oxide reductase nosZ (see Figure 4B), whereas strain IA-UW4 does, and this is differentially expressed in the anaerobic phase (Figure 6). These results suggest that strain 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 strain IIC-UW6 does contain the nirBD nitrite reductase complex for reducing nitrite to ammonium, these genes are uniformly expressed throughout the cycle (see Figure 4B).

**DISCUSSION**

In this work, we performed time-series transcriptomics to explore gene expression profiles of Accumulibacter strains belonging to clades IA and IIC, which co-existed 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 four Accumulibacter clades and used long reads to generate a single contiguous scaffold of the IIC-UW6 genome. By performing an updated clustering of orthologous groups of genes analysis, we were able to better demonstrate the similarity and uniqueness among Accumulibacter strains from this system. Unexpectedly, although clade IA dominated in abundance by ppk1 locus quantification, clade IIC was more transcriptionally active. By comparing the expression profiles of COGs shared between strains IA-UW4 and IIC-
UW6, we found that genes involved in EBPR-related pathways may differ between the strains with respect to transcription-level regulation. Notably, strain IIC-UW6 showed a strong upregulation of the high-affinity Pst phosphorus transport system in the aerobic phase. High activity of the Pst system in IIC-UW6 towards the end of the aerobic phase when phosphorus concentrations were lower combined with contrasting abundance vs. transcriptional activity patterns suggests that the IIC-UW6 strain is specialized under phosphorus-limiting conditions. Whereas lack of differential expression of the Pst system in IA-UW4 suggests this strain is more specialized under high phosphorus conditions.

Strain 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 strain IA-UW4 in nitrogen cycling under denitrifying conditions. Additionally, the IA-UW4 strain contains all subunits for encoding the Hox anaerobic hydrogenase system, which exhibits upregulated gene expression in the anaerobic phase. However, the IIC-UW6 strain is missing one of the subunits even though the other subunits were expressed (but not differentially). Production of hydrogen gas could be a differentiating role for the IA-UW4 strain, as this would impact potential interactions with diverse flanking community members in addition to replenishing NAD+ (13, 30, 78).

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, strain IIA-UW1 and strain IA-UW3 were present in the same bioreactor ecosystem and only share 85% ANI (22). In the bioreactor enrichment from this study, the dominant strains IA-UW4 and IIC-UW6 share less than 80% ANI (Figure 2B). Recently, a cutoff of >95% ANI has been
suggested to delineate distinct bacterial species or cohesive genetic units (51), which is based on observations of coverage gaps at 90-95% sequence identity when metagenomic reads were mapped back to assembled MAGs (79–82). These sequence cutoffs have also been benchmarked against homologous recombination rates and ratios of nonsynonymous to synonymous ($dN/dS$) nucleotide differences (83). 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 (84, 85). 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, 86). Because Accumulibacter has not been isolated in pure culture to date, most ecological and evolutionary inferences have been made from genome-based comparisons (23, 69) or mapping metagenomic reads back to assembled genomes with minimum percent identity cutoffs that may not
sufficiently distinguish among sub-clades (87, 88). 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’ (89). Additionally, a few genomes preliminarily classified by the Genome Taxonomy Database (GTDB) as Accumulibacter branch outside of the established nomenclature and are quite divergent based on the species and ppk1 phylogenies as well as genome-wide and 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 to 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 have been repeatedly suggested to be a niche-differentiating feature (14, 24, 29). Previously, two separate bioreactors fed with either acetate or propionate as the sole carbon source were enriched with different Accumulibacter morphotypes, which were also linked to different denitrification capabilities (25). A study with 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 IC members (29). The resident strain IC-UW-LDO also possessed the full suite of genes for denitrification and expressed them in microaerobic conditions (14). Coexisting Accumulibacter strains with differing denitrifying abilities have also been observed in a denitrifying EBPR bioreactor enriched in clade IA and clade IC (17, 28). The clade IA strain encoded genes for complete denitrification, whereas the co-occurring clade IC strain and other flanking community members did not (28). Additionally, Gao et al. performed an ancestral genome reconstruction analysis similar to Oyserman et al. 2017 (69) to specifically understand the flux of denitrification gene families (28). The nir gene cluster was identified as a core COG 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 (28). Additionally, Wang et al. subsequently observed that an internal stop codon within the open reading frame of the nosZ gene in a clade IC strain coincided with decreased transcripts mapping back (17, 28). In our study, we observed differential expression of denitrification genes in strain IA-UW4 but not strain IIC-UW6. These expression differences could explain why the two strains 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 ammonia oxidation through allylthiourea addition (and thus inhibits nitrite/nitrate production to fuel respiratory denitrification), the expression profiles of denitrification genes in strain IA-UW4 suggests this strain is equipped to have a differentiating role in a denitrifying bioreactor. We specifically observed differential
expression of nitrate reductase and nitrous-oxide reductase genes in strain IA-UW4, where homologs of these genes were not present in the strain IIC-UW6 genome and other Type II genomes. From our work and previous studies, lack of the complete genetic repertoire for denitrification among and within individual clades in addition to transcriptional-based evidence suggests that denitrification capabilities could be a niche differentiating role for specific Accumulibacter populations.

Accumulibacter’s anaerobic metabolism has also been shown to differ under varying substrate or stoichiometric conditions and possibly between the two Accumulibacter types. Recent simulations of ‘omics datasets paired with enzymatic assays hint that the routes of anaerobic metabolism for Accumulibacter may differ based on environmental conditions (75). 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 (75). Furthermore, Welles et al. found that Type II Accumulibacter can switch to partial glycogen degradation more quickly than Type I, enabling Type II to fuel VFA uptake in polyphosphate-limited conditions (90, 91). 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 between 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 also how metabolisms are distributed amongst Accumulibacter clades.
in bioreactor ecosystems that are not completely enriched in a single clade or strain and elucidate key functional differences.

CONCLUSIONS

In this study, we integrated genome-resolved metagenomics and time-series metatranscriptomics to understand gene expression patterns of co-existing Accumulibacter strains within a bioreactor ecosystem. We found evidence for denitrification gene expression in the dominating IA-UW4 strain, but higher overall transcriptional activity for the IIC-UW6 strain along with differential expression of a high-affinity phosphorus transporter. The coexistence of Accumulibacter populations exhibiting different gene expression patterns for phosphorus and nitrogen cycling and hydrogen productions suggest niche differentiating features and contribution to different metabolic programs. Detailed experiments incorporating fluorescent in situ hybridization combined with microautoradiography could be applied to confirm metabolic activities of coexisting Accumulibacter populations and how they contribute to important EBPR-related functions (92). These results also highlight the need to reevaluate the phylogenetic structure of the Accumulibacter lineage and define how potential genetic boundaries may translate to individual ecological niches that allow 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 emphasizes the need to understand the basic ecology and evolution of microorganisms underpinning biotechnological processes,
which can lead to better treatment process control and resource recovery outcomes in
the future (93).

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 (DE-SC0018409).
REFERENCES


9. Comte J, Fauteux L, Giorgio PA del. 2013. Links between metabolic plasticity and


23. Skennerton CT, Barr JJ, Slater FR, Bond PL, Tyson GW. 2015. Expanding our


30. Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C,


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


islands drive lifestyle changes in plant-associated Pseudomonas. ISME J


Reimer J, Harrington J, Rodriguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, 
Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, 
Nowaczyk N, Shebanov N, Pavlyk O, Brodkorb PA, Lee P, McGibbon RT, 
Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, 
Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, 


BMC Bioinformatics 11:119.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-


organism abundant in full-scale enhanced biological phosphorus removal plants.

Front Microbiol 7:1033.


**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. (29).

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

2A. Phylogenetic tree of ppk1 nucleotide sequences from available Accumulibacter reference genomes, 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 (98.6% completeness,
2.5% contamination, 5.18 Mbp). 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*. The pie graph 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. 1075 gene families were shared among all UW Accumulibacter clades and outgroup genomes and 277 separate gene families were found only within the UW Accumulibacter clades. 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. The IIC-UW6 genome mapped the most transcripts even though the IA-UW4 was more abundant. Reads were competitively mapped to the four assembled Accumulibacter clade genomes and normalized by transcripts per million (TPM). The sum of the
normalized counts of the three samples in the anaerobic phase and four samples in the aerobic phase were averaged, respectively, and plotted on a log\(_{10}\) scale. Numbers and percentages for adjacent to normalized bars represent the raw number of reads that are mapped to each genome. **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 (47) and KofamKOALA (48).

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 Strains IA-UW4 and IIC-UW6**

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 ±1.5 fold-change in expression between anaerobic and aerobic phases. 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 Strain IA-UW4**

Gene expression profiles of groups of accessory gene groups in strain IA-UW4 that are not present in the IIC-UW6 genome. A gene was considered differentially expressed between the anaerobic and aerobic conditions if it exhibited greater or less than $\pm 1.5$ fold-change in expression. Colored dots represent whether or not that gene contains orthologs within outgroup genomes, other Type II Accumulibacter other than the IIC-UW6 genome, within Type I Accumulibacter, or only within 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 (46).

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

Counts of raw metatranscriptomic reads mapped to each Accumulibacter genome with kallisto (67) 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 is 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 (52). A 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_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. Dechloromonas aromatica RCB
   Rhodocyclus tenuis DSM 110
   Candidatus Accumulibacter UW-LDO
   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

B. Intersection Size (Number of Orthologous Groups)

0 300 600 900

1075 887 689 399 292 277 240 164 146 144 121 115 110 109 101 98 80 70 66 67 50 47 45

Flexible gene content

Tree scale: 0.075

Outgroup

Rhodocyclus
Dechloromonas

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

Type I

Type II

Intersection Size

Type I

Type II
**A.**

![Graph showing log10 Average Expression - Transcripts per Million (TPM)]

**B.**

![Heatmap of gene expression for Strain IIC-UW6, showing log2 Fold-Change](image)

**C.**

![Diagram of oxidation state and gene expression for Strain IIC-UW6](image)
<table>
<thead>
<tr>
<th>Strain IA-UW4</th>
<th>Strain IIC-UW6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaerobic</td>
<td>Anaerobic</td>
</tr>
<tr>
<td>Anaerobic</td>
<td>Anaerobic</td>
</tr>
</tbody>
</table>

- **Cytochrome c-552**
- **Putative chemotaxis protein**
- **Low affinity K+ transporter**
- **Protein HtpX**
- **Cysteine desulfurase IscS**
- **Cysteine desulfurase NifS**
- **Transcriptional regulator IscR**
- **Coproporphyrinogen III oxidase**
- **Ammonia channel**
- **Methyltransferase**
- **Glycerol-3-phosphate dehydrogenase**
- **Phosphoribulokinase**
- **Transcriptional repressor CarH**
- **Hypothetical protein**
- **Nitrogen fixation protein NifU**
- **2Fe-2S cluster assembly protein IscU**
- **2Fe-2S ferredoxin**
- **Nitrogen regulatory protein P-II 2**
- **Hypothetical protein**
- **Hypothetical protein**
- **Ribulose bisphosphate carboxylase**
- **Flagellin**
- **Hypothetical protein**
- **Diguanylate cyclase DgcM**
- **Hypothetical protein**
- **Ribosomal protein**
- **Hypothetical protein**
- **Phosphate acetyltransferase**
- **NAD-reducing hydrogenase HoxS-B**
- **Exodeoxyribonuclease**
- **Denitrification protein NirQ**
- **Ferredoxin-NADP reductase**
- **Cytochrome c4**
- **Hypothetical protein**
- **ATP-synthase delta subunit**
- **Hypothetical protein**
- **ATP-synthase subunit c**
- **Hypothetical protein**
- **NAD-reducing hydrogenase HoxS-A**
- **Hypothetical protein**
- **NAD-reducing hydrogenase HoxS-G**
- **Bacteriohemerythrin**
- **Iron uptake protein A1**
- **Adaptive-response sensory-kinase SasA**
- **Putative redox modulator Alx**
- **Hypothetical protein**
- **Hypothetical protein**
- **Poly(3-hydroxyalkanote) polymerase PhaC**
- **Long-chain fatty-acid-CoA ligase FadD13**
- **Sensor protein QseC**
- **Phosphate transport system PhoU**
- **Phosphate transport system PstA**
- **Transcriptional regulatory protein QseB**
- **Hypothetical protein**
- **Phosphate transport system PstS**
- **Hypothetical protein**
- **Nitrite reductase**
- **Hypothetical protein**
- **Hypothetical protein**
- **Hypothetical protein**
- **Carbonic anhydrase**
- **Hypothetical protein**
- **Hypothetical protein**
- **Sulfoxide reductase MsrB**
<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>Nitrate reductase (cytochrome)</td>
<td>Hypothetical protein</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>Ferredoxin-type protein NapH</td>
<td>Nucleoid occlusion factor SlmA</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>Hypothetical protein</td>
<td>Acetylglutamate kinase</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Dethiobiotin synthase</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Hypothetical protein</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>Hypothetical protein</td>
<td>Phospholipase</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>Type IV Secretion system</td>
<td>Transcriptional regulator NtrC</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>Ammonia channel</td>
</tr>
<tr>
<td>C4-dicarboxylate transport protein</td>
<td>Glutamate racemase</td>
</tr>
<tr>
<td>Putative transposase</td>
<td>Conjugal transfer protein TraG</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>
</tbody>
</table>

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

**log2 fold-change**

2 1 0 -1

(2021) (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 February 1, 2021. ; https://doi.org/10.1101/2020.11.23.394700doi: bioRxiv preprint
<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>UW3-IA</td>
<td>UW4-IA</td>
<td>UW-LDO-IC</td>
<td>UW5-IIA</td>
<td>UW1-IIA</td>
<td>UW6-IIC</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>100.0%</td>
<td>78.3%</td>
<td>62.9%</td>
<td>40.9%</td>
<td>40.4%</td>
<td>33.3%</td>
</tr>
<tr>
<td>78.3%</td>
<td>100.0%</td>
<td>57.2%</td>
<td>34.0%</td>
<td>33.3%</td>
<td>29.2%</td>
</tr>
<tr>
<td>62.9%</td>
<td>57.2%</td>
<td>100.0%</td>
<td>40.0%</td>
<td>40.0%</td>
<td>35.7%</td>
</tr>
<tr>
<td>40.9%</td>
<td>34.0%</td>
<td>40.0%</td>
<td>100.0%</td>
<td>92.3%</td>
<td>35.7%</td>
</tr>
<tr>
<td>40.4%</td>
<td>33.3%</td>
<td>40.0%</td>
<td>92.3%</td>
<td>100.0%</td>
<td>36.0%</td>
</tr>
<tr>
<td>33.3%</td>
<td>29.2%</td>
<td>35.7%</td>
<td>35.7%</td>
<td>36.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>48.8%</td>
<td>41.5%</td>
<td>48.7%</td>
<td>52.0%</td>
<td>50.1%</td>
<td>40.9%</td>
</tr>
</tbody>
</table>
A. Clade IA

B. Clade IB

Clade IC

Clade IIA

Clade IIC

Clade IIF

Pairwise BLAST % Identity Score of ppk1

Rhodocyclus tenuis DSM 110

Tree scale: 0.01