Supplementary Information

for

Disruption in Gene Expression Cycles of Polyphosphate Accumulating Organisms is Associated with a Full-Scale Enhanced Biological Phosphorus Removal Instability Event

Jessica A. Deaver[1,2], Thomas Solon[3,4], Amy M. Grunden[2,5], and Douglas F. Call[1,2,*]

[1]Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC 27606

[2]Science and Technologies for Phosphorus Sustainability (STEPS) Center, North Carolina State University, Raleigh, NC 27606

[3]Department of Environmental Engineering and Earth Sciences, Clemson University, SC 29634

[4]Renewable Water Resources (ReWa), Greenville, SC 29607

[5]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27606

*corresponding author, dfcall@ncsu.edu

Supplementary Table Legends

Table S1: Overview of sequencing approach by sample location.

Table S2: Individual and co-assembly statistics calculated with QUAST (https://quast.sourceforge.net/). B1 = treatment train 1, B2 = treatment train 2, B3 = treatment train 3, D1 = 11/20/2023, D2 = 11/29/2023, D3 = 12/6/2023, D4 = 2/23/2024.

Table S3: Assembly statistics of metagenome-assembled genomes (MAGs) recovered in this study. MAGs presented are the species-level representatives dereplicated at a 95% average nucleotide identity (ANI) and annotated with Bakta.[1]

Table S4: Average Nucleotide Identity (ANI) between *Ca. Accumulibacter* MAGs recovered in this study and reference *Ca. Accumulibacter* MAGs from Petriglieri et al.[2] and UW 21 from Stewart et al.[3]

Table S5: FastANI comparison of *Azonexus* and family Azonexaceae MAGs recovered in this study to GTDB *Azonexus* species representatives.

Supplementary Tables

Table S1: Overview of sequencing approach by sample location.

Sample Location	Sample Location Code	Sequencing Approach
Beginning Anaerobic Zone	A	RNA Sequencing – Illumina
End Anaerobic Zone	B	RNA Sequencing – Illumina
Beginning Anoxic Zone	C	RNA Sequencing – Illumina
End Anoxic Zone	D	RNA Sequencing – Illumina
Beginning Aerobic Zone	E	RNA Sequencing – Illumina
End Aerobic Zone	G	RNA Sequencing – Illumina DNA Sequencing – Illumina & Oxford Nanopore

Tables S2-S5 are included as separate Excel files

Supplementary Figures

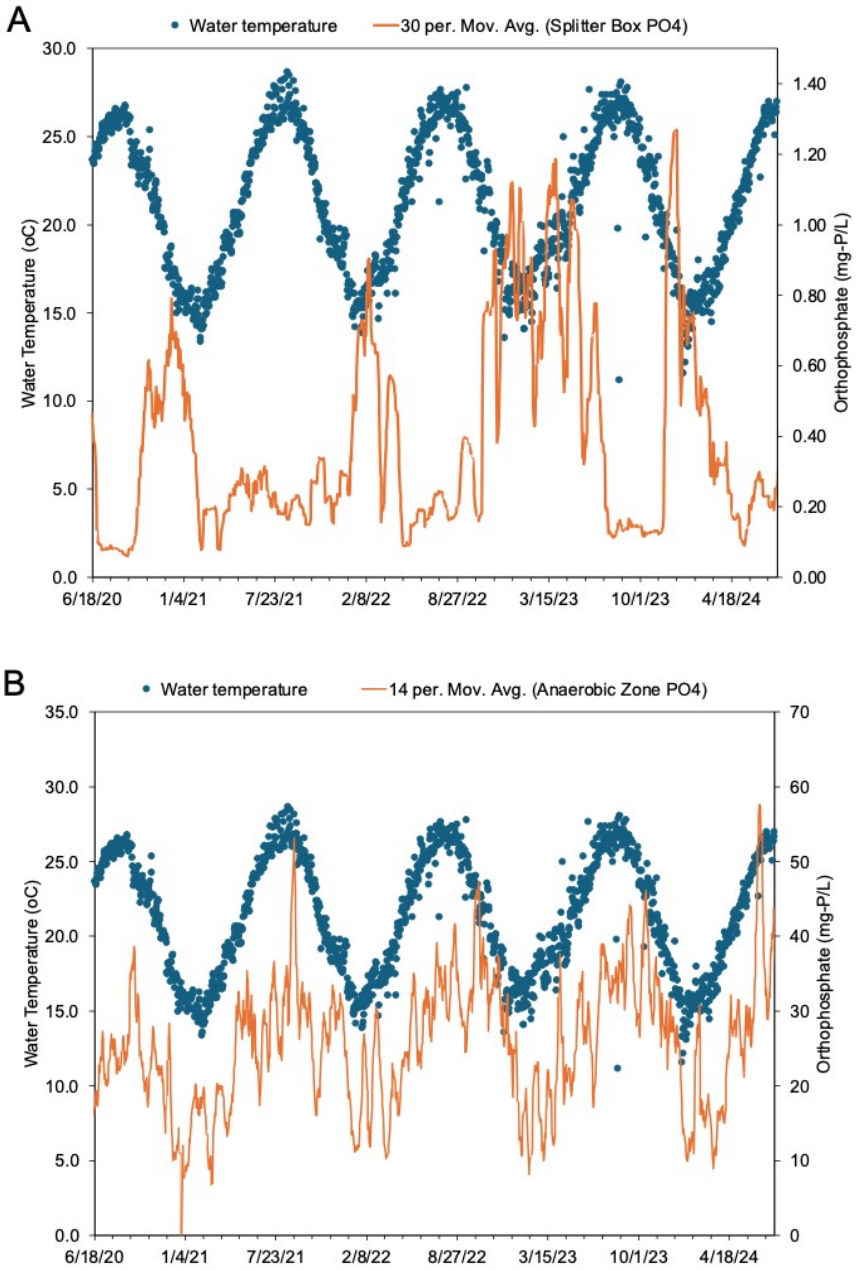


Figure S1: Approximately 3-year trends in (A) water temperature versus splitter box orthophosphate (shown as a 30-day rolling average) and (B) water temperature versus anaerobic orthophosphate measured at location B, the end of the anaerobic zone (shown as a 14-day rolling average).

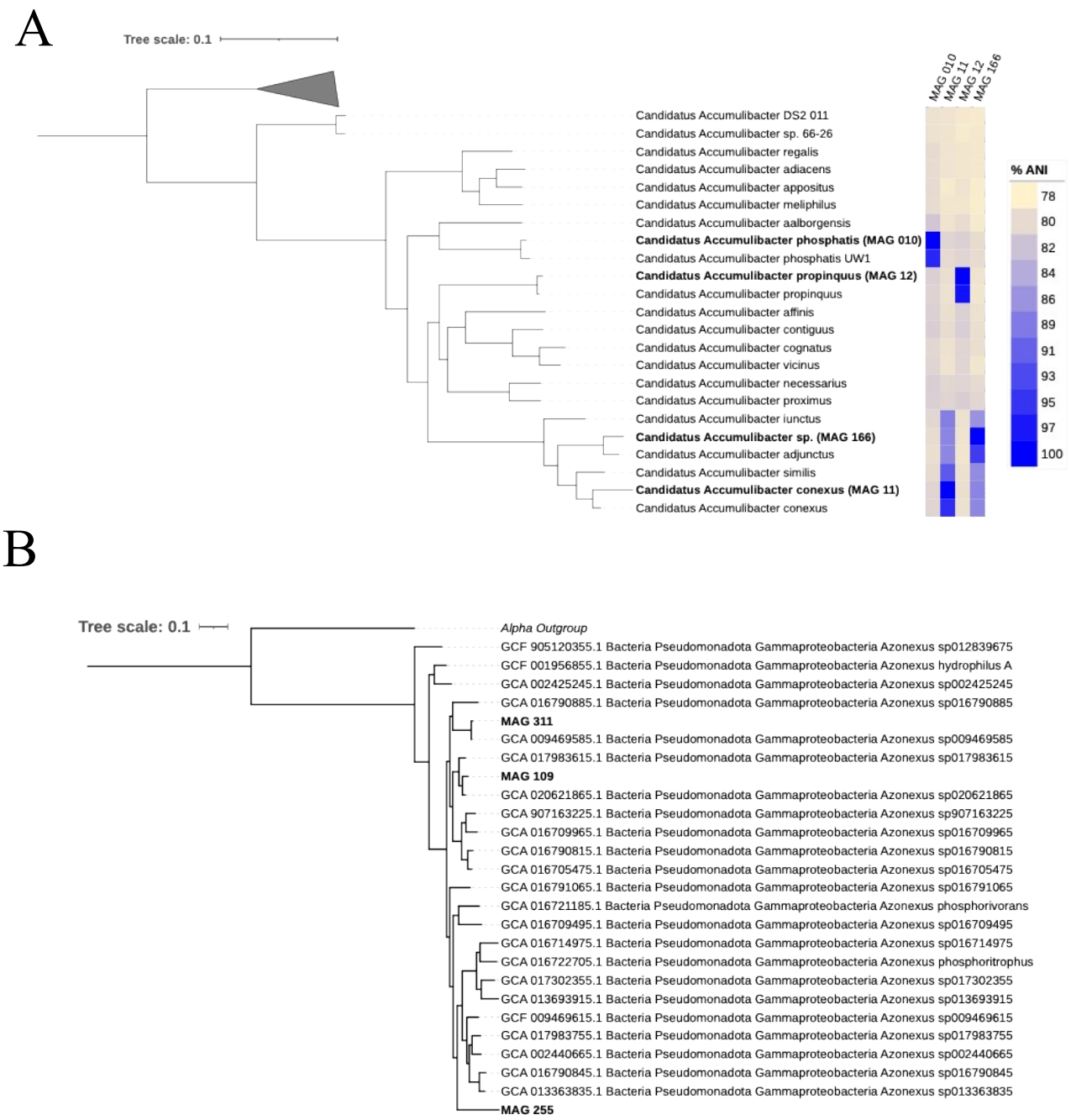


Figure S2. (A) Phylogenomic tree and pairwise ANI comparisons between *Ca. Accumulibacter* MAGs recovered in this study and *Ca. Accumulibacter* reference MAGs identified by Petriglieri et al.[2] MAGs assembled in this study are **bolded**. Maximum-likelihood tree created with GToTree based on Proteobacteria HMM profiles. Outgroup

consists of two *Azonexus* reference genomes. NCBI accessions IDs for reference genomes and % ANI values calculated with FastANI are reported in **Table S3**. (B) Phylogenomic tree containing *Azonexus* and Azonexaceae MAGs recovered in this study and GTDB *Azonexus* species representatives derived from activated sludge samples and/or that share > 90% ANI with a MAG from this study. Maximum-likelihood tree created with GToTree based on Proteobacteria HMM profiles. Outgroup consists of an alpha-proteobacterium genome. NCBI accessions IDs for reference genomes and % ANI values calculated with FastANI are reported in **Table S4.**

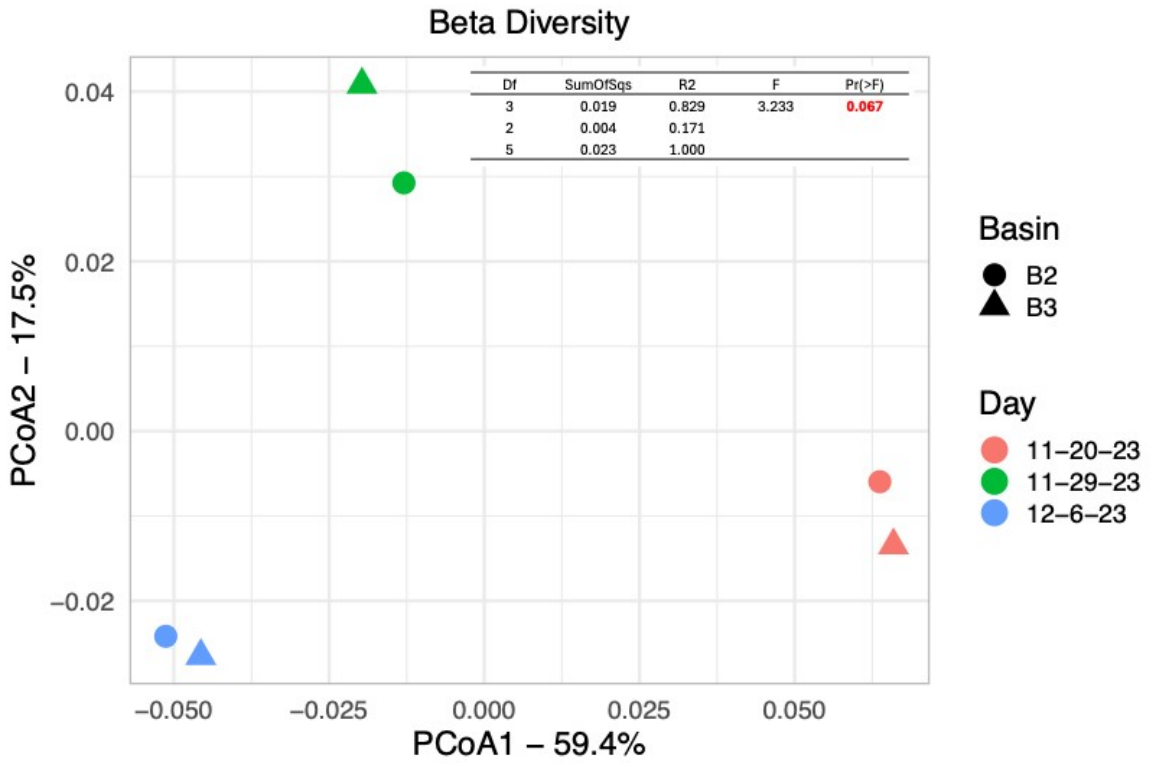


Figure S3. Genus-level β-diversity analysis based on Bray-Curtis distances and PERMANOVA statistical testing to determine significance of Bray-Curtis distances between communities.

A


Glycerol−3−phosphate regulon repressor
Multiphosphoryl transfer protein
PEP−utilising enzyme C−terminal domain−containing protein
PTS fructose−like transporter subunit IIB
Alpha/beta fold hydrolase
poly−beta−hydroxybutyrate polymerase N−terminal domain−containing protein
30S ribosomal protein S21
Preprotein translocase subunit SecD
bifunctional phosphoserine phosphatase/homoserine phosphotransferase ThrH
DUF2889 domain−containing protein
ornithine carbamoyltransferase
Phasin family protein
putative polyhydroxyalkanoate (PHA)−binding protein, phasin family
heme d1 biosynthesis radical SAM protein NirJ
Poly(R)−hydroxyalkanoic acid synthase, class III, PhaC subunit
Cytochrome b subunit of the bc complex
Cytochrome c553
glycine cleavage system aminomethyltransferase GcvT
F0F1 ATP synthase subunit A
F0F1 ATP synthase subunit B
oxygen−independent coproporphyrinogen III oxidase
cytochrome c oxidase accessory protein CcoG
Nitrate/proton symporter
Nitrate/nitrite antiporter
polyhydroxyalkanoate synthesis repressor PhaR
DUF1461 domain−containing protein
Nitric oxide reductase large subunit
Sulfate adenylyltransferase subunit 1, EFTu−like GTPase family
adenylate kinase
2−methylcitrate synthase
cytochrome c oxidase subunit II
Secreted protein
SIR2−like domain−containing protein
HTH cro/C1−type domain−containing protein
Lipoprotein SmpA/OmlA domain−containing protein
CshA−type fibril repeat protein
UDP−N−acetylmuramyl pentapeptide phosphotransferase/UDP−N−acetylglucosamine−1−phosphate transferase
L−asparaginase/archaeal Glu−tRNAGln amidotransferase subunit D
Peptidoglycan binding−like domain−containing protein
bifunctional 2−methylcitrate dehydratase/aconitate hydratase
OmpA−like domain−containing protein
von Willebrand factor type A
Aldehyde, CO or xanthine dehydrogenase, Mo−binding subunit
glutamate−5−semialdehyde dehydrogenase
Tetratricopeptide repeat−like domain−containing protein
Metal−dependent amidase/aminoacylase/carboxypeptidase
Pvc16 N−terminal domain−containing protein
aconitate hydratase AcnA
Integral membrane protein YbhL, putative Ca2+ regulator, Bax inhibitor (BI−1)/TMBIM family
Phage tail protein
Phage tail sheath protein FI
thiol peroxidase
Transcriptional regulator, contains XRE−family HTH domain
CHRD domain−containing protein
Esterase EstA
PEP−CTERM sorting domain−containing protein
ParB−like nuclease
PEP−CTERM protein−sorting domain−containing protein
Peptidase C14 caspase catalytic subunit p20
Catalase
fused isobutyryl−CoA mutase/GTPase IcmF
DNA−binding domain−containing protein
arsenosugar biosynthesis radical SAM (seleno)protein ArsS
UPF0276 protein
Dihydrolipoamide dehydrogenase (E3) component of pyruvate/2−oxoglutarate dehydrogenase complex or glutathione oxidoreductase
Flagellin
hydrogenase 2 operon protein HybA
Cysteine desulfurase/Cysteine sulfinate desulfinase IscS or related enzyme, NifS family
Fe−S cluster assembly transcriptional regulator IscR
IscS subfamily cysteine desulfurase
Integral membrane protein
Peptide methionine sulfoxide reductase MsrB
Ni,Fe−hydrogenase I large subunit
Penicillin−binding protein
acetate−−CoA ligase
SMODS and SLOG−associating 2TM effector domain−containing protein
Alpha−amylase
DUF485 domain−containing protein
Exonuclease SbcC
Ig−like domain−containing protein
Cysteine protease, C1A family
Ferritin
ANA 11−29
ANX 11−29
AER 11−29
ANA 12−6
ANX 12−6
AER 12−6
2
1
0
−1
−2

B

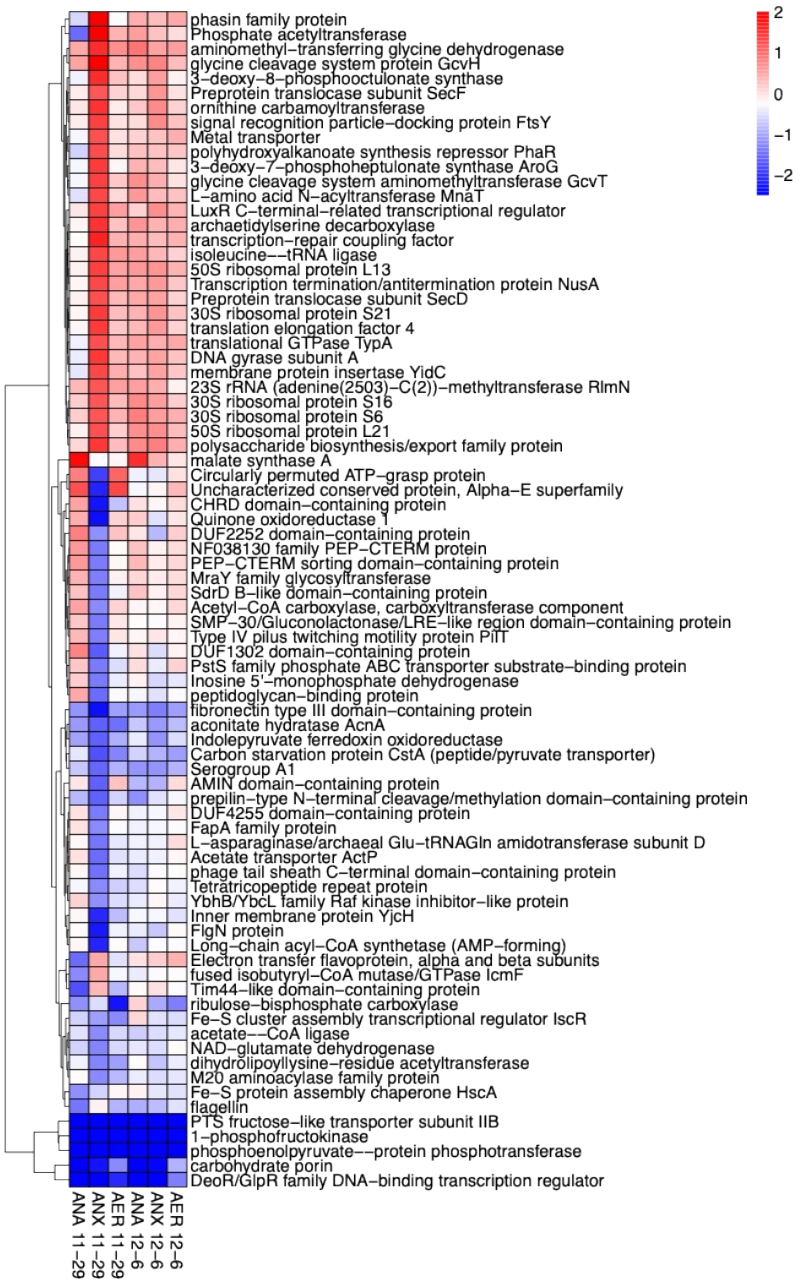


Figure S4. Differential gene expression of (A) *Ca. Accumulibacter phosphatis* and (B) *Ca. Accumulibacter propinquus* between each redox zone on 11/29/23 (unstable EBPR) and 12/6/23 to the equivalent zone on day 11/20/23. Filtered for genes with a base mean > 25 and adjusted p-value < 0.05 for at least one comparison. Heatmap colors represent $\log_2$(fold change) values.

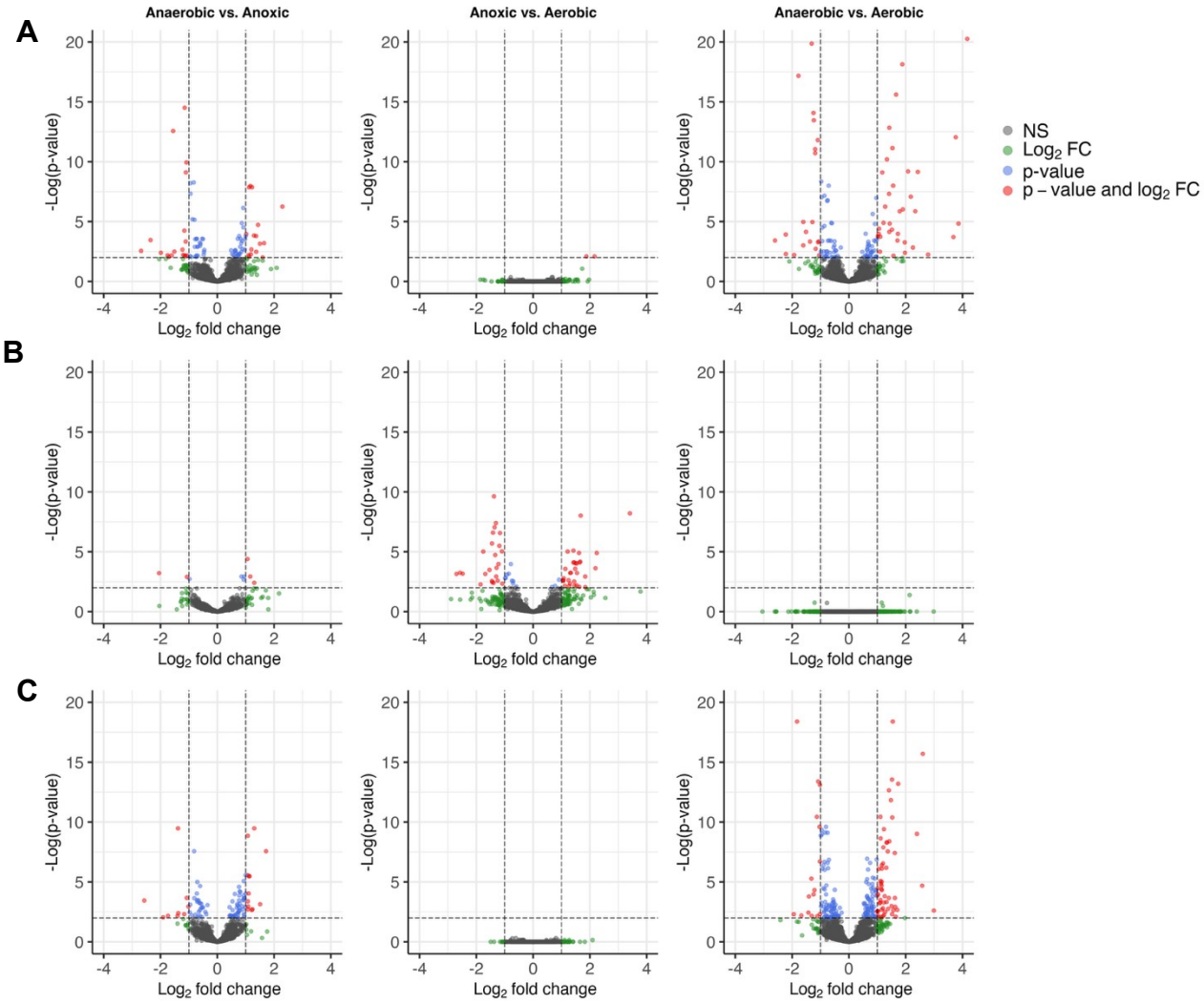


Figure S5. Differential gene expression patterns for *Ca. Accumulibacter propinquus* on (A) 11/20/23, (B) 11/29/23 (unstable EBPR), and (C) 12/6/23 between the anaerobic vs. anoxic zones (left column), anoxic vs. aerobic zones (center column), and anaerobic vs. aerobic zones (right column). The horizontal dashed lines represent an adjusted p-value equal to 0.01 and the vertical dashed lines represent a $\log_2$ fold change of +/- 1.

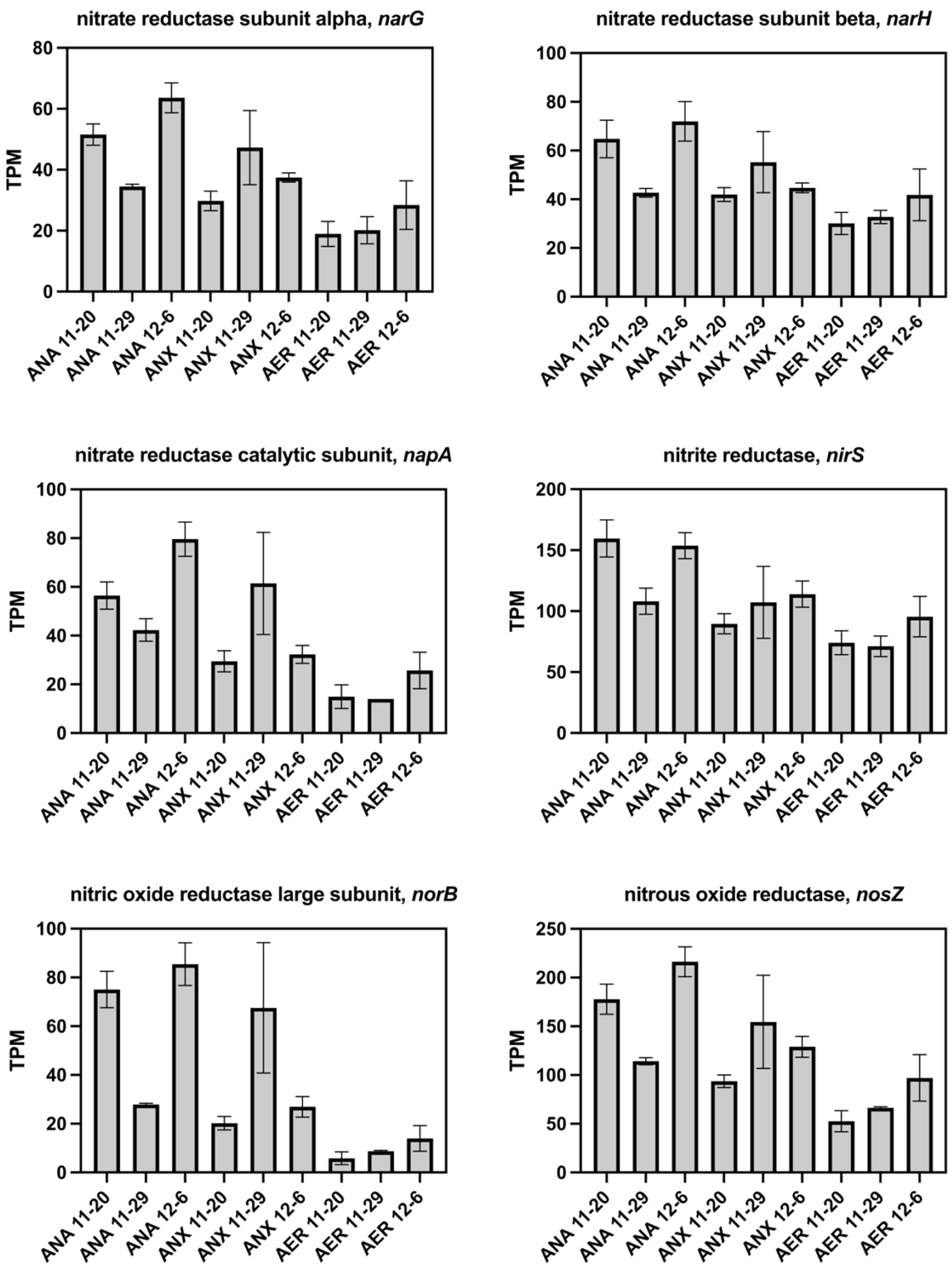
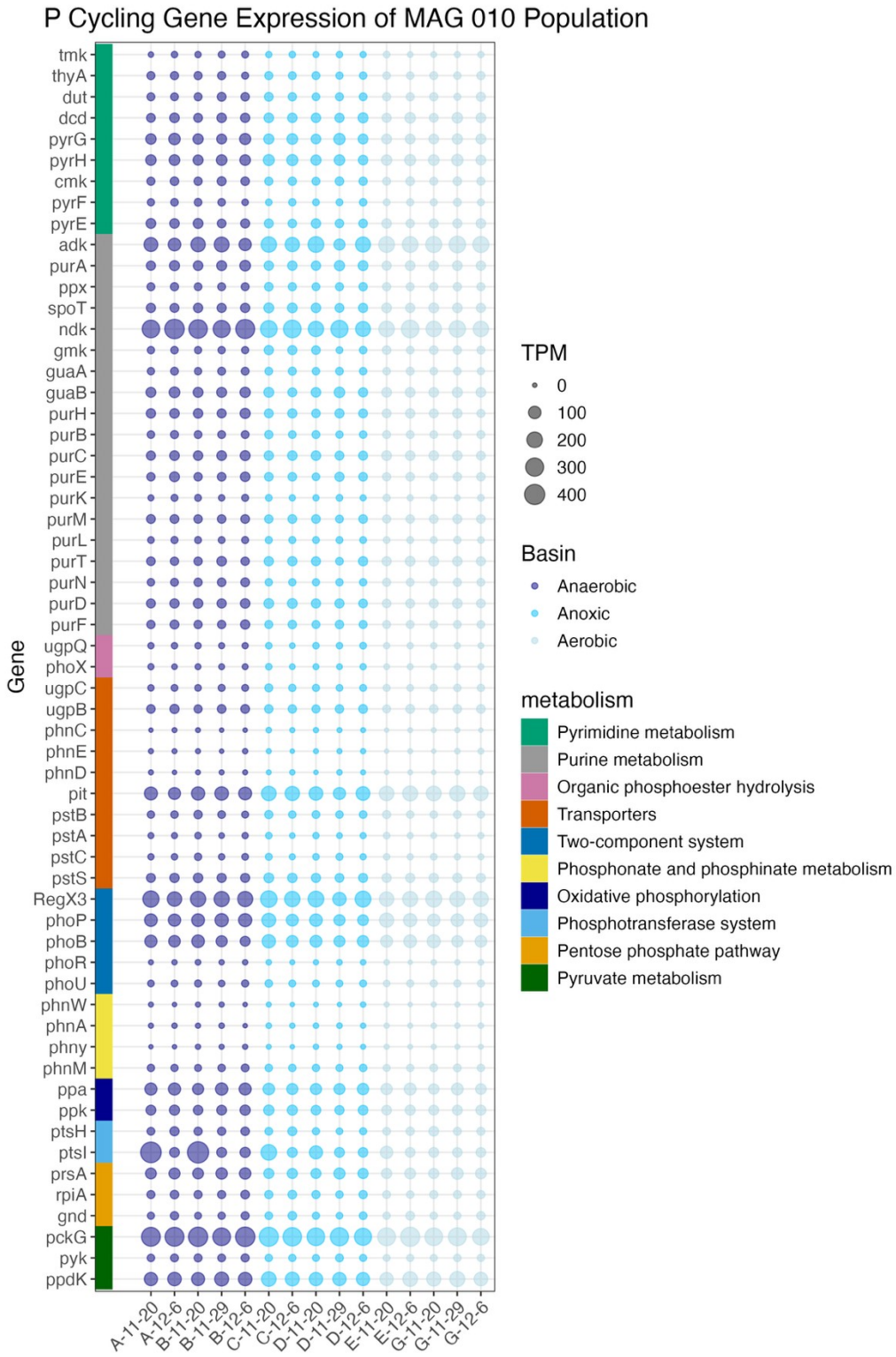


Figure S6. *Ca. Accumulibacter phosphatis* denitrification genes identified as differentially expressed in DESeq2 analyses reported in transcripts per million (TPM).

Figure S7: Differentially expressed genes between zones for *Ca. Accumulibacter propinquus*. Genes with a positive $\log_2$(Fold Change) were upregulated in the anaerobic zone (versus the aerobic zone) on 11/20/23 and 12/6/23 and in the anoxic zone (versus the aerobic zone) on 11/29/23 (unstable EBPR). Genes with a negative $\log_2$(Fold Change) were upregulated in the aerobic zone on all days versus either the anaerobic or anoxic zone. Base Mean is the average of normalized counts across all samples. ANA = Anaerobic zone, ANX = Anoxic zone, AER = Aerobic zone.

A


P Cycling Gene Expression of MAG 010 Population
Gene
tmk
thyA
dut
dcd
pyrG
pyrH
cmk
pyrF
pyrE
adk
purA
ppx
spoT
ndk
gmk
guaA
guaB
purH
purB
purC
purE
purK
purM
purL
purT
purN
purD
purF
ugpQ
phoX
ugpC
ugpB
phnC
phnE
phnD
pit
pstB
pstA
pstC
pstS
RegX3
phoP
phoB
phoR
phoU
phnW
phnA
phny
phnM
ppa
ppk
ptsH
ptsI
prsA
rpiA
gnd
pckG
pyk
ppdK
A-11-20
A-12-6
B-11-20
B-11-29
B-12-6
C-11-20
C-12-6
D-11-20
D-11-29
D-12-6
E-11-20
E-12-6
G-11-20
G-11-29
G-12-6
TPM
0
100
200
300
400
Basin
Anaerobic
Anoxic
Aerobic
metabolism
Pyrimidine metabolism
Purine metabolism
Organic phosphoester hydrolysis
Transporters
Two-component system
Phosphonate and phosphinate metabolism
Oxidative phosphorylation
Phosphotransferase system
Pentose phosphate pathway
Pyruvate metabolism

B

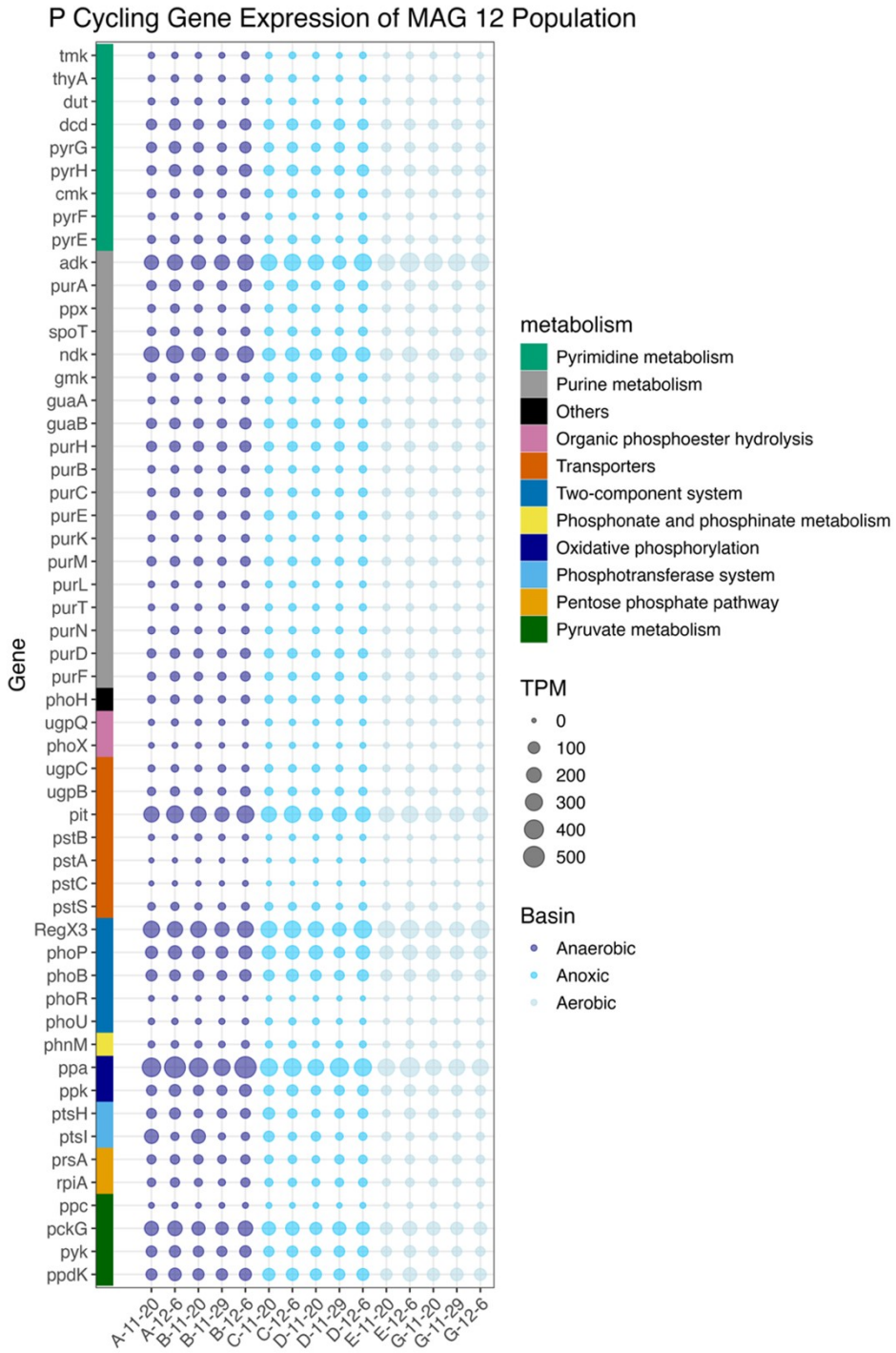


Figure S8. P Cycling gene expression in Transcripts per Million (TPM) on each sampling day (11/20/23, 11/29/23, or 12/6/23) and at each location (A-G). P cycling genes identified using the PCycDB for (A) *Ca. Accumulibacter phosphatis* and (B) *Ca. Accumulibacter propinquus*.

Supplementary Methods

Sample collection. Fresh activated sludge (AS) samples were collected and preserved for DNA and RNA sequencing. Each ~500 mL sample was collected at a location that was well-mixed, avoided foaming, and ~30 cm below the surface. A 3-mL aliquot of well-mixed AS was immediately added to a 15-mL falcon tube containing 3 mL of 2x DNA/RNA Shield Solution (Zymo Research, Irvine, CA, USA) and inverted to mix.

Chemical analyses. Orthophosphate in 0.45 µm filtered samples was either measured onsite by WRRF operators or at NC State. Onsite, operators used Hach Method 8048 for determination of reactive phosphorus (Hach Company, Loveland, CO, USA). Orthophosphate measurements at NC State used an in-house ascorbic acid assay adapted from Standard Method 4500-P E that provides for an extended range and smaller sample volumes.[5] Absorbance was measured at 880 nm using a Tecan Infinite 200 microplate reader (Tecan Life Sciences, Männedorf, Switzerland). All other measurements (nitrate, total P, BOD, etc.) were performed by operators onsite according to standard methods for lab-certified influent and effluent samples or using Hach kits for process control samples collected within the treatment train.

Sample processing and nucleic acid extractions. DNA sequencing was performed on samples collected at the end of the aerobic zones (location G in **Figure 1, Table S1**). DNA was extracted using the FastDNA Spin Kit (MP Biomedicals, Santa Ana, CA, USA) with the following modifications. For lysis, 480 µL PBS, 120 µL MT Buffer, and 500 µL of AS in DNA/RNA Shield were added to the Lysis E tube. Bead beating was performed with

a FastPrep-24 instrument for four intervals of 40 seconds at 6 m/s and cooled on ice for 2 minutes between each interval. These parameters were chosen based on the MiDAS Field Guide protocols[6] and Albertsen et al.[7] DNA was eluted in 100 µL DES. AmPure XP beads were used to clean the extracted DNA and remove residual contamination from the extraction kit following the protocol for AmPure XP PCR Purification (Beckman Coulter Life Sciences, Indianapolis, Indiana, USA). DNA was eluted in 50 µL of 10 mM Tris-HCl at pH 8.0. DNA was stored in multiple aliquots at −20°C.

RNA sequencing was performed on samples collected throughout the A2O process (**Table S1**). RNA was extracted using the ZymoBiomics RNA miniprep kit (Zymo Research) following the manufacturer's protocols with the following modifications. Samples were lysed with a FastPrep-24 instrument and cooled on ice for one minute between bead beating intervals, the DNase treatment step was included, and an extra drying step was also added after the last addition of RNA Wash Buffer. An RNA Clean & Concentrator-5 kit was used to further concentrate the extracted RNA (Zymo Research). RNA was stored in multiple aliquots at -80°C.

DNA and RNA purity was assessed using a Nanodrop Spectrophotometer ND-1000 (ThermoFisher Scientific, Waltham, MA, USA). DNA and RNA quantity was assessed using dsDNA BR assays and RNA BR assays with a Qubit 3.0 Fluorometer (ThermoFisher Scientific). DNA and RNA integrity were evaluated using gDNA and RNA ScreenTape Analyses on an Agilent 4150 TapeStation (Agilent Technologies, Inc., Santa Clara, CA, USA).

Metagenomic assembly, binning, and annotation. Computing resources available through the NC State University High Performance Computing Services Core Facility were used

to perform the following computational analyses. Demultiplexed short reads received from Admera were quality checked with FastQC (v0.12.1).[8] Reads contained no adapter sequences and < 0.1% ambiguous bases in any position, therefore quality trimming and filtering were deemed unnecessary. Raw long read sequences were basecalled with Oxford Nanopore's basecaller Dorado using the super accurate (sup) model version 4.3.0 (https://github.com/nanoporetech/dorado). Reads were quality checked with Nanoplot (v1.42.0) and filtered using Filtlong (v0.2.1) to remove reads with fewer than 1500 base pairs (bp) and a Q < 7 (https://github.com/rrwick/Filtlong).[9] The long reads were polished with short reads using Ratatosk (v0.9.0).[10] Ratatosk correct was run on long reads with the corresponding paired end short reads with the flags −Q 90 to account for use of the ONT R10 chemistries and −G to gzip output files. Polished long reads were then assembled using Flye (v.2.9.4) with the --meta and --nano-HQ options enabled.[11] Each sample was individually assembled, samples collected on the same day were co-assembled, and all samples were co-assembled for a total of 13 assemblies. Assembly statistics were checked using QUAST (v.5.2.0) (**Table S2**).[12]

Each assembly was binned using both maxbin2 (v2.2.7) and metabat2 (v2.15.2) accounting for differential coverage of all eight samples.[13,14] To prepare for binning, long and short reads were mapped to each assembly using minimap2 (v2.28) and the output alignment files were merged, sorted, and indexed using samtools (v1.20).[15,16] Metabat2 was run with default parameters. The contig name and coverage columns were extracted from the metabat2 depth text to create the correctly formatted abundance text for maxbin2, and maxbin2 was run with a minimum contig length parameter of 2500 bp. Two methods were employed for dereplication. Bins derived from the same assembly using

either binning tool were dereplicated with DasTool (v1.1.7) running default parameters to yield an optimized, non-redundant bin set per assembly.[17] Bins were further dereplicated across all assemblies using dRep (v3.5.0) at a 95% average nucleotide identity (ANI) threshold (-sa 0.95) to select the best quality representative bins at approximately the species-level.[18] MAGs with short reads mapped were visualized with Anvi'o (v8) to manually check for even coverage within samples.[19] Genomes were preliminarily classified using the GTDB-Tk classify workflow (v2.4.0) using the classify workflow, quality checked with CheckM (v1.1.6) and annotated with Bakta (v1.9.4) using the full Bakta database (v5.1).[1,20,21]

Phylogenomic analyses and relative abundance. Reference genomes from Petriglieri et al.[2] were downloaded using NCBI datasets and phylogenetic trees created with GToTree (v1.8.8) using the Proteobacteria HMM profiles.[22] Given NCBI accession numbers or genome FASTA files, GToTree identifies coding regions with Prodigal and single copy genes with HMMER, aligns and trims the sequences with Muscle and Trimal, and then uses FastTree 2 to infer approximately-maximum-likelihood trees from sequence alignments.[23–29] Trees were visualized and edited using iTOL (v7).[30] FastANI (v1.34) was used to perform pairwise average nucleotide identity (ANI) comparisons between reference genomes and MAGs assembled during this study.[31]

Taxonomic profiles of individual metagenomes and MAG recovery were assessed with SingleM (v0.18.3). SingleM is a tool used to profile metagenomes by analyzing reads that cover highly conversed regions of single copy marker genes.[32] SingleM is designed to work with short reads (Illumina sequencing); raw metagenomes were profiled with SingleM by first running 'singlem pipe' using default parameters, then summarizing

microbial community relative abundance at the genus and species levels with 'singlem summarise'. β-diversity analyses were performed using the R package vegan.[33] Bray Curtis distances were calculated and a PERMANOVA test was used to determine statistical significance.

Metagenome-enabled transcriptomics. RNA reads were quality checked with FastQC. Adapter and poly-G sequences were removed using fastp (v0.23.4) with the −g flag to remove poly-G sequences and the sequencing library adapters specified.[34] SortMeRNA (v4.3.7) was used to filter out remaining rRNA sequences.[35] 30-45 million paired end reads remained after removing rRNA sequences. When possible, MAGs recovered from this study were used as references for transcript quantification. For *Ca. Accumulibacter* species identified as present in the metagenomes but not recovered during MAG assembly, species representative (as defined by Petriglieri et al.[2]) MAGs available from NCBI were used as references. The reference MAGs obtained from NCBI were also annotated with Bakta. The coding regions predicted with prodigal during MAG annotation with Bakta were concatenated together to create a mapping index of *Ca. Accumulibacter* MAGs. Quality processed metatranscriptomic reads from all samples were competitively pseudoaligned to the mapping index and quantified with kallisto (v0.51.0).[36] Kallisto output includes transcript counts normalized for gene length and sequencing depth using the transcripts per million (TPM) method. PCycDB, a database designed to facilitate the identification and analysis of phosphorus cycling genes, was used to evaluated phosphorus cycling related gene expression.[4] DESeq2 was used for statistical analyses of differentially expressed genes.[37] Transcript abundance estimates from kallisto were imported into R using tximport and then summarized to gene-level count matrices.[38]

Multiple contrasts were performed between groups defined as the zone and day (i.e., “Anaerobic 11/20/23”, “Anoxic 12/6/23”, etc.) and the results reported here. As a note, multiple contrasts were performed between location groups defined as the exact sample location and day (i.e., “Location B 11/20/23”, “Location D 12/6/23”, etc.) to examine how considering precise location versus multiple sample locations within the same zone impact results. Results were similar, and therefore we present the data from zone/day groups here. R scripts used for analyses are available at *https://github.com/jadeaver/bioP_MT*.

Supplementary Notes

Supplementary Note 1 - Phosphorus cycling gene expression extended discussion.

To further explore phosphorus gene expression patterns beyond DESeq2 statistical analyses, we used PCycDB, a database designed to facilitate the identification and analysis of phosphorus cycling genes, to more broadly evaluate changes related to phosphorus (**Figure S9**).[4] Most phosphorus cycling genes showed only minor changes both among the redox zones and between sampling days. For both MAGs, expression of the high affinity phosphorus transport system genes (*pstABCS*) varied little across zones and days. Our results contrast with prior lab-based studies showing cyclical expression of phosphorus genes in *Ca. Accumulibacter* across anaerobic and aerobic zones.[39,40] McDaniel et al.[40] showed that *Ca. Accumulibacter* strain IIC-UW6 differentially expressed *pstABCS* aerobically, and strain IA-UW4 increased *pstABCS* expression at the anaerobic/aerobic transition.[39] One explanation for why our results differ from these prior studies is that our MAGs belong to different clades (IIA and IIB rather than IA and IIC). Different strains may have different expression patterns. Differences between environmental conditions in full-scale systems and lab-scale SBRs may also play a role. Lab-scale systems have better defined environmental conditions and operational parameters, which may elicit clearer responses to environmental cues, like phosphorus concentrations.

The low affinity inorganic phosphorus transport gene (*pit*), which is suggested to contribute to PAO's unique ability to hyperaccumulate phosphorus, was also expressed consistently across redox zones and days and at higher expression levels than *pstABCS*.[41] Notably, *pit* expression by MAG 010 and MAG 12 was 141$\pm$31 transcripts per

million (TPM) and 212±42 TPM, respectively, on average across all redox zones and days. (**Figure S8**). Comparatively, the highest average expression of any *pstABCS* subunit in either MAG was only 35±4 TPM on average across all redox zones and days. The *pit* gene is thought to be constitutively expressed rather than under fine-tuned control, such as how the *pst* transporter is controlled by the *pho* regulon in response to environmental cues like low phosphorus concentrations.[42] Xie et al.[43] suggested that not only is the *pit* transporter a key defining feature of *Ca. Accumulibacter*, but also that *pho* dysregulation may have contributed to the emergence of the phosphorus hyperaccumulation phenotype. Targeting regulation of *pit* expression via gene editing approaches could be one strategy to control phosphorus uptake by PAOs to levels exceeding the current limits of phosphorus hyperaccumulation. In summary, there were no significant changes in expression of phosphorus cycling genes between the stable and unstable periods that could explain why unstable phosphorus removal occurred.

References

1. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genomics*. 2021;7(000685). doi:10.1099/mgen.0.000685

2. Petriglieri F, Singleton CM, Kondrotaite Z, et al. Reevaluation of the Phylogenetic Diversity and Global Distribution of the Genus "*Candidatus* Accumulibacter." *mSystems*. 2022;7(3):e00016-22. doi:10.1128/msystems.00016-22

3. Stewart RD, Myers KS, Amstadt C, Seib M, McMahon KD, Noguera DR. Refinement of the "*Candidatus* Accumulibacter" genus based on metagenomic analysis of biological nutrient removal (BNR) pilot-scale plants operated with reduced aeration. *mSystems*. 2024;9(3):e01188-23. doi:10.1128/msystems.01188-23

4. Zeng J, Tu Q, Yu X, et al. PCycDB: a comprehensive and accurate database for fast analysis of phosphorus cycling genes. *Microbiome*. 2022;10(1):101. doi:10.1186/s40168-022-01292-1

5. Ding H, Good C, Call DF. Ascorbic Acid Method for Ortho-P Measurement. *protocols.io*. Published online August 4, 2025. doi:dx.doi.org/10.17504/protocols.io.dm6gpzkr8lzp/v1

6. Dueholm MKD, Nierychlo M, Andersen KS, et al. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat Commun*. 2022;13(1908). doi:10.1038/s41467-022-29438-7

7. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to Basics – The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLOS ONE*. 2015;10(7):e0132783. doi:10.1371/journal.pone.0132783

8. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

9. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*. 2023;39(5):btad311. doi:10.1093/bioinformatics/btad311

10. Holley G, Beyter D, Ingimundardottir H, et al. Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol*. 2021;22(28). doi:10.1186/s13059-020-02244-4

11. Kolmogorov M, Bickhart DM, Behsaz B, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods*. 2020;17(11):1103-1110. doi:10.1038/s41592-020-00971-x

12. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086

13. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32(4):605-607. doi:10.1093/bioinformatics/btv638

14. Kang DD, Li F, Kirton E, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7(e7359). doi:10.7717/peerj.7359

15. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, ed. *Bioinformatics*. 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191

16. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

17. Sieber CMK, Probst AJ, Sharrar A, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018;3(7):836-843. doi:10.1038/s41564-018-0171-1

18. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017;11(12):2864-2868. doi:10.1038/ismej.2017.126

19. Eren AM, Esen OC, Quince C, et al. Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ*. 2015;2015(10):1-29. doi:10.7717/peerj.1319

20. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043-1055. doi:10.1101/gr.186072.114

21. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Borgwardt K, ed. *Bioinformatics*. 2022;38(23):5315-5316. doi:10.1093/bioinformatics/btac672

22. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*. 2019;35(20):4162-4164. doi:10.1093/bioinformatics/btz188

23. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, ed. *PLoS Comput Biol*. 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195

24. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun*. 2022;13(1):6968. doi:10.1038/s41467-022-34630-w

25. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972-1973. doi:10.1093/bioinformatics/btp348

26. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11(1):119. doi:10.1186/1471-2105-11-119

27. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*. 2020;38(9):1079-1086. doi:10.1038/s41587-020-0501-8

28. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*. 2010;5(3):e9490-e9490. doi:10.1371/journal.pone.0009490

29. Tange O. *Gnu Parallel 2018*. Zenodo; 2018. doi:10.5281/ZENODO.1146014

30. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49(W1):W293-W296. doi:10.1093/nar/gkab301

31. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114. doi:10.1038/s41467-018-07641-9

32. Woodcroft BJ, Aroney STN, Zhao R, et al. Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat Biotechnol*. Published online June 12, 2025. doi:10.1038/s41587-025-02738-1

33. Oksanen J, Blanchet FG, Friendly M, et al. vegan: Community Ecology Package. Published online 2019. https://cran.r-project.org/package=vegan

34. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

35. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*. 2012;28(24):3211-3217. doi:10.1093/bioinformatics/bts611

36. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(550). doi:10.1186/s13059-014-0550-8

38. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000 Res*. 2016;4(1521). doi:doi:10.12688/f1000research.7563

39. McDaniel EA, Van Steenbrugge JJM, Noguera DR, et al. TbasCO: trait-based comparative ‘omics identifies ecosystem-level and niche-differentiating adaptations of an engineered microbiome. *ISME Commun*. 2022;2(1). doi:10.1038/s43705-022-00189-2

40. McDaniel EA, Moya-Flores F, Beach NK, et al. Metabolic Differentiation of Co-occurring Accumulibacter Clades Revealed through Genome-Resolved Metatranscriptomics. *mSystems*. 2021;6(e00474-21). doi:10 .1128/mSystems.00474-21

41. McIlroy SJ, Albertsen M, Andresen EK, et al. ‘*Candidatus* Competibacter’-lineage genomes retrieved from metagenomes reveal functional metabolic diversity. *ISME J*. 2014;8(3):613-624. doi:10.1038/ismej.2013.162

42. Santos-Beneit F. The Pho regulon: a huge regulatory network in bacteria. *Front Microbiol*. 2015;6. doi:10.3389/fmicb.2015.00402

43. Xie X, Deng X, Chen L, et al. Integrated genomics provides insights into the evolution of the polyphosphate accumulation trait of Ca. Accumulibacter. *Environ Sci Ecotechnology*. 2024;20:100353. doi:10.1016/j.ese.2023.100353