Journal Pre-proof

Integrated genomics provides insights into the evolution of the polyphosphate accumulation trait of *Ca.* Accumulibacter

Xiaojing Xie, Xuhan Deng, Liping Chen, Jing Yuan, Hang Chen, Chaohai Wei, Xianghui Liu, Stefan Wuertz, Guanglei Qiu

PII: S2666-4984(23)00118-7

DOI: https://doi.org/10.1016/j.ese.2023.100353

Reference: ESE 100353

To appear in: *Environmental Science and Ecotechnology*

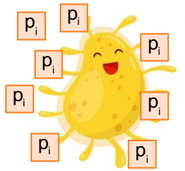
Received Date: 14 March 2023

Revised Date: 18 November 2023

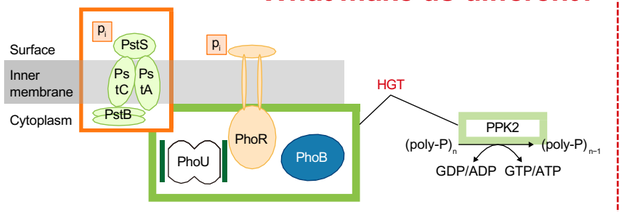
Accepted Date: 23 November 2023

Please cite this article as: X. Xie, X. Deng, L. Chen, J. Yuan, H. Chen, C. Wei, X. Liu, S. Wuertz, G. Qiu, Integrated genomics provides insights into the evolution of the polyphosphate accumulation trait of *Ca.* Accumulibacter, *Environmental Science and Ecotechnology* (2024), doi: https://doi.org/10.1016/j.ese.2023.100353.

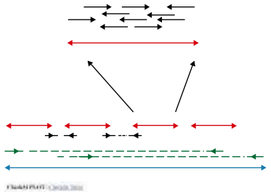
Integrated genomics provide insights for the evolution of the polyphosphate accumulation trait of *Ca.* Accumulibacter



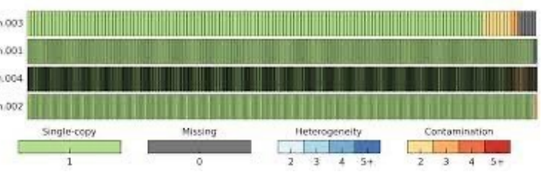
What make us different?



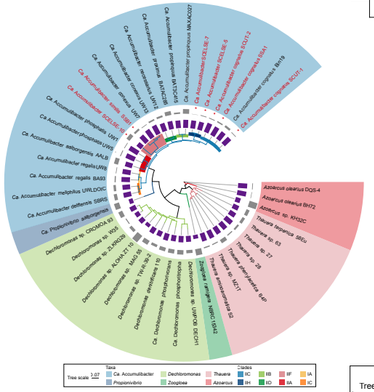
01 Data acquisition and evaluation

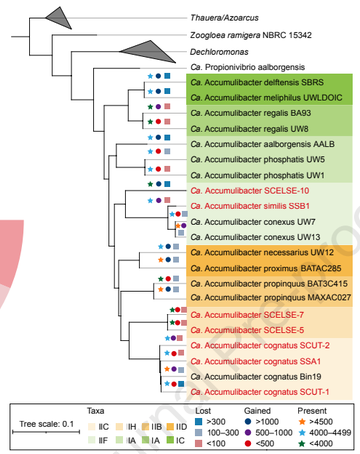


National Center for Biotechnology Information

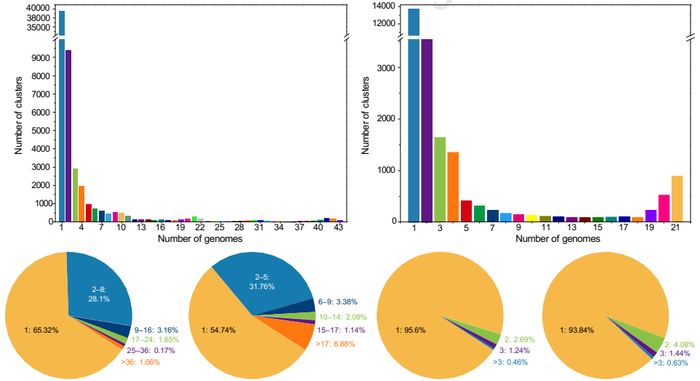


02 Orthologue analyses

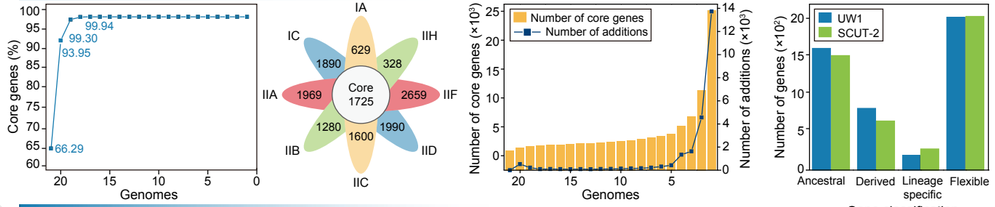




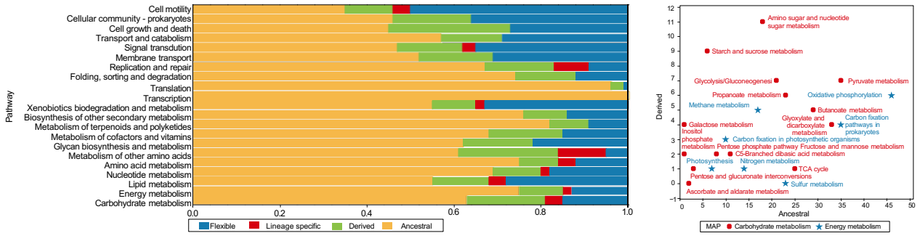
03 Pangenome analysis



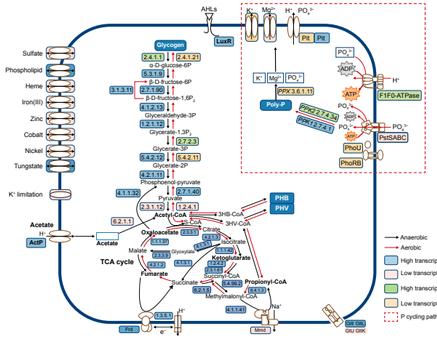
04 Gene flux analysis



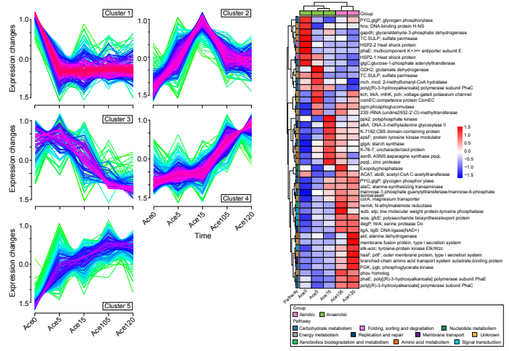
05 Metablic function analysis



06 Horizontal gene transfer



07 Metatranscriptomic analysis



Integrated genomics provides insights into the evolution of the polyphosphate accumulation trait of *Ca.* Accumulibacter

Xiaojing Xie[a, 1], Xuhan Deng[a, 1], Liping Chen[a], Jing Yuan[a], Hang Chen[a], Chaohai Wei [a,c], Xianghui Liu[b,c], Stefan Wuertz[b,c*], Guanglei Qiu[a,b,d,e*]

[a] *School of Environment and Energy, South China University of Technology, Guangzhou 510006, China.*

[b] *Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore 637551, Singapore.*

[c] *School of Civil and Environmental Engineering, Nanyang Technological University, Singapore 639798, Singapore*

[d] *Guangdong Provincial Key Laboratory of Solid Wastes Pollution Control and Recycling, Guangzhou 510006, China*

[e] *The Key Lab of Pollution Control and Ecosystem Restoration in Industry Clusters, Ministry of Education, Guangzhou 510006, China*

[*] Corresponding Author: qiugl@scut.edu.cn (G.Q.), SWuertz@ntu.edu.sg (S.W.)

[1] Authors contributed equally towards this study

Abstract: *Candidatus* Accumulibacter, a prominent polyphosphate-accumulating organism (PAO) in wastewater treatment, plays a crucial role in enhanced biological phosphorus removal (EBPR). The genetic underpinnings of its polyphosphate accumulation capabilities, however, remain largely unknown. Here, we conducted a comprehensive genomic analysis of *Ca*. Accumulibacter-PAOs and their relatives within the Rhodocyclaceae family, identifying 124 core genes acquired via horizontal gene transfer (HGT) at its least common ancestor. Metatranscriptomic analysis of an enrichment culture of *Ca*. Accumulibacter revealed active transcription of 44 of these genes during an EBPR cycle, notably including the polyphosphate kinase 2 (PPK2) gene instead of the commonly recognized polyphosphate kinase 1 (PPK1) gene. Intriguingly, the phosphate regulon (Pho) genes showed minimal transcriptions, pointing to a distinctive fact of Pho dysregulation, where PhoU, the phosphate signaling complex protein, was not regulating the high-affinity phosphate transport (Pst) system, resulting in continuous phosphate uptake. To prevent phosphate toxicity, *Ca*. Accumulibacter utilized the laterally acquired PPK2 to condense phosphate into polyphosphate, resulting in the polyphosphate-accumulating feature. This study provides novel insights into the evolutionary emergence of the polyphosphate-accumulating trait in *Ca*. Accumulibacter, offering potential advancements in understanding the PAO phenotype in the EBPR process.

Keywords: *Candidatus* Accumulibacter; Comparative genomics; Horizontal gene transfer (HGT); PhoU; Polyphosphate kinase 2 (PPK2)

1 Introduction

With the rapid development of industry and the economy, there has been a significant surge in wastewater generation. This escalating wastewater production has, in turn, resulted in excessive phosphorus (P) discharge, leading to adverse consequences such as eutrophication, water quality deterioration, and aquatic ecosystem degeneration [1-3]. Enhanced biological phosphorus removal (EBPR) is an environmentally friendly and economical process widely applied in municipal wastewater treatment plants (WWTPs) for P removal [4-10]. This process is mediated by a group of microorganisms, namely polyphosphate-accumulating organisms (PAOs) [11-14]. *Candidatus* Accumulibacter is a model genus of PAOs commonly found in lab- and full-scale EBPR systems [15-18]. Under anaerobic conditions, *Ca.* Accumulibacter uses intracellularly stored polyphosphate (poly-P) as an energy source to power the uptake of volatile fatty acids (VFAs). This metabolic process results in the release of phosphate. The assimilated VFAs are then polymerized and stored as polyhydroxyalkanoates (PHAs). In the subsequent aerobic phase, PHAs are oxidized for cell metabolism and reproduction. Excess phosphate is removed from the aquatic phase to synthesize poly-P, achieving P removal [19-22]. This unique metabolic feature allows PAOs to thrive in alternating anaerobic-aerobic conditions, conferring sustainable P removal. However, the key genetic basis affording PAOs the ability to P cycling is unclear. Genes known to be indispensable for the P cycling feature, e.g., the polyphosphate kinase gene (*ppk*) and exopolyphosphatase gene (*ppx*) for poly-P synthesis and hydrolysis, respectively, and the inorganic phosphate transporter gene (*pit*) and the high-affinity phosphate

63 transporter gene (*pst*) for phosphate transport, are widely preserved in the bacterial
64 domain, including in non-PAOs [23, 24]. Their presence does not guarantee the P
65 cycling ability, and the key genes have yet to be identified. The transition from non-
66 PAOs to PAOs may be driven by adaptive evolution [25, 26].
67 The need to understand the gain and loss of genes in different strains and the genome
68 diversification in a given lineage of organisms gave rise to pangenomics. A pangenome
69 encompasses the entire set of genes from all individuals of a specific lineage [27, 28].
70 Genes in a pangenome are divided into core genes and variable genes [29]. The
71 collection of genes commonly present in all individuals of a specific lineage is called
72 the core pangenome, representing the common genetic features of a microbial lineage
73 [30]. The variable genes can be further divided into unique genes (found in a single
74 strain/genome) and dispensable genes (shared in at least two but not all strains/genomes)
75 [31]. Dispensable genes represent the intra-lineage diversity encoded among different
76 members [29]. By avoiding single sample bias and ensuring full representation of
77 genomic diversity of different lineage members, the analysis of the pangenome provides
78 insight into the genetic basis of common phenotypic characteristics shared in a group
79 of bacteria, greatly improving our ability to solve complex phenotypic problems [32-
80 34]. Comparative genomics has been applied to study the evolution and development
81 of many bacterial species [35-40]. Via comparative genomic analysis, Fernandez-Fueyo
82 et al. [41] found a subset of potentially important genes for selective lignin
83 decomposition in *Ceriporipsis subvermispora*.
84 Oyserman et al. [42] previously constructed a pangenome of the Rhodocyclaceae family

(including ten *Ca.* Accumulibacter and 16 out-group genomes) to explore the genetic composition and evolutionary changes in metabolic pathways of the *Ca.* Accumulibacter genus. However, at the time, limited numbers of *Ca.* Accumulibacter genomes were available, with more than half having low completeness (<90%). The deficiency in genome quality and quantity may result in an inadequate representation of the lineage pangenome and affect the downstream analysis of genes. With the advance in EBPR research and sequencing techniques, increasing numbers of high-quality *Ca*. Accumulibacter genomes have been obtained [7, 14, 18, 43-49]. New PAOs and glycogen-accumulating organisms (GAOs) were also identified in genera phylogenetically closely related to *Ca*. Accumulibacter. GAOs occupy a similar ecological niche as PAOs in EBPR systems. They use glycogen instead of polyphosphate as an energy source for anaerobic carbon source uptake, thus competing with PAOs. For instance, a *Propionivibrio* member was shown to perform as a GAO in full-scale WWTPs in Denmark [50]. Two *Dechloromonas* members in the same WWTPs (i.e., *Ca.* Dechloromonas phosphoritropha and *Ca.* Dechloromonas phosphorivorans) were revealed to be PAOs [51]. The identification of *Dechloromonas*-related PAOs raises the possibility that the emergence of the PAO phenotype may have occurred before the *Ca.* Accumulibacter's last common ancestor (LCA). The evolution in the P cycling feature needs to be re-evaluated and traced. Combined with the analysis of gene transcriptional characteristics of representative PAO strains, the key genomic characteristics distinguishing PAOs and non-PAOs may be further identified and determined, which would significantly advance the understanding of the genomic basis

of the PAO phenotype.

To understand the emergence of the PAO phenotype of *Ca.* Accumulibacter, we selected 43 high-quality genomes within the Rhodocyclaceae family for comparative genomic analysis. A pangenome of the Rhodocyclaceae family, including 21 *Ca*. Accumulibacter genomes, seven of which were recovered from our EBPR reactors, 22 out-group genomes, including two confirmed *Dechloromonas* PAOs, i.e., *Ca.* Dechloromonas phosphoritropha and *Ca.* Dechloromonas phosphorivorans [51], and one *Propionivibrio* GAO genome, *Ca.* Propionivibrio aalborgensis [50], was constructed. In the analysis of genes within the pangenome, genes were classified as ancestral, derived, flexible, or lineage-specific genes. The dynamics in these genes in the evolutionary process were analyzed, and metatranscriptomic analyses were performed on an enrichment culture of *Ca*. Accumulibacter Clade IIC SCUT-2 for identifying their active genes in a typical anaerobic-aerobic cycle to narrow down the range of genes important for the PAO phenotype of *Ca*. Accumulibacter. Genomic comparisons were further performed between *Ca*. Accumulibacter, two *Dechloromonas*-related PAOs, and the *Propionivibrio* GAO. Among the numerous genes investigated, two key players emerged: the phosphate signaling complex protein gene (*pho*U) in the Pho regulon and the laterally derived polyphosphate kinase 2 gene (*ppk*2). These genes were identified as instrumental in the emergence of the PAO phenotype of *Ca.* Accumulibacter. This study provides new insights into the development of the P cycling trait of *Ca.* Accumulibacter.

2 Materials and Methods

2.1 Data acquisition and evaluation

The genomes used for analysis included seven high-quality genomes recovered from our EBPR reactors and 36 genomes obtained from the National Center for Biotechnology Information (NCBI) database. All 43 genomes belong to the Rhodocyclaceae family, including 21 *Ca.* Accumulibacter genomes and 22 out-group genomes (ten *Dechloromonas*, seven *Thauera*, three *Azoarcus*, one *Propionivibrio*, and one *Zooglea ramigera* genomes). The completeness and contamination of the genomes were evaluated using CheckM [52]. The GenBank assembly accession, corresponding species names, and additional details about the qualities of these genomes can be found in the Supplementary Materials Table S1–S3.

2.2 Orthologue analyses

Orthologous gene clustering is necessary for the reconstruction of the ancestral state. To find orthologous gene clusters based on the protein sequences, all vs. all BLAST of each Rhodocyclaceae genome was conducted using Orthofinder 2.5.4 [53] with parameters *-evalue 1e-5, -seg yes, -soft_masking true, -use_sw_tback*. The results were filtered to the query coverage ≥ 75% and the percent identity ≥ 70%. Orthologous gene clusters were identified using MCL version 14–137 with an inflation value of 1.1 [54].

2.3 Phylogenetic analysis of pangenome

Orthofinder was used to identify the pan single-copy genes for reliable phylogenic tree

construction and gene flux analysis. The pan single-copy genes were aligned using the linsi option in MAFFT version 7.508 [55] and masked in Gblocks version 0.91b [56]. Seqkit (version 2.3.0) [57] was used to sort the single-copy gene sequences and convert the multi-line sequences into a one-line sequence. Iqtree version 2.2.0.3 [58] was used to predict the best phylogenetic tree model. Finally, the tree was constructed with model Q. insect+F+I+I+R4. Landscaping of the phylogenetic tree was achieved using iTOL version 6.6 [59].

2.4 Pangenome analysis

When a genome set has incomplete genomes, it is necessary to determine a threshold number of genomes in which a gene must be observed in order to call it 'core'. The probability that a gene was observed in all *Ca.* Accumulibacter genomes are the product of the completeness of each genome. The probability of a gene's absence in one genome while being present in all other genomes was computed by multiplying the completeness of the remaining genomes with the incompleteness (i.e., 1 minus the completeness) of the incomplete genomes. Cut-off values were calculated using the R script [60] (Supplementary Materials Spreadsheet 1). The maximum number of genomes allowing an effective calculation of the cutoff value was 21. Via a comprehensive evaluation of the quality and the clade distribution of all available genomes, 21 high-quality *Ca.* Accumulibacter genomes covering eight different clades were used for pangenomic analysis (The completeness and contamination of these genomes are documented in the Supplemental Materials Table S1–S2).

2.5 Gene gain/loss analysis

Gene flux was analyzed using Count [61] based on the matrix of orthologous gene family abundance obtained in the previous analyses. For a more comprehensive examination of gene gain and loss dynamics, we applied a Wagner parsimony penalty of 2 [62, 63]. Genes acquired before the node of the LCA of *Ca.* Accumulibacter were defined as ancestral, while those acquired at the node of *Ca.* Accumulibacter LCA were defined as derived genes. Genes determined to be obtained via horizontal gene transfer (HGT) in the derived genes were classified as laterally derived genes. Lineage-specific genes were present in a single *Ca.* Accumulibacter genome. Flexible genes were present in more than one but less than 18 *Ca.* Accumulibacter genomes. Genetic comparisons were performed between PAO and GAO genomes to understand the differences in their genetic makeup better. The pangenome is composed of 21 *Ca.* Accumulibacter and two *Dechloromonas* PAOs were denoted as the pan PAO genome. Core genes of the pan PAO genome were defined as genes belonging to the core genes of the pan *Ca.* Accumulibacter genome and were also present in two *Dechloromonas* PAO genomes. Differential genes were defined as core genes present in the pan PAO genome but absent in the *Ca.* Propionivibrio aalborgensis GAO genome.

2.6 Metabolic function analysis

The ancestral, derived, flexible, and lineage-specific genes were annotated and classified based on KEGG annotations [64] of clade IIC member SCUT-2 [49] and clade IIA member UW1 [4, 26]. The number of genes annotated in each metabolic pathway

was counted, with the number of each type of gene being divided by the total number of genes in the pathway. Metabolic pathways with high proportions of derived genes were considered to have undergone major changes during evolution.

2.7 Horizontal gene transfer (HGT) identification

Parametric and phylogenetic methods are commonly used to infer HGT [65]. This study used the phylogenetic method for HGT identification. Each derived gene was queried in the non-redundant (NR) database (published on May 7, 2015) [66] using the following BLASTP parameter [*-max_target_seqs 100-value 1E-6*] to preserve the first 100 BLAST results. The representative species were obtained from the first 100 BLAST results. Subsequently, the numbers and percentages of *Ca.* Accumulibacter, non-*Ca.* Accumulibacter Rhodocyclaceae, and non-Rhodocyclaceae members in the first 100 BLAST results were then calculated. A gene was considered a laterally derived gene if the numbers of *Ca.* Accumulibacter or non-*Ca.* Accumulibacter Rhodocyclaceae-related hits were less than 10%. All core and differential-derived genes in each metabolic pathway were analyzed to determine if they were obtained via HGT. The derived genes that were classified as HGT-originated are referred to as laterally derived genes. The origination of key genes (*ppk*2 and the homolog of *pho*U) was further confirmed using the phylogenetic method based on best-match analysis.

2.8 Metatranscriptomic analysis

An anaerobic-aerobic full-cycle study was performed on an enrichment culture of *Ca.*

Accumulibacter Clade IIC SCUT-2 in the lab-scale EBPR reactor SCUT (Supplementary Materials). The P cycling activities and the transformation of carbon compounds were monitored. Activated sludge samples were collected just before the start of a sequencing batch reactor (SBR) cycle (0 min), and at 5 min (anaerobic phase), 30 min (anaerobic phase), 105 min (aerobic phase), and 120 min (aerobic phase) of the SBR cycle. The samples were snap-frozen in liquid N_2 and stored at −80 °C before the extraction of ribonucleic acid (RNA) for metatranscriptomic analysis.

For metatranscriptomic analysis, total RNA was extracted using the RNA PowerSoil® Total RNA Isolation Kit (Omega Bio-Tek, GA, USA). Fastp [67] and SortMeRNA [68] removed adaptation sequences and ribosomal ribonucleic acids (rRNAs). Filtered reads were mapped to the corresponding *Ca.* Accumulibacter draft genome (i.e., SCUT-2) using BBMap version 38.96 [69] and were normalized to transcript per million (TPM). Genes with TPM > 100 were considered to be highly transcribed. Details on the reactor operation, full-cycle study, sample collection, metagenomic analysis, and metatranscriptomic analysis are found in the Supplementary Materials. Raw reads and draft genomes obtained were submitted to NCBI under BioProject No. PRJNA807832 and No. PRJNA771771.

3 Results

3.1 Identification of orthologous gene clusters

A total of 60722 pan Rhodocyclaceae orthologous gene clusters were identified, including 25080 homologous genes in the *Ca.* Accumulibacter pangenome

(Supplementary Materials Spreadsheet 2, Sheets 1 and 3). Large proportions (63.8% and 54.7%) of gene families in the pan Rhodocyclaceae and pan *Ca.* Accumulibacter genomes were present in only single genomes (Fig. 2a,b). Approximately 1% (626) of gene families were present in ≥ 37 of the 43 genomes, which were used to define the core pan Rhodocyclaceae genome (Fig. 2c). In the pan *Ca.* Accumulibacter genome, 6.9% of genes were shared in ≥ 18 genomes (Fig. 2d). Non-paralogous genes (average gene copy per genome = 1) account for high proportions of pan Rhodocyclaceae and pan *Ca.* Accumulibacter genomes (95.6% and 93.8%, respectively) (Fig. 2e,f). The orthologous gene cluster identification results include the number of representative genes in each genome and summary statistics of pan Rhodocyclaceae and pan *Ca.* Accumulibacter gene clusters are provided in Supplementary Materials Spreadsheet 2 (Sheets 2 and 4).

3.2 Gene flux analysis

Among the 25080 gene clusters in the pan *Ca.* Accumulibacter genome, 2499 (9.96%) were inferred to occur in the genome of the LCA, and 1668 (6.73%) occurred before the LCA. Eight hundred eighteen (3.26%) were acquired at the node of LCA. Gene occurrence possibility calculation suggested that with a genome-number cutoff of 18, 99.94% of core genes could be identified (Fig. 3a). At this cutoff value, 1725 (6.88%) core genes were identified in the pan *Ca.* Accumulibacter genome (Fig. 3b,c). By further reducing the cutoff value to 17, the number of core genes increased from 1725 to 1829, and those with known functions increased from 298 to 318. As this study

mainly focused on the changes in the genetic content, i.e., new core-derived genes and horizontally transferred genes, looser cutoff values did not seem to bring new gains. Thus, a relatively stricter cut-off value (i.e., 18) was used to ensure the accuracy of the results. The gene gain or loss of a pangenome needs to be characterized in specific lineage member genomes. To facilitate a subsequent combination with the transcriptome data, SCUT-2 and UW1 were used as representative genomes for gene flux analysis. Each gene in Clade IIC SCUT-2 and Clade IIA UW1 genomes was classified as ancestral, derived, lineage-specific, or flexible genes. There were no significant differences in the numbers and proportions of ancestral and flexible genes in these two genomes (ancestral genes accounted for 32.6% and 34.7%; flexible genes accounted for 43.8% and 43.6% in SCUT-2 and UW-1, respectively). Six hundred thirty eight and eight hundred and two derived genes were found in the SCUT-2 and UW1 genomes (17.6% and 14.0%, respectively). One hundred eighty nine lineage-specific genes (genes occurred only in UW1) were observed in UW1, which was slightly less than those (i.e., 275) in the SCUT-2 genome (Fig. 3d). Figure 4 and Supplementary Materials Spreadsheet 3 provided additional details about the presence, gain and loss of genes, and the discrete categories to which they were assigned.

3.3 Evolution of *Ca.* Accumulibacter metabolic pathways

The collections of genes identified as ancestral, derived, flexible, and lineage-specific genes were annotated using KEGG [64] and were grouped into different metabolic pathways. In SCUT-2, 2293 genes were annotated to various metabolic pathways. The

translation metabolic pathway had the highest proportion of ancestral genes (77, accounting for 96%). The largest number of ancestral genes (224) and derived genes (63) was observed in the carbohydrate metabolism pathway, accounting for 63% and 18%, respectively. The highest proportion (15 out of 53, 28.0%) of derived genes was observed in the cell growth and death metabolic pathway (Fig. 5a). Within each primary pathway, ancestral and derived genes also showed distinct proportions in different secondary pathways. For instance, within the carbohydrate metabolism, the galactose metabolism pathways had the highest proportion (4 out of 5, 80%) of derived genes. Whereas ancestral genes dominated the citric acid cycle (TCA cycle) (25 out of 30, 83%) and the glyoxylate and dicarboxylate metabolism pathways (33 out of 45, 73%). In signal transduction, the two-component system contained the highest proportion of derived genes (27 out of 182, 15%). In membrane transport, among the 122 ABC-transporter encoding genes, 18 were derived (15%) (Fig. 5). Similar number and proportion of genes assigning to various metabolic pathways were observed in the *Ca*. Accumulibacter clade IIA UW1 genome with only two metabolic pathways (transport and catabolism, cell growth and death) showing significant differences in the proportions of derived genes (28% and 40% in SCUT-2 and 14% and 23% in UW1, respectively) (Fig. 5 and Supplementary Materials Fig. S1). These results indicated that different strains of *Ca.* Accumulibacter underwent comparable developmental changes during evolution but, at the same time, preserved a certain degree of gene diversity. Detailed annotation of each gene in SCUT-2 and UW1 can be viewed in Supplementary Spreadsheet 4.

3.4 Pan *Ca.* Accumulibacter phylogenetic analysis of derived genes

Relatively strict parameters (i.e., 70% identity and 75% coverage) were used to identify homologous gene clusters. The derived genes were manually classified into those derived from accumulative mutations and HGT. Phylogenetic analysis was further performed to confirm that *ppk*2 and the homolog of *pho*U are horizontally derived (Supplementary Materials Fig. S4). Among 298 core-derived genes successfully annotated in KEGG, 124 were shown to have been acquired via HGT. Among the 124 genes, 67 were involved in KEGG pathways. The carbohydrate metabolism pathway harbors the highest numbers (25) of derived genes via HGT, including these in glycolysis/gluconeogenesis (e.g., genes encoding the phosphoglucomutase, the glucokinase, and the phosphoglycerate kinase), starch and sucrose metabolism (e.g., the starch synthase, and the glycogen phosphorylase genes), and in butanoate metabolism (genes encoding the poly[(R)-3-hydroxyalkanoate polymerase subunits). In signal transduction, the two-component system contained ten laterally derived genes, such as genes encoding the REDOX signal transduction system proteins RegA/B and the phosphate regulon proteins PhoR-PhoB. Another remarkable set of genes derived via HGT was oxidative phosphorylation in the energy metabolism pathway, including these encoding the NADH-quinone oxidoreductase subunit, the polyphosphate kinase, and the cytochrome C. The inorganic phosphate transporter gene (*pit*) was also acquired via HGT. Similar results were observed for UW1. In the two-component system, genes encoding the REDOX signal transduction system proteins RegA/B and the phosphate regulon proteins PhoR-PhoB were laterally derived. More details about the BLAST

comparison results can be found in Supplementary Materials Spreadsheet 5.

3.5 Comparison of genetic compositions in PAOs and non-PAOs

In the context of our investigation, the presence or absence of specific genes in *Ca.* Accumulibacter, compared to closely related PAOs and non-PAOs, holds significant implications for elucidating the genetic basis of the P cycling phenotype. If a gene was present in *Ca.* Accumulibacter, but absent in other closely related PAOs, may also not be a key to developing the P cycling phenotype. Conversely, if a gene was present in *Ca.* Accumulibacter, or their closely related PAOs but absent in non-PAOs, might be a key gene to the emergence of the PAO phenotype. For a better understanding of the genomic difference between closely related PAOs and non-PAOs, a pan PAO genome (composed of 21 *Ca.* Accumulibacter and two *Dechloromonas* PAOs) [51] analysis was performed. The pan PAO genome was compared to the *Ca.* Propionivibrio aalborgensis (a closely related GAO) [50] genome to identify differential genes (defined as core genes present in the pan PAO genome but absent in the *Ca.* Propionivibrio aalborgensis genome). In the pan PAO genome, 124 differential genes were identified. Alkaline phosphatase synthesis response regulator (PhoP) and polyphosphate kinase 2 (PPK2) genes were both differential genes. Other genes in the operon or the genes regulated by PhoP were not differential genes. Carbohydrate metabolism had the largest differential genes (16), including those encoding the acetyl-CoA C-acetyltransferase and the enoyl-CoA hydratase. The cofactor and vitamin metabolic pathway harbored the second largest number of differential genes (11), followed by energy metabolism (9),

replication and repair (6), and signal transduction (5) metabolic pathways. The lowest number (1) of differential genes was observed in the transcription and metabolism of other amino acid pathways. Further analysis of another 21 available *Propionivibrio* genomes confirmed that *ppk*2 and *pho*U are differential genes between *Ca.* Accumulibacter and *Propionivibrio*. HGT analysis was aimed at gene acquisition in *Ca.* Accumulibacter during evolution, based on the hypothesis that the emergence of the P cycling ability by PAOs resulted from the acquisition of certain key genes. However, the hypothesis ignored the possibility that non-PAOs may have lost certain key genes in the process of evolution, leading to their inability to remove P. Differential genes included gene loss in non-PAOs during evolution. The analysis in this part allows us to understand the evolutionary process from a different perspective more comprehensively. More details about the differential genes (metabolic pathway and functional annotation) can be found in Supplementary Materials Spreadsheet 6.

3.6 Metatranscriptomic profiles

By analysis of the gene transcription levels of *Ca*. Accumulibacter in a typical EBPR cycle, we excluded genes that displayed no remarkable transcription in the comparative genome may be excluded. Thus, the range of genes could be further narrowed down, facilitating the identification of key genes important to the PAO phenotype. Metatranscriptomic analysis was performed on an enrichment culture of *Ca*. Accumulibacter clade IIC strain SCUT-2 (with a relative abundance of 37.1%, as suggested by the metagenomic analysis). In the SCUT-2 genome, out of 5037 annotated

359 genes, 906 were highly transcribed (TPM > 100). There were 298 core-derived genes,
360 84 of which were highly transcribed (Supplementary Materials Spreadsheet 7). To
361 understand the dynamic patterns and functional relationships of 905 core genes with
362 known function, they were classified into five clusters using the Mfuzz [70] (Fig. 6b).
363 Most genes (e.g., the acetate permease gene *act*P, NOF05_02545) in Cluster 1 were
364 related to the transporter for carbon uptake and energy utilization. Cluster 2 showed a
365 pattern of increased transcription throughout the anaerobic period, peaking after oxygen
366 exposure. Key members of this cluster included the phosphate transport system
367 substrate-binding protein (*pstS*, NOF05_04305) and the laterally derived polyphosphate
368 kinase 2 gene (*ppk2*, NOF05_17285). Cluster 3 genes showed high transcription at the
369 beginning of the anaerobic stage and reduced towards the end of the anaerobic cycle,
370 correlating with the depletion of acetate (Fig. 6a). Their high transcription in the aerobic
371 stage was mostly related to the routing of anaerobically stored carbon to the TCA cycle
372 and glycogenesis [7, 26]. Cluster 4 contained genes encoding the distant homolog of
373 PhoU (NOF05_17860, NOF05_12350) and antitoxin CptB (NOF05_13125), which
374 showed low transcription during the anaerobic stage but were upregulated during the
375 aerobic phase. These genes possibly play a role in sustaining vital activities and
376 controlling homeostatic environments [71]. Finally, genes in Cluster 5 may be
377 associated with the maintenance of stable intracellular environments or cell growth,
378 including genes encoding the ion transporters, such as the magnesium transporter gene
379 (NOF05_18175) and the low-affinity inorganic phosphate transporter (*pit*,

NOF05_12345). These clustering patterns aligned with the metabolic characteristics of *Ca.* Accumulibacter in EBPR (Fig. 6a).

The transcription of horizontally transferred genes in SCUT-2 was further analyzed. 44 genes, which were identified to be obtained via HGT, were highly transcribed (Fig. 6c). These genes were involved in pathways, such as glycolysis/gluconeogenesis (phosphoglycerate kinase, and phosphoglucomutase), ABC transporters (branched-chain amino acid transport system substrate-binding protein), butanoate metabolism (poly[(R)-3-hydroxyalkanoate] polymerase subunit), the two-component system (low molecular weight protein-tyrosine phosphatase, polysaccharide biosynthesis/export protein, tyrosine-protein kinase, and serine protease), transporters for inorganic salts (sulfate permease, and magnesium transporter), and showed high transcription throughout the EBPR cycle. Polyphosphate kinase 2 gene (*ppk*2) was also highly transcribed and was significantly upregulated in the anaerobic phase. The transcription of the phosphate transport regulator (a distant homolog of PhoU) was significantly upregulated in the aerobic stage. PHA synthesis-related genes were also highly transcribed. A full list of the SCUT-2 gene transcription data can be found in Supplementary Materials Spreadsheet 7.

Comparisons were further made to the gene transcription characteristics of UW1 [26]. 35 horizontally derived gene families were highly transcribed in both SCUT-2 and UW1 (Supplementary Materials Fig. S2). Apart from the homolog *pho*U genes and *pit*, which are related to phosphate regulation and transport, 42 laterally derived gene families were under-transcribed in SCUT-2 but highly transcribed in UW1, including the acetate

kinase gene. These 42 gene families may not play a key role in the evolution of non-PAO to PAO due to their different transcription behaviors in SCUT-2 and UW1. Combined with transcriptomic analysis, the range of key genes can be effectively reduced, and a metabolic model of *Ca*. Accumulibacter can be constructed (Fig. 6d). Most genes in the central carbon metabolic pathway were highly transcribed non-HGT genes, indicating that this pathway is indispensable for *Ca.* Accumulibacter, yet raises doubts about its direct involvement in the evolution from a non-PAO metabolism to a PAO. In the P cycling pathway, several laterally acquired genes were involved, suggesting their potential pivotal role in the evolution of *Ca*. Accumulibacter. Some of them were highly transcribed, further implying their importance in the evolution of *Ca.* Accumulibacter (Fig. 6).

4 Discussion

Previous research suggested that the transition of PAO from non-PAO may have occurred at the node of *Ca.* Accumulibacter LCA [42]. However, a recent investigation has put forth compelling evidence indicating the presence of PAOs in the *Dechloromonas* genus (i.e., *C*a. Dechloromonas phosphoritropha, *C*a. Dechloromonas phosphorivorans) [51], raising a possibility that the emergence of the PAO phenotype may have occurred before the *Ca.* Accumulibacter LCA. Here, we discuss the function of key laterally derived genes in the context of pangenomics and known PAO metabolism. A metatranscriptomic analysis of an enrichment culture of *Ca*. Accumulibacter Clade IIC member SCUT-2 contrasting those of *Ca*. Accumulibacter

Clade IIA UW1 was performed to study the transcriptional dynamics of key genes in *Ca.* Accumulibacter. This approach allowed the exclusion of genes that were not highly transcribed in the large collection of laterally derived genes to narrow down the range of key genes to obtain new insights on key genomic features of the polyphosphate accumulating trait.

4.1 Carbon substrate uptake

The largest number of genes were annotated to the carbohydrate metabolism pathway in both SCUT-2 and UW1 genomes (354 and 369, respectively). The SCUT-2 genome contained 224 ancestral genes, 63 derived genes, and 49 laterally derived genes. Transcriptomic analysis suggested that when acetate was used as a carbon source, genes directly related to intracellular acetate processing and PHA synthesis were remarkably upregulated in SCUT-2 (Supplementary Materials Spreadsheet 7). The high-affinity acetyl-CoA synthetase (NOF05_02565) and low-affinity phosphate acetyltransferase (NOF05_11790) are responsible for acetate activation [11, 72]. Other genes involved in the acetyl-CoA pathway, including the pyruvate kinase gene (NOF05_14290) and the phosphoenolpyruvate carboxykinase gene (NOF05_14615), maintained high levels of transcription throughout the anaerobic-aerobic cycle. However, these genes are all ancestral genes. Only one horizontally transferred gene (i.e., the acetate kinase gene, NOF05_16845) was barely transcribed. Therefore, genes related to acetate processing may not be pivotal factors contributing to the emergence of the PAO phenotype. In addition, in the TCA cycle [73], there were 30 genes. Among them, only the

444 dihydrolipoamide dehydrogenase gene (NOF05_18520) was laterally derived, whereas
445 transcribed at a low level. This indicates that the gain/loss of genes in the TCA cycle
446 might not have contributed remarkably to the evolution of non-PAOs to PAOs. Four
447 laterally derived genes occurred in the PHA synthesis pathway (*pha*C NOF05_18015,
448 NOF05_21650, NOF05_21620, and *pha*A NOF05_18020), NOF05_21650 and
449 NOF05_21620 were highly transcribed throughout the EBPR cycle (Fig. 6). Whereas
450 *Ca.* Propionivibrio aalborgensis also encoded these genes [50]. Their contribution to
451 the evolution from a non-PAO metabolism to a PAO metabolism was unlikely.

452 **4.2 Two-component systems**

453 The two-component signal transduction system enables bacteria to sense, respond, and
454 adapt to diverse and dynamic environmental conditions [74]. This system is commonly
455 preserved in the bacterial domain. The number of genes in the two-component system
456 was considered to be closely related to the bacteria's living environment [75]. Bacteria
457 living in extreme environments tend to encode many signaling proteins for improved
458 adaption [76]. In the SCUT-2 genome, a total of 182 genes were annotated to the two-
459 component system, including 81 ancestral genes and 27 derived genes. Notably, 12 of
460 these genes have been acquired laterally. In both SCUT-2 and UW1, phosphate regulon
461 response regulator gene *pho*B (NOF05_18105), phosphate regulon sensor histidine
462 kinase gene *pho*R (NOF05_18105), and redox signaling genes *reg*A and *reg*B
463 (NOF05_11115, NOF05_11120) were laterally derived. RegB/RegA was shown to
464 control and regulate a variety of basic metabolic processes in *Rhodobacter*, *Capsulatus*,

and *Sphaeroides*, such as photosynthesis, CO_2 fixation, N_2 assimilation, denitrification, and electron transport [77] via direct or indirect control of respective operons [78, 79]. However, both *reg*A and *reg*B were absent in two *Dechloromonas* PAO genomes (GCA_016722705.1 and GCA_016721185.1) [51], suggesting that the redox signaling RegA/B were not indispensable for a PAO phenotype. PhoR-PhoB is present in both *Ca.* Accumulibacter and two *Dechloromonas* PAO genomes can potentially play a role in PAO phenotype evolution. Since the PhoR-PhoB system is a part of the Pho regulon, further discussion was provided in the following subsection.

4.3 Phosphate regulatory system

The phosphate regulator (Pho) is a regulatory mechanism to maintain and manage inorganic phosphate concentrations in bacterial cells. The system typically consists of extracellular enzymes, transporters, and enzymes involved in the intracellular storage of phosphate [80]. Signal transduction of Pho regulators requires seven proteins, including PhoR, PhoB, four components of the ABC transporter Pst (PstS, PstA, PstB, and PstC), and PhoU. An increase in the extracellular phosphate concentration near the PstSCAB transporter would increase phosphate binding to PhoU, inhibiting the PhoR kinase activity and the PstSCAB transporter activity. In the absence of phosphate input, PhoU dissociates with phosphate, allowing the phosphate transport (Pst) to return to a normal working state [81]. The above feedback control enables bacteria to maintain and control a relatively stable intracellular phosphate concentration. Most of the genes in the Pho regulatory system in *Ca.* Accumulibacter are laterally derived, including those

encoding PhoR and PhoB. In addition, within the genomes of *Ca*. Accumulibacter, three distant homologs of the phoU gene (designated as NOF05_17860, NOF05_09930, and NOF05_09935) were identified. Distant homologs are protein pairs with similar structures and functions but low gene sequence similarity [82]. The homolog *pho*U is located in the *pit* operon within *Ca.* Accumulibacter genomes. Moreover, PhoR-PhoB is also present in two *Dechloromonas* PAO genomes (*Ca.* Dechloromonas phosphoritropha and *Ca.* Dechloromonas phosphorivorans).

In SCUT-2, the transcription of *pho*R (NOF05_18110) and *pho*B (NOF05_18105, NOF05_19100) was negligible. The transcription level of *pho*R (CAP2UW1_1997) in UW1 was also low. The transcription of *pho*B (CAP2UW1_1996) in the aerobic phase was slightly upregulated (with TPM values from 12 to 92) but was still at relatively low levels (Supplementary Materials Spreadsheet 5). These results suggest that PhoR-PhoB in *Ca.* Accumulibacter was probably not active in perceiving phosphate concentrations. Similarly, the *pho*U genes were almost not transcribed (with the maximum TPM values < 12, Fig. 6). Although the homolog *pho*U genes showed high transcription, the trend was not in line with *pst*, indicating that the PhoU or their laterally derived homologs were not effectively regulating Pst (Supplementary Materials Fig. S3). The same phenomenon was observed in UW1 [26] and UW6 [45] metatranscriptome (Supplementary Materials Spreadsheet 7). In *Staphylococcus aureus*, the absence of *pho*U homolog, located in the *pit* operon, leads to the upregulation of phosphate transporter genes (*pst*), increasing intracellular polyphosphate levels [71]. In *Sinorhizobium meliloti*, the absence of *pho*U resulted in excessive accumulation of

508 phosphate, which inactivates cells due to P poisoning, resulting in poor cell growth [83,
509 84]. Based on these results, we proposed two hypotheses. (1) PhoU in *Ca.*
510 Accumulibacter was ineffective in regulating Pst even under high intracellular
511 phosphate concentrations (no transcription of the *pho*U, and the unmatched
512 transcription of *pho*U homolog and *pst*, Supplementary Materials Fig. S3). Pst
513 continued to operate (as indicated by the high transcription of *pst* in the transcriptome,
514 Supplementary Materials Fig. S3), resulting in excessive phosphate accumulation in
515 cells (Fig. 6a). The laterally derived PPK2 functioned (as suggested by the high
516 transcription of *ppk*2, Supplementary Materials Fig. S3) to condense excess phosphate
517 into poly-P to avoid P poisoning. The second is that, in *Ca*. Accumulibacter, since *pho*U,
518 the homolog of *pho*U and *ppk*2 were derived from different donor bacteria
519 (Rhodocyclaceae, *Burkholderia*, and Gramaproteobacteria, respectively, as suggested
520 by the BLAST results, Supplementary Materials SpreadSheet 5), their encoding
521 proteins (i.e., PhoU, PhoU homolog, and PPK2) may have incompatible phosphate
522 activation/inactivation thresholds. PPK2 continued to synthesize poly-P by consuming
523 intracellular phosphate transported via Pst, resulting in consistently low intracellular
524 phosphate concentration, which was insufficient to combine with PhoU and/or its
525 homologs to downregulate Pst. In the SCUT-2 and UW1 transcriptomes (Fig. 6), PPK2
526 showed high levels of transcription during the entire EBPR cycle (with TPM values up
527 to 12481 in SCUT-2), which was further up-regulated in the aerobic stage, suggesting
528 that PPK2 worked to synthesize poly-P by consuming phosphate which was imported
529 via Pst, avoiding possible cell inactivation and poisoning due to elevated intracellular

phosphate concentrations and achieved poly-P accumulation. In addition, *Ca.* Dechloromonas phosporitropha lacked *pst*, *pho*U, *pho*B, and *pho*R genes in the Pho regulon, which is consistent with our hypothesis that the Pho regulation may not work properly in PAOs. The transcriptomics data of *Microlunatus phosphovorus* (BioProject No. PRJNA984968) and proteomics data of *Tetrasphaera elongate* (obtained from Herbst et al. [85]) were further analyzed to check whether the same hypothesis could apply to other PAOs. In the transcriptome of *Microlunatus phosphovorus*, we found that the transcriptional patterns of *pst* were also inconsistent with those of *pho*U during an anaerobic and aerobic cycle (Supplementary Materials). From the proteome of *Tetrasphaera elongate*, the relative abundances of Pst and PhoU did not vary significantly between anaerobic and oxic conditions; hence, they were not significantly affected by changes in phosphate concentrations [85]. Taken together, these results suggest that in *Microlunatus phosphovorus* and *Tetrasphaera elongate*, the regulation of Pst by PhoU was not effective and that the Pho dysregulation hypothesis may also apply to non-*Ca.* Accumulibacter PAOs. However, additional work is needed to confirm its broad applicability.

Despite that, there is limited research on the Pho regulatory system in *Ca.* Accumulibacter, the transcriptomics and gene origination analysis in the Pho regulon suggested that it may represent a key link in the emergence of the PAO phenotype.

4.4 Transport of phosphate

Phosphorus (organic and/or inorganic) is a typical restricting nutrient. Therefore, microorganisms developed adaptive mechanisms to cope with ordinary P deficiency. Low-affinity inorganic phosphate transport systems (Pit) and high-affinity phosphate transport systems (Pst) are key transporters used for inorganic phosphate transport [86, 87]. In the pan *Ca.* Accumulibacter genomes, genes encoding the Pst transporter, are neither core nor laterally derived. Furthermore, *Ca.* Dechloromonas phosporitropha (PAO) do not encode any *pst* [51]. These results suggested that the Pst transport system may not be indispensable for a PAO phenotype. *Ca.* Dechloromonas phosporitropha encoded a phosphonates/phosphate transport system (Phn), which was shown to be a high-affinity phosphate transporter in *Mycobacterium smegmatis* [88]. This system may serve as a backup for the Pst transport system in *Ca.* Dechloromonas phosporitropha.

In the pan PAO genome, the low-affinity inorganic phosphate transporter gene (*pit*, NOF05_09925, NOF05_09940) was laterally derived. The efflux of phosphate in symport with H^+ via Pit produces proton motive force, which is a key driving force for the uptake of VFAs, lactate, succinate and amino acids by *Ca.* Accumulibacter [7, 8, 89]. Therefore, *pit* is an important feature gene for the PAO phenotype. In SCUT-2 transcriptomes, the transcription of the *pit* was upregulated during the transition from anaerobic to aerobic conditions (Supplementary Materials Fig. S3). The confirmed GAO, *Ca.* Propionivibrio aalborgensis, which is closely related to *Ca.* Accumulibacter (Fig. 1), are lack of *pit*. But *pit* is present in the genomes of other GAOs, for example, *Defluviicoccus* GAO-HK [90], *Ca.* Competibacter denitrificans, and *Ca.*

Contendobacter odensis [91]. In addition, we analyzed 21 *Propionivibrio* genomes in the NCBI database. Pit transporter was encoded in 13 of 21 *Propionivibrio* genomes (Supplementary Materials Table S4). Anyhow, *pit* may not be a key feature driving the evolution of non-PAO into PAOs and may neither be used as a marker gene for the PAO phenotype, although it is indispensable for the P cycling trait.

5. Conclusion

In this study, we conducted pangomics with metatranscriptomic analysis on an enrichment culture of *Ca.* Accumulibacter clade IIC member SCUT-2. The primary objectives of this investigation were to understand the genomic transition in the evolution of *Ca.* Accumulibacter and to identify the key genes responsible for the emergence of the P-accumulating traits. Our study has brought forth several noteworthy findings:

(1) A total of 298 core genes were identified as novel acquisitions in the ancestral lineage of *Ca.* Accumulibacter, with 124 of them being derived via HGT. Notably, 44 of these laterally derived core genes were highly transcribed in a typical EBPR cycle.

(2) A high-affinity phosphate transport system (Pst) may not be indispensable for the PAO phenotype. Inorganic phosphate transporter (Pit) may not be a key feature driving non-PAO evolution into PAOs. Consequently, their encoding genes may not be reliable markers for the PAO phenotype.

(3) Low transcription of the *pho*R-*pho*B two-component system genes and the unmatched transcription of *pst* and *pho*U implied that the Pho regulon may not function properly in *Ca.* Accumulibacter.

(4) A Pho dysregulation hypothesis was proposed. The PhoU and laterally derived PhoU homologs in *Ca.* Accumulibacter were ineffective in regulating Pst, resulting in excessive P uptake. To avoid P poisoning, the laterally derived PPK2 was employed to condense excess phosphate into poly-P. Alternatively, PhoU and PPK2 genes were derived from different donor bacteria, resulting in unmatched activation/inactivation thresholds. PPK2 tends to reduce the intracellular phosphate concentration to levels perceived by PhoU as low-phosphate states, thereby promoting continuous phosphate uptake.

This study is expected to provide a new perspective for understanding the development and evolution of the P cycling traits for *Ca.* Accumulibacter.

CRediT authorship contribution statement

Xiaojing Xie: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. **Xuhan Deng:** Data Curation, Resources, Visualization. **Liping Chen:** Data Curation, Resources, Visualization. **Jing Yuan:** Investigation, Resources, Data Curation. **Hang Chen:** Investigation, Resources, Data Curation. **Chaohai Wei:** Writing - Review & Editing, Supervision. **Xianghui Liu:** Investigation, Resources, Data Curation. **Stefan Wuertz:** Supervision, Writing - Review & Editing, Project Administration, Funding Acquisition. **Guanglei Qiu:** Conceptualization, Methodology, Investigation, Supervision, Writing - Review & Editing, Validation, Project Administration, Funding Acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (52270035 and 51808297), the Natural Science Foundation of Guangdong Province (2021A1515010494), the Guangzhou Key Research and Development Program (2023B03J1334), and the Pearl River Talent Recruitment Program (2019QN01L125).

Data available

All data generated or analyzed during this study are included in this published article. Metagenomic raw reads and draft genomes were submitted to NCBI under BioProject No. PRJNA807832 and No. PRJNA771771. Metatranscriptomic data were submitted to NCBI under the submitted No. PRJNA807832. Other data were documented in the Supplementary Materials.

Reference

[1]. Bunce JT, Ndam E, Ofiteru ID, Moore A, Graham DW: **A review of phosphorus removal technologies and their applicability to small-scale domestic wastewater treatment systems**. *Frontiers in Environmental Science* 2018, **6:**8.

[2]. Abdelfattah A, Ali SS, Ramadan H, El-Aswar EI, Eltawab R, Ho S-H, Elsamahy T, Li S, El-Sheekh MM, Schagerl M, Kornaros M, Sun J: **Microalgae-based wastewater treatment: Mechanisms, challenges, recent advances, and future prospects**. *Environmental Science and Ecotechnology* 2023, **13**:100205.

[3]. Qiu G, Law Y, Zuniga-Montanez R, Deng X, Lu Y, Roy S, Thi SS, Hoon HY, Nguyen TQN, Eganathan K, Liu X, Nielsen PH, Williams RBH, Wuertz S: **Global warming readiness: Feasibility of enhanced biological phosphorus removal at 35 °C**. *Water Research* 2022, **216**:118301.

[4]. Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities**. *Nature Biotechnology* 2006, **24**(10):1263-1269.

[5]. Oehmen A, Lemos PC, Carvalho G, Yuan Z, Keller J, Blackall LL, Reis MAM: **Advances in enhanced biological phosphorus removal: From micro to macro scale**. *Water Research* 2007, **41**(11):2271-2300.

[6]. Qiu G, Zuniga-Montanez R, Law Y, Thi SS, Nguyen TQN, Eganathan K, Liu X, Nielsen PH, Williams RBH, Wuertz S: **Polyphosphate-accumulating organisms in full-scale tropical wastewater treatment plants use diverse carbon sources**. *Water Research* 2019, **149**:496-510.

[7]. Qiu G, Liu X, Saw NMMT, Law Y, Zuniga-Montanez R, Thi SS, Ngoc Nguyen TQ, Nielsen PH, Williams RBH, Wuertz S: **Metabolic traits of *Candidatus* Accumulibacter clade IIF strain SCELSE-1 using amino acids as carbon sources for enhanced biological phosphorus removal**. *Environmental Science & Technology* 2020, **54**(4):2448-2458.

[8]. Chen L, Chen H, Hu Z, Tian Y, Wang C, Xie P, Deng X, Zhang Y, Tang X, Lin X, Li B, Wei C, Qiu G: **Carbon uptake bioenergetics of PAOs and GAOs in full-scale enhanced biological phosphorus removal systems**. *Water Research* 2022, **216**:118258.

[9]. Diaz R, Mackey B, Chadalavada S, kainthola J, Heck P, Goel R: **Enhanced Bio-P removal: Past, present, and future – A comprehensive review**. *Chemosphere* 2022, **309**:136518.

[10]. Zhang C, Guisasola A, Baeza JA: **A review on the integration of mainstream P-recovery strategies with enhanced biological phosphorus removal**. *Water Research* 2022, **212**:118102.

[11]. He S, McMahon KD: **Microbiology of '*Candidatus* Accumulibacter' in activated sludge**. *Microbial Biotechnology* 2011, **4**(5):603-619.

[12]. Nielsen PH, McIlroy SJ, Albertsen M, Nierychlo M: **Re-evaluating the microbiology of the enhanced biological phosphorus removal process**. *Current Opinion in Biotechnology* 2019, **57**:111-118.

[13]. Dorofeev AG, Nikolaev YA, Mardanov AV, Pimenov NV: **Role of phosphate-accumulating bacteria in biological phosphorus removal from wastewater**. *Applied Biochemistry and Microbiology* 2020, **56**(1):1-14.

[14]. Zhang C, Chen X, Han M, Li X, Chang H, Ren N, Ho S-H: **Revealing the role of microalgae-bacteria niche for boosting wastewater treatment and energy reclamation in response to temperature**. *Environmental Science and Ecotechnology* 2023, **14**:100230.

[15]. Seviour RJ, Mino T, Onuki M: **The microbiology of biological phosphorus**

removal in activated sludge systems. *FEMS Microbiology Reviews* 2003, **27**(1):99-127.

[16]. Mao Y, Graham DW, Tamaki H, Zhang T: **Dominant and novel clades of *Candidatus* Accumulibacter phosphatis in 18 globally distributed full-scale wastewater treatment plants**. *Scientific Reports* 2015, **5**(1):11857.

[17]. Roy S, Guanglei Q, Zuniga-Montanez R, Williams RBH, Wuertz S: **Recent advances in understanding the ecophysiology of enhanced biological phosphorus removal**. *Current Opinion in Biotechnology* 2021, **67**:166-174.

[18]. Petriglieri F, Singleton CM, Kondrotaite Z, Dueholm MKD, McDaniel EA, McMahon KD, Nielsen PH: **Reevaluation of the phylogenetic diversity and global distribution of the genus *Candidatus* Accumulibacter**. *mSystems* 2022, **7**(3):e00016-00022.

[19]. Oehmen A, Zeng RJ, Yuan Z, Keller J: **Anaerobic metabolism of propionate by polyphosphate-accumulating organisms in enhanced biological phosphorus removal systems**. *Biotechnology and Bioengineering* 2005, **91**(1):43-53.

[20]. Kolakovic S, Freitas EB, Reis MAM, Carvalho G, Oehmen A: **Accumulibacter diversity at the sub-clade level impacts enhanced biological phosphorus removal performance**. *Water Research* 2021, **199**:117210.

[21]. Zhao W, Bi X, Peng Y, Bai M: **Research advances of the phosphorus-accumulating organisms of *Candidatus* Accumulibacter, *Dechloromonas* and *Tetrasphaera*: Metabolic mechanisms, applications and influencing factors**. *Chemosphere* 2022, **307**:135675.

[22]. Páez-Watson T, van Loosdrecht MCM, Wahl SA: **Predicting the impact of temperature on metabolic fluxes using resource allocation modelling: Application to polyphosphate accumulating organisms**. *Water Research* 2023, **228**:119365.

[23]. Bessarab I, Maszenan AM, Haryono MAS, Arumugam K, Saw NMMT, Seviour RJ, Williams RBH: **Comparative genomics of members of the genus *Defluviicoccus* with insights into their ecophysiological importance**. *Frontiers in Microbiology* 2022, **13:**834906.

[24]. Maszenan AM, Bessarab I, Williams RBH, Petrovski S, Seviour RJ: **The phylogeny, ecology and ecophysiology of the glycogen accumulating organism (GAO) *Defluviicoccus* in wastewater treatment plants**. *Water Research* 2022, **221**:118729.

[25]. Turcotte MM, Corrin MSC, Johnson MTJ: **Adaptive evolution in ecological communities**. *PLOS Biology* 2012, **10**(5):e1001332.

[26]. Oyserman BO, Noguera DR, del Rio TG, Tringe SG, McMahon KD: **Metatranscriptomic insights on gene expression and regulatory controls in *Candidatus* Accumulibacter phosphatis**. *The ISME Journal* 2016, **10**(4):810-822.

[27]. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson

WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"**. *Proceedings of the National Academy of Sciences* 2005, **102**(39):13950-13955.

[28]. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, Xie W-Z, Cheng Y, Zhang Y, Liu K, Yang Q-Y, Chen L-L, Guo L: **Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus***. *Nature Plants* 2020, **6**(1):34-45.

[29]. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Current Opinion in Genetics & Development* 2005, **15**(6):589-594.

[30]. Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN: **How the pan-genome is changing crop genomics and improvement**. *Genome Biology* 2021, **22**(1):3.

[31]. Aggarwal SK, Singh A, Choudhary M, Kumar A, Rakshit S, Kumar P, Bohra A, Varshney RK: **Pangenomics in microbial and crop research: progress, applications, and perspectives**. *Genes* 2022, **13**(4):598.

[32]. Golicz AA, Batley J, Edwards D: **Towards plant pangenomics**. *Plant Biotechnology Journal* 2016, **14**(4):1099-1105.

[33]. Flowers JJ, He S, Malfatti S, del Rio TG, Tringe SG, Hugenholtz P, McMahon KD: **Comparative genomics of two '*Candidatus* Accumulibacter' clades performing biological phosphorus removal**. *The ISME Journal* 2013, **7**(12):2301-2314.

[34]. Camejo PY, Oyserman BO, McMahon KD, Noguera DR: **Integrated omic analyses provide evidence that a "*Candidatus* Accumulibacter phosphatis" strain performs denitrification under microaerobic conditions**. *mSystems* 2019, **4**(1):e00193-00118.

[35]. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran A-N, Wortman JR, Alsmark UCM, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N: **Comparative genomics of trypanosomatid parasitic protozoa**. *Science* 2005, **309**(5733):404-409.

[36]. Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM, Pedrós-Alió C: **Ecology of marine *Bacteroidetes*: a comparative genomics approach**. *The ISME Journal* 2013, **7**(5):1026-1037.

[37]. Coghlan A, Tyagi R, Cotton JA, Holroyd N, Rosa BA, Tsai IJ, Laetsch DR, Beech RN, Day TA, Hallsworth-Pepin K, Ke H-M, Kuo T-H, Lee TJ, Martin J, Maizels RM, Mutowo P, Ozersky P, Parkinson J, Reid AJ, Rawlings ND,

Ribeiro DM, Swapna LS, Stanley E, Taylor DW, Wheeler NJ, Zamanian M, Zhang X, Allan F, Allen JE, Asano K, Babayan SA, Bah G, Beasley H, Bennett HM, Bisset SA, Castillo E, Cook J, Cooper PJ, Cruz-Bustos T, Cuéllar C, Devaney E, Doyle SR, Eberhard ML, Emery A, Eom KS, Gilleard JS, Gordon D, Harcus Y, Harsha B, Hawdon JM, Hill DE, Hodgkinson J, Horák P, Howe KL, Huckvale T, Kalbe M, Kaur G, Kikuchi T, Koutsovoulos G, Kumar S, Leach AR, Lomax J, Makepeace B, Matthews JB, Muro A, O'Boyle NM, Olson PD, Osuna A, Partono F, Pfarr K, Rinaldi G, Foronda P, Rollinson D, Samblas MG, Sato H, Schnyder M, Scholz T, Shafie M, Tanya VN, Toledo R, Tracey A, Urban JF, Wang L-C, Zarlenga D, Blaxter ML, Mitreva M, Berriman M, International Helminth Genomes C: **Comparative genomics of the major parasitic worms**. *Nature genetics* 2019, **51**(1):163-174.

[38]. Kjærbølling I, Vesth T, Frisvad JC, Nybo JL, Theobald S, Kildgaard S, Petersen TI, Kuo A, Sato A, Lyhne EK, Kogle ME, Wiebenga A, Kun RS, Lubbers RJM, Mäkelä MR, Barry K, Chovatia M, Clum A, Daum C, Haridas S, He G, LaButti K, Lipzen A, Mondo S, Pangilinan J, Riley R, Salamov A, Simmons BA, Magnuson JK, Henrissat B, Mortensen UH, Larsen TO, de Vries RP, Grigoriev IV, Machida M, Baker SE, Andersen MR: **A comparative genomics study of 23 *Aspergillus* species from section Flavi**. *Nature Communications* 2020, **11**(1):1106.

[39]. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, Petersen B, Wang Z, Zhou Q, Diekhans M, Chen W, Andreu-Sánchez S, Margaryan A, Howard JT, Parent C, Pacheco G, Sinding M-HS, Puetz L, Cavill E, Ribeiro ÂM, Eckhart L, Fjeldså J, Hosner PA, Brumfield RT, Christidis L, Bertelsen MF, Sicheritz-Ponten T, Tietze DT, Robertson BC, Song G, Borgia G, Claramunt S, Lovette IJ, Cowen SJ, Njoroge P, Dumbacher JP, Ryder OA, Fuchs J, Bunce M, Burt DW, Cracraft J, Meng G, Hackett SJ, Ryan PG, Jønsson KA, Jamieson IG, da Fonseca RR, Braun EL, Houde P, Mirarab S, Suh A, Hansson B, Ponnikas S, Sigeman H, Stervander M, Frandsen PB, van der Zwan H, van der Sluis R, Visser C, Balakrishnan CN, Clark AG, Fitzpatrick JW, Bowman R, Chen N, Cloutier A, Sackton TB, Edwards SV, Foote DJ, Shakya SB, Sheldon FH, Vignal A, Soares AER, Shapiro B, González-Solís J, Ferrer-Obiol J, Rozas J, Riutort M, Tigano A, Friesen V, Dalén L, Urrutia AO, Székely T, Liu Y, Campana MG, Corvelo A, Fleischer RC, Rutherford KM, Gemmell NJ, Dussex N, Mouritsen H, Thiele N, Delmore K, Liedvogel M, Franke A, Hoeppner MP, Krone O, Fudickar AM, Milá B, Ketterson ED, Fidler AE, Friis G, Parody-Merino ÁM, Battley PF, Cox MP, Lima NCB, Prosdocimi F, Parchman TL, Schlinger BA, Loiselle BA, Blake JG, Lim HC, Day LB, Fuxjager MJ, Baldwin MW, Braun MJ, Wirthlin M, Dikow RB, Ryder TB, Camenisch G, Keller LF, DaCosta JM, Hauber ME, Louder MIM, Witt CC, McGuire JA, Mudge J, Megna LC, Carling MD, Wang B, Taylor SA, Del-Rio G, Aleixo A, Vasconcelos ATR, Mello CV, Weir JT, Haussler D, Li Q, Yang H, Wang J, Lei F, Rahbek C, Gilbert MTP, Graves GR, Jarvis ED, Paten B, Zhang G: **Dense sampling of bird diversity increases power of comparative**

genomics. *Nature* 2020, **587**(7833):252-257.

[40]. Zhang Z, Guo Y, Yang F, Li J: **Pan-Genome analysis reveals functional divergences in gut-restricted *Gilliamella* and *Snodgrassella***. *Bioengineering* 2022, **9**(10):544.

[41]. Fernandez-Fueyo E, Ruiz-Dueñas FJ, Ferreira P, Floudas D, Hibbett DS, Canessa P, Larrondo LF, James TY, Seelenfreund D, Lobos S, Polanco R, Tello M, Honda Y, Watanabe T, Watanabe T, Ryu JS, Kubicek CP, Schmoll M, Gaskell J, Hammel KE, St. John FJ, Vanden Wymelenberg A, Sabat G, Splinter BonDurant S, Syed K, Yadav JS, Doddapaneni H, Subramanian V, Lavín JL, Oguiza JA, Perez G, Pisabarro AG, Ramirez L, Santoyo F, Master E, Coutinho PM, Henrissat B, Lombard V, Magnuson JK, Kües U, Hori C, Igarashi K, Samejima M, Held BW, Barry KW, LaButti KM, Lapidus A, Lindquist EA, Lucas SM, Riley R, Salamov AA, Hoffmeister D, Schwenk D, Hadar Y, Yarden O, de Vries RP, Wiebenga A, Stenlid J, Eastwood D, Grigoriev IV, Berka RM, Blanchette RA, Kersten P, Martinez AT, Vicuna R, Cullen D: **Comparative genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis**. *Proceedings of the National Academy of Sciences* 2012, **109**(14):5458-5463.

[42]. Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, Noguera DR, McMahon KD: **Ancestral genome reconstruction identifies the evolutionary basis for trait acquisition in polyphosphate accumulating bacteria**. *The ISME Journal* 2016, **10**(12):2931-2945.

[43]. Arumugam K, Bağcı C, Bessarab I, Beier S, Buchfink B, Górska A, Qiu G, Huson DH, Williams RBH: **Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data**. *Microbiome* 2019, **7**(1):61.

[44]. Rubio-Rincón FJ, Weissbrodt DG, Lopez-Vazquez CM, Welles L, Abbas B, Albertsen M, Nielsen PH, van Loosdrecht MCM, Brdjanovic D: **"*Candidatus* Accumulibacter delftensis": A clade IC novel polyphosphate-accumulating organism without denitrifying activity on nitrate**. *Water Research* 2019, **161**:136-151.

[45]. McDaniel EA, Moya-Flores F, Keene Beach N, Camejo PY, Oyserman BO, Kizaric M, Khor EH, Noguera DR, McMahon KD: **Metabolic differentiation of co-occurring Accumulibacter clades revealed through genome-resolved metatranscriptomics**. *mSystems* 2021, **6**(4):e0047421.

[46]. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M: **Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing**. *Nature Communications* 2021, **12**(1):2009.

[47]. Srinivasan VN, Li G, Wang D, Tooker NB, Dai Z, Onnis-Hayden A, Bott C, Dombrowski P, Schauer P, Pinto A, Gu AZ: **Oligotyping and metagenomics reveal distinct *Candidatus* Accumulibacter communities in side-stream versus conventional full-scale enhanced biological phosphorus removal (EBPR) systems**. *Water Research* 2021, **206**:117725.

[48]. Tian Y, Chen H, Chen L, Deng X, Hu Z, Wang C, Wei C, Qiu G, Wuertz S: **Glycine adversely affects enhanced biological phosphorus removal**. *Water Research* 2022, **209**:117894.

[49]. Deng X, Yuan J, Chen L, Chen H, Wei C, Nielsen PH, Wuertz S, Qiu G: **CRISPR-Cas phage defense systems and prophages in *Candidatus* Accumulibacter**. *Water Research* 2023, **235**:119906.

[50]. Albertsen M, McIlroy SJ, Stokholm-Bjerregaard M, Karst SM, Nielsen PH: **"*Candidatus* Propionivibrio aalborgensis": a novel glycogen accumulating organism abundant in full-scale enhanced biological phosphorus removal plants**. *Frontiers in Microbiology* 2016, **7**:1033.

[51]. Petriglieri F, Singleton C, Peces M, Petersen JF, Nierychlo M, Nielsen PH: **"*Candidatus* Dechloromonas phosphoritropha" and "*Ca.* D. phosphorivorans", novel polyphosphate accumulating organisms abundant in wastewater treatment systems**. *The ISME Journal* 2021, **15**(12):3605-3614.

[52]. 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 Research* 2015, **25**(7):1043-1055.

[53]. Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative genomics**. *Genome Biology* 2019, **20**(1):238.

[54]. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Research* 2002, **30**(7):1575-1584.

[55]. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: Improvements in performance and usability**. *Molecular Biology and Evolution* 2013, **30**(4):772-780.

[56]. Castresana J: **Selection of conserved blocks from multiple alignments for their sse in phylogenetic analysis**. *Molecular Biology and Evolution* 2000, **17**(4):540-552.

[57]. Shen W, Le S, Li Y, Hu F: **SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation**. *PLOS ONE* 2016, **11**(10):e0163962.

[58]. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R: **IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era**. *Molecular Biology and Evolution* 2020, **37**(5):1530-1534.

[59]. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation**. *Nucleic Acids Research* 2021, **49**(W1):W293-W296.

[60]. Zhang A-N, Mao Y, Wang Y, Zhang T: **Mining traits for the enrichment and isolation of not-yet-cultured populations**. *Microbiome* 2019, **7**(1):96.

[61]. Csűös M: **Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood**. *Bioinformatics* 2010, **26**(15):1910-1912.

[62]. Pál C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer**. *Nature genetics* 2005, **37**(12):1372-

1375.

[63]. Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon K, Bertilsson S, Stepanauskas R, Andersson SG: **Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade**. *Genome Biology* 2013, **14**(11):1-14.

[64]. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Research* 2013, **42**(D1):D199-D205.

[65]. Ravenhall M, Škunca N, Lassalle F, Dessimoz C: **Inferring horizontal gene transfer**. *PLOS Computational Biology* 2015, **11**(5):e1004095.

[66]. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Research* 2006, **35**(suppl_1):D61-D65.

[67]. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.

[68]. Kopylova E, Noé L, Touzet H: **SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data**. *Bioinformatics* 2012, **28**(24):3211-3217.

[69]. Bushnell B: **BBMap: A fast, accurate, splice-aware aligner**. In: *Conference: 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA, March 17-20, 2014; United States*. DE-AC02-05CH11231 2016-04-08: 2014: Medium: ED.

[70]. Kumar L, M EF: **Mfuzz: a software package for soft clustering of microarray data**. *Bioinformation* 2007, **2**(1):5-7.

[71]. Shang Y, Wang X, Chen Z, Lyu Z, Lin Z, Zheng J, Wu Y, Deng Q, Yu Z, Zhang Y, Qu D: ***Staphylococcus* aureus PhoU homologs regulate persister formation and virulence**. *Frontier of Microbiology* 2020, **11**:865.

[72]. Chen L, Wei G, Zhang Y, Wang K, Wang C, Deng X, Li Y, Xie X, Chen J, Huang F, Chen H, Zhang B, Wei C, Qiu G: ***Candidatus* Accumulibacter use fermentation products for enhanced biological phosphorus removal**. *Water Research* 2023, **246**:120713.

[73]. Zhou Y, Pijuan M, Zeng RJ, Yuan Z: **Involvement of the TCA cycle in the anaerobic metabolism of polyphosphate accumulating organisms (PAOs)**. *Water Research* 2009, **43**(5):1330-1340.

[74]. Capra EJ, Laub MT: **Evolution of two-component signal transduction systems**. *Annual Review of Microbiology* 2012, **66**:325-347.

[75]. Alm E, Huang K, Arkin A: **The evolution of two-component systems in bacteria reveals different strategies for niche adaptation**. *PLOS Computational Biology* 2006, **2**(11):e143.

[76]. Ulrich LE, Zhulin IB: **The MiST2 database: a comprehensive genomics resource on microbial signal transduction**. *Nucleic Acids Research* 2010, **38**(Database issue):D401-407.

[77]. Elsen S, Swem LR, Swem DL, Bauer CE: **RegB/RegA, a highly conserved redox-responding global two-component regulatory system**. *Microbiology*

and Molecular Biology Reviews 2004, **68**(2):263-279.

[78]. Elsen S, Dischert W, Colbeau A, Bauer CE: **Expression of uptake hydrogenase and molybdenum nitrogenase in *Rhodobacter capsulatus* is coregulated by the RegB-RegA two-component regulatory system**. *Journal of Bacteriology* 2000, **182**(10):2831-2837.

[79]. Dubbs JM, Bird TH, Bauer CE, Tabita FR: **Interaction of CbbR and RegA* transcription regulators with the *Rhodobacter sphaeroides* cbb promoter-operator region ***. *Journal of Biological Chemistry* 2000, **275**(25):19224-19230.

[80]. Santos-Beneit F: **The Pho regulon: a huge regulatory network in bacteria**. *Frontiers in Microbiology* 2015, **6**:402.

[81]. Choi S, Jeong G, Choi E, Lee E-J: **A dual regulatory role of the PhoU protein in *Salmonella Typhimurium***. *mBio* 2022, **13**(3):e00811-00822.

[82]. Monzon V, Paysan-Lafosse T, Wood V, Bateman A: **Reciprocal best structure hits: using AlphaFold models to discover distant homologues**. *Bioinformatics Advances* 2022, **2**(1):vbac072.

[83]. diCenzo GC, Sharthiya H, Nanda A, Zamani M, Finan TM: **PhoU allows rapid adaptation to high phosphate concentrations by modulating PstSCAB transport rate in *Sinorhizobium meliloti***. *Journal of Bacteriology* 2017, **199**(18).

[84]. Li Y, Zhang Y: **PhoU is a persistence switch involved in persister formation and tolerance to multiple antibiotics and stresses in *Escherichia coli***. *Antimicrobial Agents and Chemotherapy* 2007, **51**(6):2092-2099.

[85]. Herbst FA, Dueholm MS, Wimmer R, Nielsen PH: **The proteome of *Tetrasphaera elongata* is adapted to changing conditions in wastewater treatment plants**. *Proteomes* 2019, **7**(2).

[86]. Willsky GR, Malamy MH: **Characterization of two genetically separable inorganic phosphate transport systems in *Escherichia coli***. *Journal of Bacteriology* 1980, **144**(1):356-365.

[87]. Martín JF, Liras P: **Molecular mechanisms of phosphate sensing, transport and signalling in *Streptomyces* and related *Actinobacteria***. *International Journal of Molecular Sciences* 2021, **22**(3).

[88]. Gebhard S, Tran SL, Cook GM: **The Phn system of *Mycobacterium smegmatis*: a second high-affinity ABC-transporter for phosphate**. *Microbiology* 2006, **152**(11):3453-3465.

[89]. Saunders AM, Mabbett AN, McEwan AG, Blackall LL: **Proton motive force generation from stored polymers for the uptake of acetate under anaerobic conditions**. *FEMS Microbiology Letters* 2007, **274**(2):245-251.

[90]. Wang Z, Guo F, Mao Y, Xia Y, Zhang T: **Metabolic characteristics of a glycogen-accumulating organism in *Defluviicoccus* cluster II revealed by comparative genomics**. *Microbial Ecology* 2014, **68**(4):716-728.

[91]. McIlroy SJ, Albertsen M, Andresen EK, Saunders AM, Kristiansen R, Stokholm-Bjerregaard M, Nielsen KL, Nielsen PH: **'*Candidatus Competibacter*'-lineage genomes retrieved from metagenomes reveal**

functional metabolic diversity. *The ISME Journal* 2014, **8**(3):613-624.

Figure 1. A phylogenetic tree of 43 Rhodocyclaceae members was built based on the concatenation of 59 single-copy genes. The genomes in red were recovered from our lab-scale reactors. SSA1, SSB1, and SCUT-1 were recovered in our previous work [43, 48]. SCELSE-5, SCELSE-7, SCELSE-10, and SCUT-2 were recovered from three of our lab-scale EBPR reactors (Supplementary Materials). The purple bars represent the number of shared orthogroups. The gray bars represent the number of unassigned genes.

Figure 2. a–**b**, The number of gene clusters at different frequencies in the pan Rhodocyclaceae genome (**a**) and the pan *Ca.* Accumulibacter genome (**b**). **c**–**d**, The proportion of clusters at different frequencies in the pan Rhodocyclaceae genome (**c**) and the pan *Ca.* Accumulibacter genome (**d**). **e**–**f**, The proportion of different average gene copies per genome in the pan Rhodocyclaceae genome (**e**) and the pan *Ca.* Accumulibacter genome (**f**). In each orthogroup, the average gene copies per genome are defined as the number of genes divided by the number of genomes.

Figure 3. **a**, Using the genome integrity estimate, about 99.94% of the core genes could be identified with a cut-off value of 18. Only gene families that appear in ≥18 *Ca.* Accumulibacter genomes are considered core genes. **b**, A Venn diagram describing the numbers of core genes and lineage-specific genes in the pan *Ca.* Accumulibacter genome. **c**, The number of core genes observed at different cutoff values. **d**, The number of genes assigned as ancestral, derived, lineage-specific, and flexible genes in SCUT-2

1011 and UW1.

1012

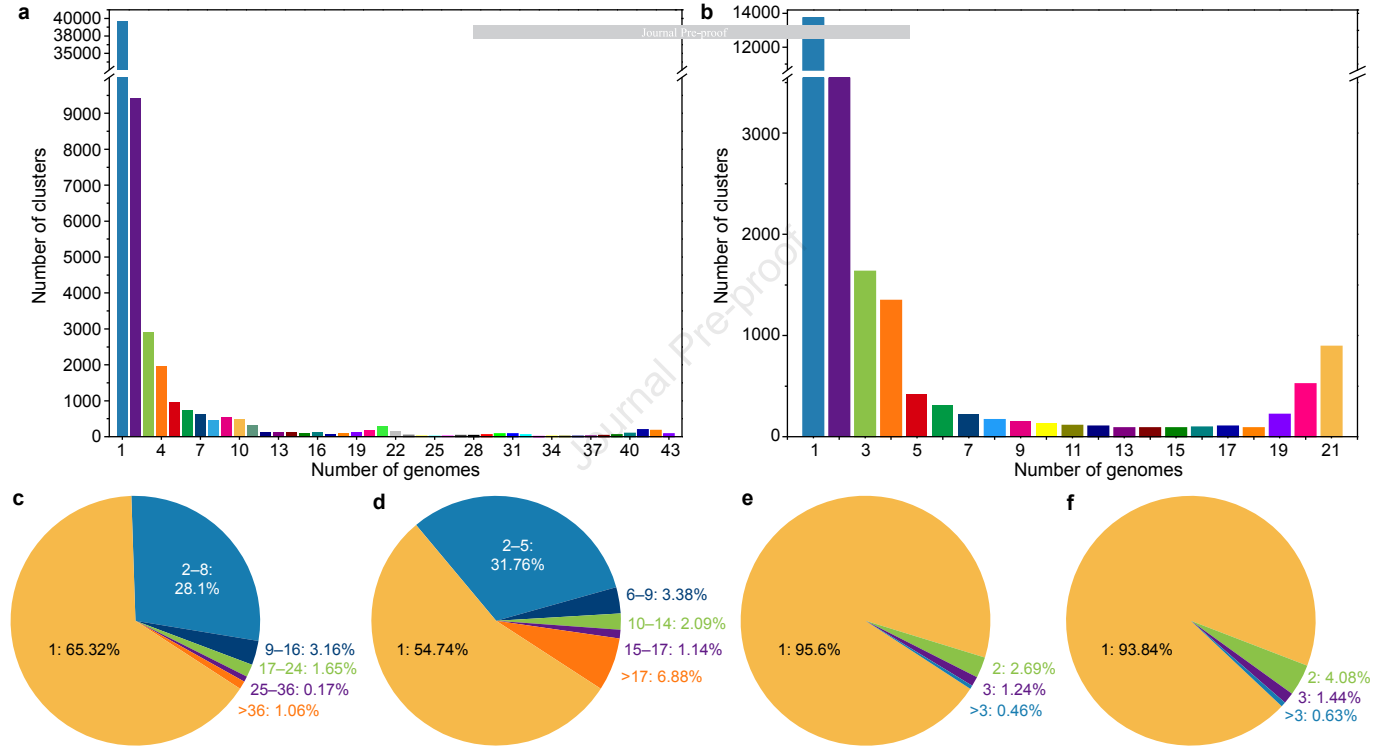
1013 **Figure 4.** Gain or loss of genes at various nodes of the *Ca.* Accumulibacter lineage. A
1014 maximum likelihood tree was built based on the concatenation of single-copy genes
1015 with model Q. insect+F+I+I+R4. Genomes in red are those recovered from our
1016 bioreactors [7, 43, 48, 49].

1017

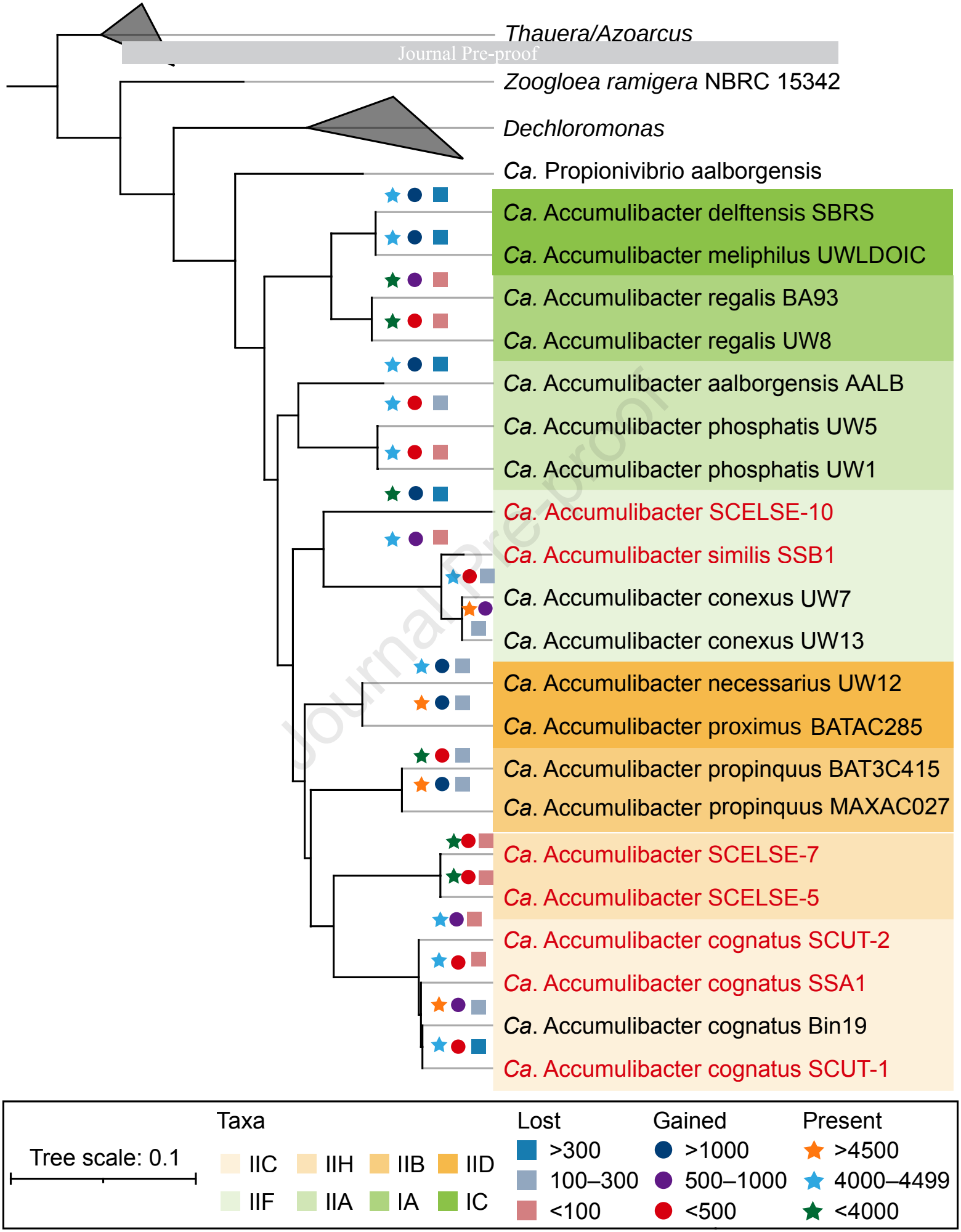
1018 **Figure 5**. **a**, The ratio of ancestral, derived, lineage-specific , and flexible genes in
1019 different primary metabolic pathways (MAP) of SCUT-2. **b**, The number of ancestral
1020 and derived genes in representative secondary MAP of SCUT-2.

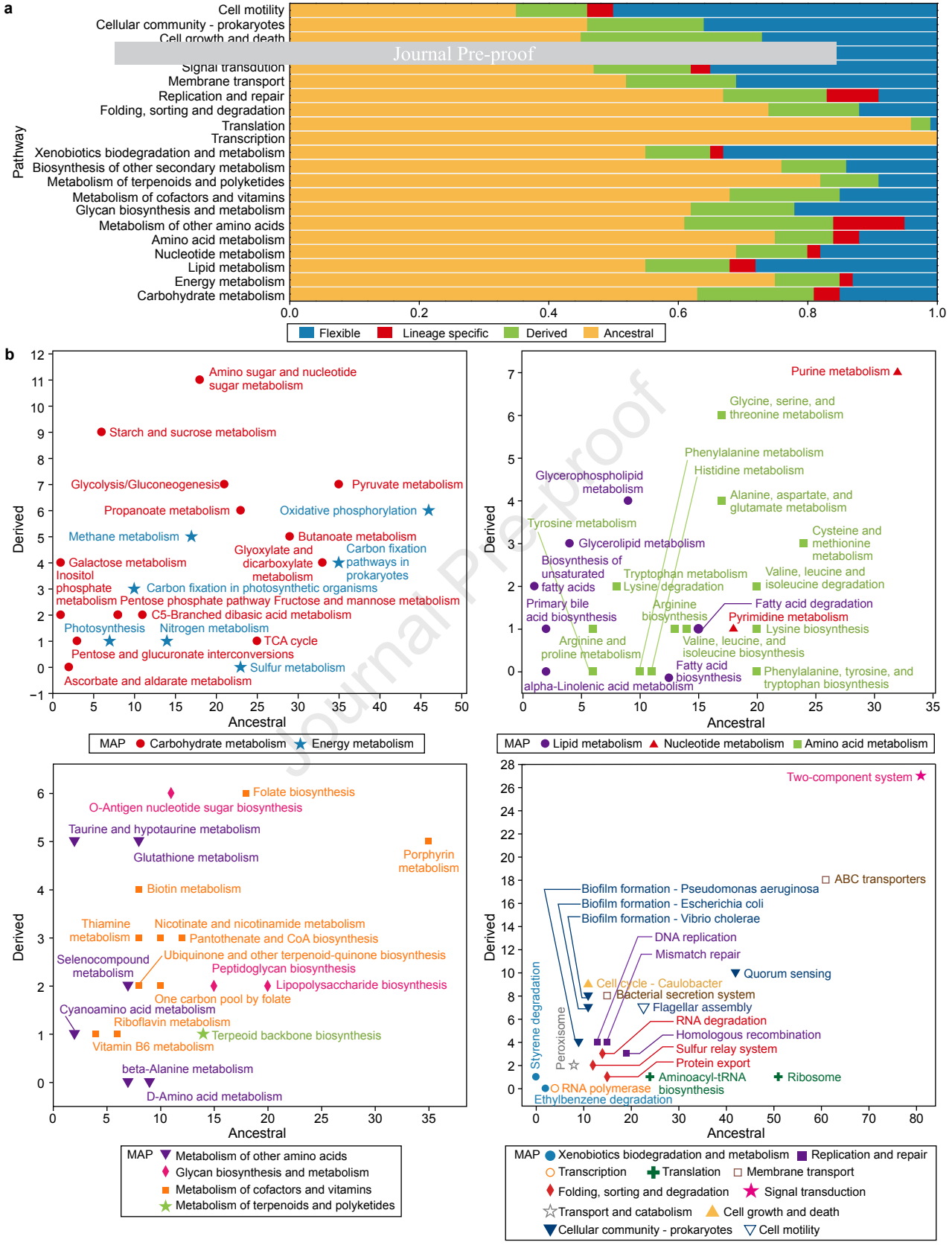
1021

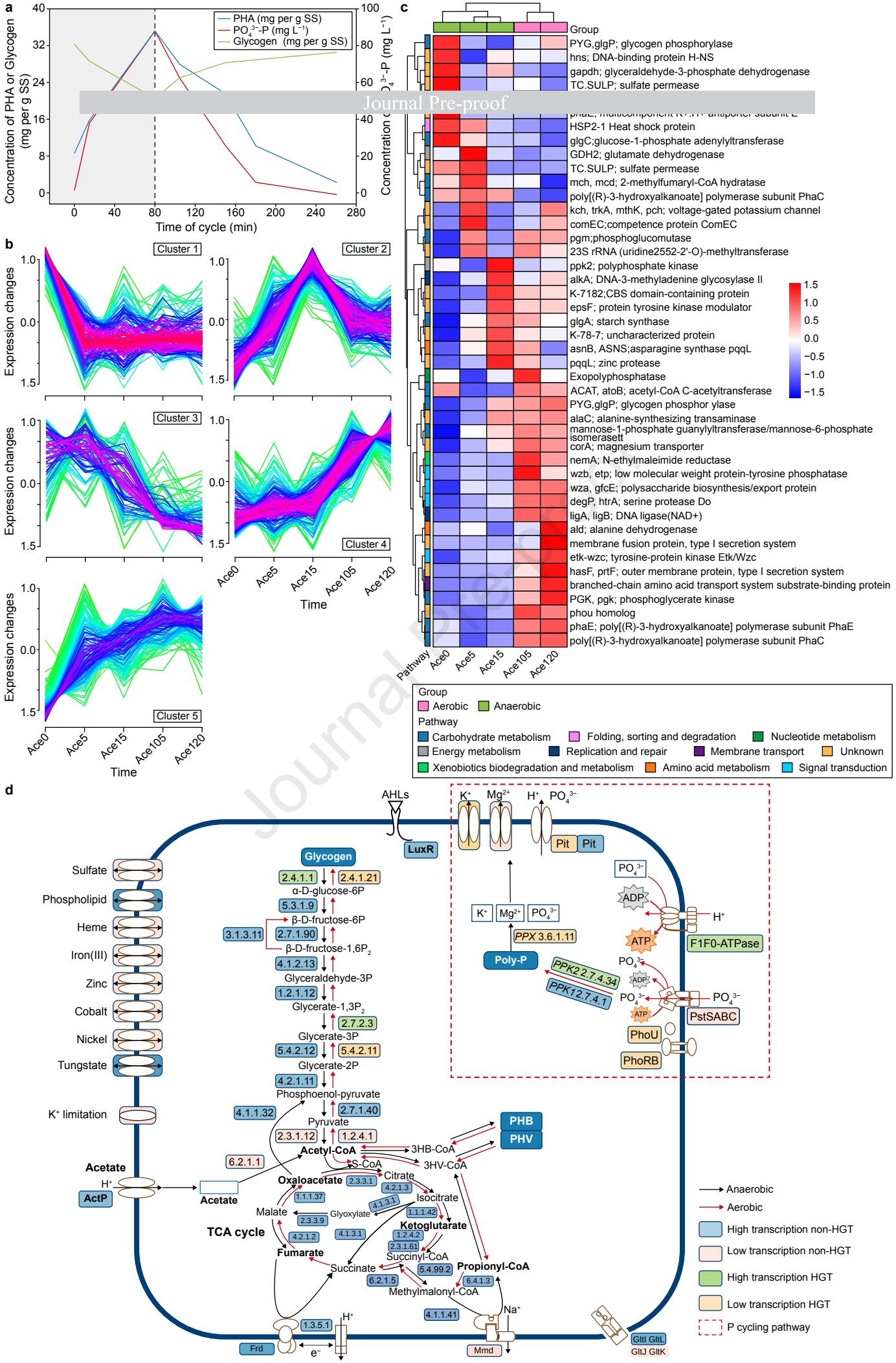
1022 **Figure 6. a**, Changes in phosphate, PHA, and glycogen concentrations during an
1023 anaerobic-aerobic full cycle. **b**, Cluster analysis of transcriptome data at different time
1024 points for transcription pattern identification. **c**, 44 highly transcribed and laterally
1025 derived genes (via HGT) in the SCUT-2 genome during the anaerobic-aerobic full cycle.
1026 **d**, A metabolic model of *Ca*. Accumulibacter. Black and red solid arrows represent
1027 active metabolic pathways in the anaerobic and aerobic phases. Genes in blue and
1028 pink are genes not acquired via HGT with high and low transcription, respectively.
1029 Genes in green and yellow represent genes acquired via HGT with high and low
1030 transcription, respectively. The red dashed line denotes the key P cycling pathway.
1031 The enzyme commission (EC) number indicates the key enzyme involved in each
1032 pathway/reaction.



a
Core genes (%)
Genomes
99.94
99.30
93.95
66.29
b
IA
IB
IC
IIA
IIB
IIC
IID
IIE
IIF
IIH
629
328
2659
1990
1600
1280
1969
1890
Core
1725
c
Number of core genes
Number of additions
Number of core genes (×10³)
Number of additions (×10³)
Genomes
d
UW1
SCUT-2
Number of genes (×10²)
Ancestral
Derived
Lineage specific
Flexible
Gene classification




a
Pathway
Cell motility
Cellular community - prokaryotes
Cell growth and death
Signal transduction
Membrane transport
Replication and repair
Folding, sorting and degradation
Translation
Transcription
Xenobiotics biodegradation and metabolism
Biosynthesis of other secondary metabolism
Metabolism of terpenoids and polyketides
Metabolism of cofactors and vitamins
Glycan biosynthesis and metabolism
Metabolism of other amino acids
Amino acid metabolism
Nucleotide metabolism
Lipid metabolism
Energy metabolism
Carbohydrate metabolism
0.0
0.2
0.4
0.6
0.8
1.0
Flexible
Lineage specific
Derived
Ancestral
Journal Pre-proof
b
Derived
Ancestral
Amino sugar and nucleotide sugar metabolism
Starch and sucrose metabolism
Glycolysis/Gluconeogenesis
Pyruvate metabolism
Propanoate metabolism
Oxidative phosphorylation
Methane metabolism
Butanoate metabolism
Galactose metabolism
Glyoxylate and dicarboxylate metabolism
Carbon fixation pathways in prokaryotes
Inositol phosphate metabolism
Carbon fixation in photosynthetic organisms
Pentose phosphate pathway
Fructose and mannose metabolism
C5-Branched dibasic acid metabolism
Photosynthesis
Nitrogen metabolism
TCA cycle
Pentose and glucuronate interconversions
Sulfur metabolism
Ascorbate and aldarate metabolism
MAP
Carbohydrate metabolism
Energy metabolism
Purine metabolism
Glycine, serine, and threonine metabolism
Phenylalanine metabolism
Histidine metabolism
Glycerophospholipid metabolism
Alanine, aspartate, and glutamate metabolism
Tyrosine metabolism
Glycerolipid metabolism
Cysteine and methionine metabolism
Biosynthesis of unsaturated fatty acids
Tryptophan metabolism
Lysine degradation
Valine, leucine and isoleucine degradation
Primary bile acid biosynthesis
Arginine biosynthesis
Fatty acid degradation
Pyrimidine metabolism
Lysine biosynthesis
Arginine and proline metabolism
Valine, leucine, and isoleucine biosynthesis
Fatty acid biosynthesis
Phenylalanine, tyrosine, and tryptophan biosynthesis
alpha-Linolenic acid metabolism
MAP
Lipid metabolism
Nucleotide metabolism
Amino acid metabolism
Folate biosynthesis
O-Antigen nucleotide sugar biosynthesis
Taurine and hypotaurine metabolism
Glutathione metabolism
Porphyrin metabolism
Biotin metabolism
Thiamine metabolism
Nicotinate and nicotinamide metabolism
Pantothenate and CoA biosynthesis
Selenocompound metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Peptidoglycan biosynthesis
Lipopolysaccharide biosynthesis
One carbon pool by folate
Cyanoamino acid metabolism
Riboflavin metabolism
Terpeoid backbone biosynthesis
Vitamin B6 metabolism
beta-Alanine metabolism
D-Amino acid metabolism
MAP
Metabolism of other amino acids
Glycan biosynthesis and metabolism
Metabolism of cofactors and vitamins
Metabolism of terpenoids and polyketides
Two-component system
ABC transporters
Biofilm formation - Pseudomonas aeruginosa
Biofilm formation - Escherichia coli
Biofilm formation - Vibrio cholerae
DNA replication
Mismatch repair
Quorum sensing
Cell cycle - Caulobacter
Bacterial secretion system
Flagellar assembly
RNA degradation
Homologous recombination
Sulfur relay system
Protein export
Aminoacyl-tRNA biosynthesis
Ribosome
RNA polymerase
Styrene degradation
Peroxisome
Ethylbenzene degradation
MAP
Xenobiotics biodegradation and metabolism
Replication and repair
Transcription
Translation
Membrane transport
Folding, sorting and degradation
Signal transduction
Transport and catabolism
Cell growth and death
Cellular community - prokaryotes
Cell motility



Highlights

- 298 core genes were acquired by *Ca.* Accumulibacter at their least common ancestor
- 124 of these core genes were obtained via horizontal gene transfer (HGT)
- *pho*R, *pho*B, *pho*U homologs, *pit* and *ppk*2 in *Ca.* Accumulibacter were laterally derived
- Incompatible transcriptions of *pho*R, *pho*B, *pho*U, and *pst* were observed
- Pho regulon disorder may be a key to the P accumulating trait of *Ca.* Accumulibacter

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: