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/](https://doi.org/10.1016/j.ese.2023.100353) [j.ese.2023.100353](https://doi.org/10.1016/j.ese.2023.100353).

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published

in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences.

**Integrated genomics provide insights for t accumulation trait of *Ca.* Accumulibacter**

**he evolution of the polyphosphate**

**02 Orthologue analyses 04 Gene flux analysis**

p

p

i

i 100 IA 25 14

Number of core genes (×103)

pi *Thauera/Azoarcus*

Number of core genes Number of additions

Number of additions (×103)

Number of genes (×102)

p

i *Zoogloea ramigera* NBRC 15342

*Dechloromonas*

95 99.94

99.30

Core genes (%)

90 93.95

IC IIH 20

629

20

12

10 15

UW1 SCUT-2

pi *Ca.* Propionivibrio aalborgensis 85

p

i *Ca.* Accumulibacter delftensis SBRS 80

*Ca.* Accumulibacter meliphilus UWLDOIC 75

IIA

1890

1969

Core 1725

328

2659

15

IIF 10

8

6 10

pi

p

i

PAO Non-PAO

*Ca.* Accumulibacter regalis BA93

*Ca.* Accumulibacter regalis UW8

*Ca.* Accumulibacter aalborgensis AALB *Ca.* Accumulibacter phosphatis UW5 *Ca.* Accumulibacter phosphatis UW1

70

65 66.29

60

20 15 10 5 0

1280 1990

1600

IIB IID

IIC

5 4

2

0

0

20 15 10 5

5

0

Ancestral Derived Lineage Flexible

pi

Surface

Inner membrane

Cytoplasm

**What make us different?**

pi

PstS

HGT

*Ca.* Accumulibacter SCELSE-10 *Ca.* Accumulibacter similis SSB1 *Ca.* Accumulibacter conexus UW7 *Ca.* Accumulibacter conexus UW13

*Ca.* Accumulibacter necessarius UW12

*Ca.* Accumulibacter proximus BATAC285 *Ca*. Accumulibacter propinquus BAT3C415 *Ca*. Accumulibacter propinquus MAXAC027

*Ca*. Accumulibacter SCELSE-7

*Ca*. Accumulibacter SCELSE-5

Genomes

**05 Metablic function analysis**

Cell motility Cellular community - prokaryotes

Cell growth and death Transport and catabolism

Signal transdution Membrane transport Replication and repair

Genomes

specific

Gene classification

12

Amino sugar and nucleotide

11 sugar metabolism

10

9 Starch and sucrose metabolism 8

PhoU

PstB

Ps Ps

tC tA

PhoR

PhoB

PPK2

(poly-P) (poly-P)

n

n−1

Folding, sorting and degradation

Translation Transcription

*Ca*. Accumulibacter cognatus SCUT-2 *Ca*. Accumulibacter cognatus SSA1 *Ca*. Accumulibacter cognatus Bin19

*Ca*. Accumulibacter cognatus SCUT-1

Pathway

Xenobiotics biodegradation and metabolism

7 Glycolysis/Gluconeogenesi

6 Propanoate metabolism

Derived

5 Methane metabolism

Pyruvate metabolism Oxidative phosphorylation

Butanoate metabolism

GDP/ADP GTP/ATP

Biosynthesis of other secondary metabolism Metabolism of terpenoids and polyketides

4 Galactose metabolism

Glyoxylate and dicarboxylate

Carbon fixation pathways in

Metabolism of cofactors and vitamins

Inositol

3 phosphate

metabolism prokaryotes

Carbon fixation in photosynthetic organisms

Glycan biosynthesis and metabolism Metabolism of other amino acids Amino acid metabolism Nucleotide metabolism

Lipid metabolism Energy metabolism Carbohydrate metabolism

metabolism Pentose phosphate pathway Fructose and mannose metabolism

2 C5-Branched dibasic acid metabolism Photosynthesis Nitrogen metabolism

1 TCA cycle

Pentose and glucuronate interconversions

0 Sulfur metabolism

Ascorbate and aldarate metabolism

−1

5 10 15 20 25 30 35 40 45 50

**01 Data acquisition and evaluation**

**03 Pangenome analysis**

0.0 0.2 0.4

0.6 0.8 1.0 0

Ancestral

40000

38000

35000

14000

MAP Carbohydrate metabolism Energy metabolism

Flexible Lineage specific Derived Ancestral

12000

**06 Horizontal gene transfer 07 Metatranscriptomic analysis**

9000

8000

Number of clusters

7000

3000

Sulfate

Number of clusters

**Glycogen**

AHLs

**LuxR**

K+ Mg2+

H+ PO 3−

4

Pit Pit

PO 3−

4

1.0

0.0

Expression changes

Cluster 1 Cluster 2

1.0

0.0

Group

PYG,glgP; glycogen phosphorylase hns; DNA-binding protein H-NS

gapdh; glyceraldehyde-3-phosphate dehydrogenase TC.SULP; sulfate permease

HSP2-2 Heat shock protein

phaE; multicomponent K+:H+ antiporter subunit E HSP2-1 Heat shock protein

2.4.1.1

2.4.1.21

glgC;glucose-1-phosphate adenylyltransferase

6000

5000

4000

3000

2000

Phospholipid

Heme Iron(III) Zinc

α-D-glucose-6P

5.3.1.9

β-D-fructose-6P

3.1.3.11 2.7.1.90

β-D-fructose-1,6P2

4.1.2.13

Glyceraldehyde-3P

K+ Mg2+ PO 3−

*PPX* 3.6.1.11

4

**Poly-P**

ADP

ATP PO 3−

4

ADP

H+

F1F0-ATPase

1.5

1.0

Expression changes

1.5

1.0

Cluster 3

GDH2; glutamate dehydrogenase TC.SULP; sulfate permease

mch, mcd; 2-methylfumaryl-CoA hydratase

poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaC kch, trkA, mthK, pch; voltage-gated potassium channel comEC;competence protein ComEC pgm;phosphoglucomutase

23S rRNA (uridine2552-2'-O)-methyltransferase ppk2; polyphosphate kinase

1000

1.2.1.12

PO 3− PO 3−

alkA; DNA-3-methyladenine glycosylase Il

1.5

2000

Cobalt

Nickel

Glycerate-1,3P2

2.7.2.3

Glycerate-3P

4

ATP

PhoU

4

PstSABC

K-7182;CBS domain-containing protein epsF; protein tyrosine kinase modulator glgA; starch synthase

K-78-7; uncharacterized protein

1.0

0.5

0.0

1000

5.4.2.12

5.4.2.11

0.0

0.0

asnB, ASNS;asparagine synthase pqqL pqqL; zinc protease

−0.5

0 1 4 7 10 13 16 19 22 25 28 31 34 37 40 43

0

1 3 5 7 9 11 13 15 17 19 21

Tungstate

Glycerate-2P

4.2.1.11

Phosphoenol-pyruvate

PhoRB

Exopolyphosphatase

ACAT, atoB; acetyl-CoA C-acetyltransferase PYG,glgP; glycogen phosphor ylase

alaC; alanine-synthesizing transaminase

−1.0

−1.5

Number of genomes

Number of genomes

K+ limitation

4.1.1.32

2.7.1.40

Pyruvate

**PHB**

1.5

1.5

Cluster 4

mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerasett

corA; magnesium transporter

nemA; N-ethylmaleimide reductase

2.3.1.12

1.2.4.1

**PHV**

wzb, etp; low molecular weight protein-tyrosine phosphatase

2–8:

2–5:

31.76%

**Acetate**

H+

**ActP**

6.2.1.1

**Acetate**

**Acetyl-CoA**

S-CoA

**Oxaloacetate** 2.3.3.1

1.1.1.37

3HB-CoA

3HV-CoA

Citrate

4.2.1.3

Isocitrate

Anaerobic

1.0

Time

wza, gfcE; polysaccharide biosynthesis/export protein degP, htrA; serine protease Do

ligA, ligB; DNA ligase(NAD+) ald; alanine dehydrogenase

membrane fusion protein, type I secretion system etk-wzc; tyrosine-protein kinase Etk/Wzc

28.1%

6–9: 3.38%

Malate

**TCA cycle**

2.3.3.9

4.2.1.2

Glyoxylate

4.1.3.1

1.1.1.42

**Ketoglutarate**

1.2.4.2

Aerobic

High transcription non-HGT

0.0

hasF, prtF; outer membrane protein, type I secretion system

branched-chain amino acid transport system substrate-binding protein PGK, pgk; phosphoglycerate kinase

1: 65.32%

9–16: 3.16%

1: 54.74%

10–14: 2.09%

15–17: 1.14%

1: 95.6%

1: 93.84%

**Fumarate**

Succinate

2.3.1.61

Succinyl-CoA

5.4.99.2 **Propionyl-CoA**

Low transcription non-HGT

phou homolog

phaE; poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaE poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaC

17–24: 1.65%

2: 2.69%

6.2.1.5

6.4.1.3

High transcription HGT

25–36: 0.17%

>36: 1.06%

>17: 6.88%

3: 1.24%

>3: 0.46%

2: 4.08%

3: 1.44%

>3: 0.63%

H+

1.3.5.1

Methylmalonyl-CoA

4.1.1.41

Na+

Expression changes

GltI GltL

Low transcription HGT P cycling pathway

1.5

Cluster 5

Group

Aerobic Anaerobic Pathway

Carbohydrate metabolism

Folding, sorting and degradation

Nucleotide metabolism

Frd e−

Mmd

GltJ GltK

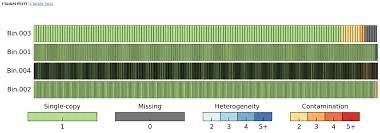
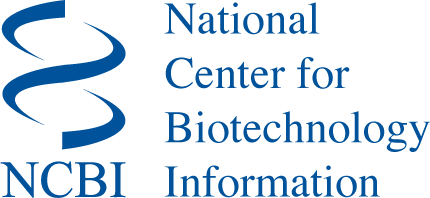
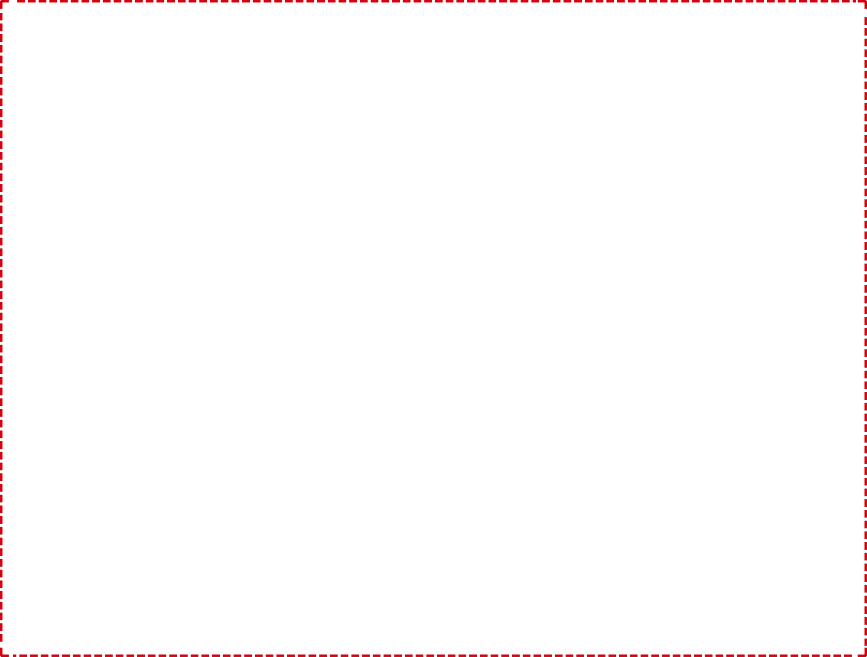
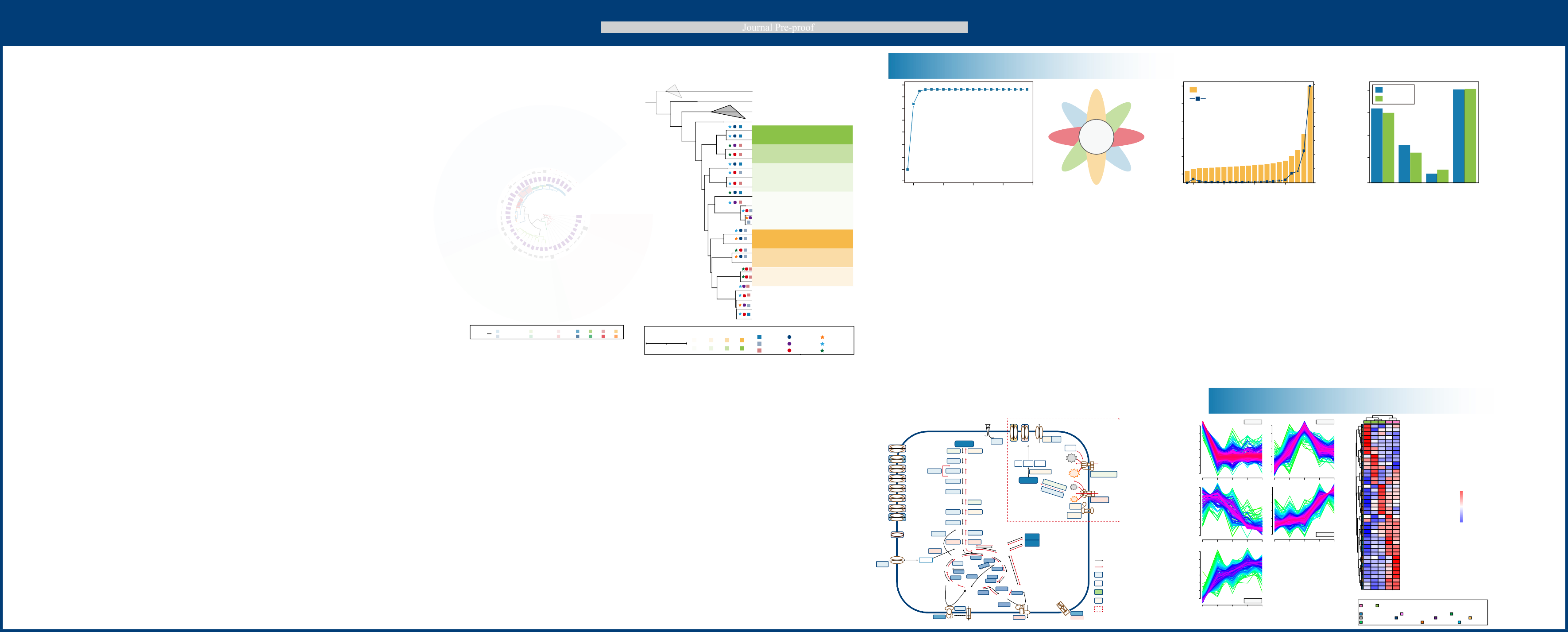
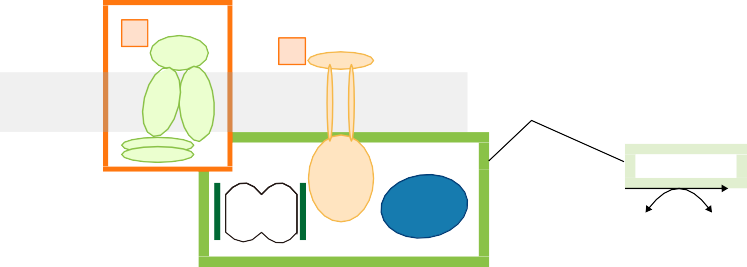
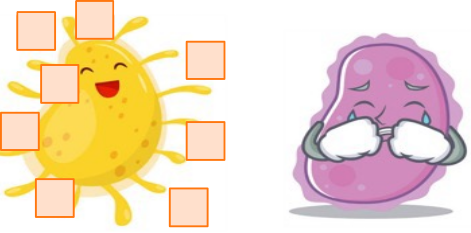
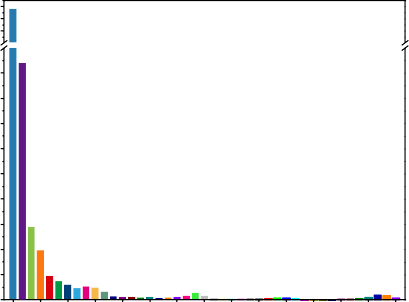
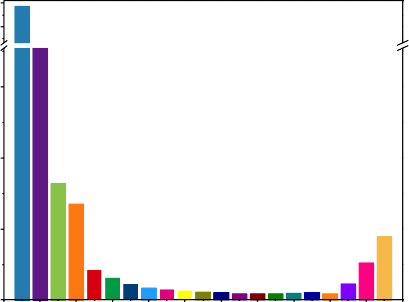
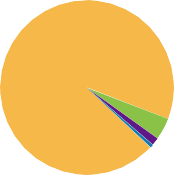
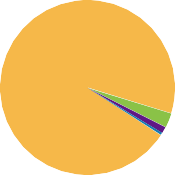
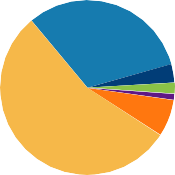
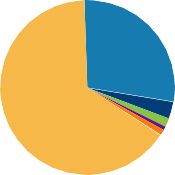
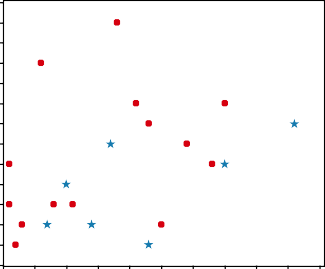
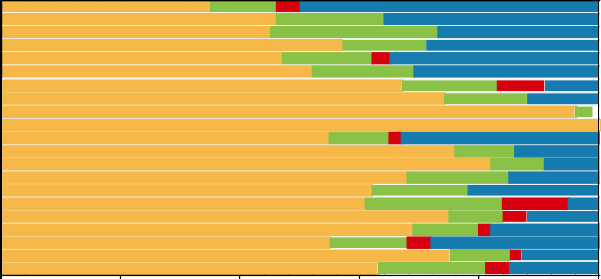
Energy metabolism

Replication and repair

Membrane transport

Unknown

Xenobiotics biodegradation and metabolism



Tree scale 0.07

Taxa

*Ca.* Accumulibacter

*Propionivibrio*

Clades

*Dechloromonas Thauera* IIC IIB IIF IA

*Zoogloea*

*Azoarcus* IIH IID IIA IC

Taxa

IIC IIF

Lost

Gained

Present

Tree scale: 0.1

IIH IIB IID

IIA IA IC

>300 >1000 >4500

100–300 500–1000 4000–4499

<100 <500 <4000

Amino acid metabolism

Signal transduction

# Integrated genomics provides insights into the evolution of the polyphosphate

1. **accumulation trait of *Ca.* Accumulibacter**
2. Xiaojing Xiea, 1, Xuhan Denga, 1, Liping Chena, Jing Yuana, Hang Chena, Chaohai Wei
3. a,c, Xianghui Liub,c, Stefan Wuertzb,c\*, Guanglei Qiua,b,d,e\*
4. a *School of Environment and Energy, South China University of Technology,*
5. *Guangzhou 510006, China.*
6. b *Singapore Centre for Environmental Life Sciences Engineering, Nanyang*
7. *Technological University, Singapore 637551, Singapore.*
8. c *School of Civil and Environmental Engineering, Nanyang Technological University,*
9. *Singapore 639798, Singapore*
10. d *Guangdong Provincial Key Laboratory of Solid Wastes Pollution Control and*
11. *Recycling, Guangzhou 510006, China*
12. e *The Key Lab of Pollution Control and Ecosystem Restoration in Industry Clusters,*
13. *Ministry of Education, Guangzhou 510006, China*

15

1. \* Corresponding Author: [qiugl@scut.edu.cn](mailto:qiugl@scut.edu.cn) (G.Q.), [SWuertz@ntu.edu.sg](mailto:SWuertz@ntu.edu.sg) (S.W.)
2. 1 Authors contributed equally towards this study 18
3. **Abstract:** *Candidatus* Accumulibacter, a prominent polyphosphate-accumulating
4. organism (PAO) in wastewater treatment, plays a crucial role in enhanced biological
5. phosphorus removal (EBPR). The genetic underpinnings of its polyphosphate
6. accumulation capabilities, however, remain largely unknown. Here, we conducted a
7. comprehensive genomic analysis of *Ca*. Accumulibacter-PAOs and their relatives
8. within the Rhodocyclaceae family, identifying 124 core genes acquired via horizontal
9. gene transfer (HGT) at its least common ancestor. Metatranscriptomic analysis of an
10. enrichment culture of *Ca*. Accumulibacter revealed active transcription of 44 of these
11. genes during an EBPR cycle, notably including the polyphosphate kinase 2 (PPK2)
12. gene instead of the commonly recognized polyphosphate kinase 1 (PPK1) gene.
13. Intriguingly, the phosphate regulon (Pho) genes showed minimal transcriptions,
14. pointing to a distinctive fact of Pho dysregulation, where PhoU, the phosphate signaling
15. complex protein, was not regulating the high-affinity phosphate transport (Pst) system,
16. resulting in continuous phosphate uptake. To prevent phosphate toxicity, *Ca*.
17. Accumulibacter utilized the laterally acquired PPK2 to condense phosphate into
18. polyphosphate, resulting in the polyphosphate-accumulating feature. This study
19. provides novel insights into the evolutionary emergence of the polyphosphate-
20. accumulating trait in *Ca*. Accumulibacter, offering potential advancements in
21. understanding the PAO phenotype in the EBPR process.
22. **Keywords:** *Candidatus* Accumulibacter; Comparative genomics; Horizontal gene
23. transfer (HGT); PhoU; Polyphosphate kinase 2 (PPK2) 40

# 1 Introduction

1. With the rapid development of industry and the economy, there has been a significant
2. surge in wastewater generation. This escalating wastewater production has, in turn,
3. resulted in excessive phosphorus (P) discharge, leading to adverse consequences such
4. as eutrophication, water quality deterioration, and aquatic ecosystem degeneration [1-
5. 3]. Enhanced biological phosphorus removal (EBPR) is an environmentally friendly
6. and economical process widely applied in municipal wastewater treatment plants
7. (WWTPs) for P removal [4-10]. This process is mediated by a group of microorganisms,
8. namely polyphosphate-accumulating organisms (PAOs) [11-14]. *Candidatus*
9. Accumulibacter is a model genus of PAOs commonly found in lab- and full-scale EBPR
10. systems [15-18]. Under anaerobic conditions, *Ca.* Accumulibacter uses intracellularly
11. stored polyphosphate (poly-P) as an energy source to power the uptake of volatile fatty
12. acids (VFAs). This metabolic process results in the release of phosphate. The
13. assimilated VFAs are then polymerized and stored as polyhydroxyalkanoates (PHAs).
14. In the subsequent aerobic phase, PHAs are oxidized for cell metabolism and
15. reproduction. Excess phosphate is removed from the aquatic phase to synthesize poly-
16. P, achieving P removal [19-22]. This unique metabolic feature allows PAOs to thrive in
17. alternating anaerobic-aerobic conditions, conferring sustainable P removal. However,
18. the key genetic basis affording PAOs the ability to P cycling is unclear. Genes known
19. to be indispensable for the P cycling feature, e.g., the polyphosphate kinase gene (*ppk*)
20. and exopolyphosphatase gene (*ppx*) for poly-P synthesis and hydrolysis, respectively,
21. and the inorganic phosphate transporter gene (*pit*) and the high-affinity phosphate
22. transporter gene (*pst*) for phosphate transport, are widely preserved in the bacterial
23. domain, including in non-PAOs [23, 24]. Their presence does not guarantee the P
24. cycling ability, and the key genes have yet to be identified. The transition from non-
25. PAOs to PAOs may be driven by adaptive evolution [25, 26].
26. The need to understand the gain and loss of genes in different strains and the genome
27. diversification in a given lineage of organisms gave rise to pangenomics. A pangenome
28. encompasses the entire set of genes from all individuals of a specific lineage [27, 28].
29. Genes in a pangenome are divided into core genes and variable genes [29]. The
30. collection of genes commonly present in all individuals of a specific lineage is called
31. the core pangenome, representing the common genetic features of a microbial lineage
32. [30]. The variable genes can be further divided into unique genes (found in a single
33. strain/genome) and dispensable genes (shared in at least two but not all strains/genomes)
34. [31]. Dispensable genes represent the intra-lineage diversity encoded among different
35. members [29]. By avoiding single sample bias and ensuring full representation of
36. genomic diversity of different lineage members, the analysis of the pangenome provides
37. insight into the genetic basis of common phenotypic characteristics shared in a group
38. of bacteria, greatly improving our ability to solve complex phenotypic problems [32-
39. 34]. Comparative genomics has been applied to study the evolution and development
40. of many bacterial species [35-40]. Via comparative genomic analysis, Fernandez-Fueyo
41. et al. [41] found a subset of potentially important genes for selective lignin
42. decomposition in *Ceriporipsis subvermispora*.
43. Oyserman et al. [42] previously constructed a pangenome of the Rhodocyclaceae family
44. (including ten *Ca.* Accumulibacter and 16 out-group genomes) to explore the genetic
45. composition and evolutionary changes in metabolic pathways of the *Ca.*
46. Accumulibacter genus. However, at the time, limited numbers of *Ca.* Accumulibacter
47. genomes were available, with more than half having low completeness (<90%). The
48. deficiency in genome quality and quantity may result in an inadequate representation
49. of the lineage pangenome and affect the downstream analysis of genes. With the
50. advance in EBPR research and sequencing techniques, increasing numbers of high-
51. quality *Ca*. Accumulibacter genomes have been obtained [7, 14, 18, 43-49]. New PAOs
52. and glycogen-accumulating organisms (GAOs) were also identified in genera
53. phylogenetically closely related to *Ca*. Accumulibacter. GAOs occupy a similar
54. ecological niche as PAOs in EBPR systems. They use glycogen instead of
55. polyphosphate as an energy source for anaerobic carbon source uptake, thus competing
56. with PAOs. For instance, a *Propionivibrio* member was shown to perform as a GAO in
57. full-scale WWTPs in Denmark [50]. Two *Dechloromonas* members in the same
58. WWTPs (i.e., *Ca.* Dechloromonas phosphoritropha and *Ca.* Dechloromonas
59. phosphorivorans) were revealed to be PAOs [51]. The identification of *Dechloromonas*-
60. related PAOs raises the possibility that the emergence of the PAO phenotype may have
61. occurred before the *Ca.* Accumulibacter's last common ancestor (LCA). The evolution
62. in the P cycling feature needs to be re-evaluated and traced. Combined with the analysis
63. of gene transcriptional characteristics of representative PAO strains, the key genomic
64. characteristics distinguishing PAOs and non-PAOs may be further identified and
65. determined, which would significantly advance the understanding of the genomic basis
66. of the PAO phenotype.
67. To understand the emergence of the PAO phenotype of *Ca.* Accumulibacter, we selected
68. 43 high-quality genomes within the Rhodocyclaceae family for comparative genomic
69. analysis. A pangenome of the Rhodocyclaceae family, including 21 *Ca*. Accumulibacter
70. genomes, seven of which were recovered from our EBPR reactors, 22 out-group
71. genomes, including two confirmed *Dechloromonas* PAOs, i.e., *Ca.* Dechloromonas
72. phosphoritropha and *Ca.* Dechloromonas phosphorivorans [51], and one *Propionivibrio*
73. GAO genome, *Ca.* Propionivibrio aalborgensis [50], was constructed. In the analysis of
74. genes within the pangenome, genes were classified as ancestral, derived, flexible, or
75. lineage-specific genes. The dynamics in these genes in the evolutionary process were
76. analyzed, and metatranscriptomic analyses were performed on an enrichment culture of
77. *Ca*. Accumulibacter Clade IIC SCUT-2 for identifying their active genes in a typical
78. anaerobic-aerobic cycle to narrow down the range of genes important for the PAO
79. phenotype of *Ca*. Accumulibacter. Genomic comparisons were further performed
80. between *Ca*. Accumulibacter, two *Dechloromonas*-related PAOs, and the
81. *Propionivibrio* GAO. Among the numerous genes investigated, two key players
82. emerged: the phosphate signaling complex protein gene (*pho*U) in the Pho regulon and
83. the laterally derived polyphosphate kinase 2 gene (*ppk*2). These genes were identified
84. as instrumental in the emergence of the PAO phenotype of *Ca.* Accumulibacter. This
85. study provides new insights into the development of the P cycling trait of *Ca.*
86. Accumulibacter.

# 2 Materials and Methods

1. **2.1 Data acquisition and evaluation**
2. The genomes used for analysis included seven high-quality genomes recovered from
3. our EBPR reactors and 36 genomes obtained from the National Center for
4. Biotechnology Information (NCBI) database. All 43 genomes belong to the
5. Rhodocyclaceae family, including 21 *Ca.* Accumulibacter genomes and 22 out-group
6. genomes (ten *Dechloromonas*, seven *Thauera*, three *Azoarcus*, one *Propionivibrio*, and
7. one *Zooglea ramigera* genomes). The completeness and contamination of the genomes
8. were evaluated using CheckM [52]. The GenBank assembly accession, corresponding
9. species names, and additional details about the qualities of these genomes can be found
10. in the Supplementary Materials Table S1–S3.

# 2.2 Orthologue analyses

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

# 2.3 Phylogenetic analysis of pangenome

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

# 2.4 Pangenome analysis

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

# 2.5 Gene gain/loss analysis

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

# 2.6 Metabolic function analysis

1. The ancestral, derived, flexible, and lineage-specific genes were annotated and
2. classified based on KEGG annotations [64] of clade IIC member SCUT-2 [49] and clade
3. IIA member UW1 [4, 26]. The number of genes annotated in each metabolic pathway
4. was counted, with the number of each type of gene being divided by the total number
5. of genes in the pathway. Metabolic pathways with high proportions of derived genes
6. were considered to have undergone major changes during evolution.

# 2.7 Horizontal gene transfer (HGT) identification

1. Parametric and phylogenetic methods are commonly used to infer HGT [65]. This study
2. used the phylogenetic method for HGT identification. Each derived gene was queried
3. in the non-redundant (NR) database (published on May 7, 2015) [66] using the
4. following BLASTP parameter [*-max\_target\_seqs 100-value 1E-6*] to preserve the first
5. 100 BLAST results. The representative species were obtained from the first 100 BLAST
6. results. Subsequently, the numbers and percentages of *Ca.* Accumulibacter, non-*Ca.*
7. Accumulibacter Rhodocyclaceae, and non-Rhodocyclaceae members in the first 100
8. BLAST results were then calculated. A gene was considered a laterally derived gene if
9. the numbers of *Ca.* Accumulibacter or non-*Ca.* Accumulibacter Rhodocyclaceae-
10. related hits were less than 10%. All core and differential-derived genes in each
11. metabolic pathway were analyzed to determine if they were obtained via HGT. The
12. derived genes that were classified as HGT-originated are referred to as laterally derived
13. genes. The origination of key genes (*ppk*2 and the homolog of *pho*U) was further
14. confirmed using the phylogenetic method based on best-match analysis.

# 2.8 Metatranscriptomic analysis

1. An anaerobic-aerobic full-cycle study was performed on an enrichment culture of *Ca.*
2. Accumulibacter Clade IIC SCUT-2 in the lab-scale EBPR reactor SCUT
3. (Supplementary Materials). The P cycling activities and the transformation of carbon
4. compounds were monitored. Activated sludge samples were collected just before the
5. start of a sequencing batch reactor (SBR) cycle (0 min), and at 5 min (anaerobic phase),
6. 30 min (anaerobic phase), 105 min (aerobic phase), and 120 min (aerobic phase) of the
7. SBR cycle. The samples were snap-frozen in liquid N2 and stored at −80 °C before the
8. extraction of ribonucleic acid (RNA) for metatranscriptomic analysis.
9. For metatranscriptomic analysis, total RNA was extracted using the RNA PowerSoil®
10. Total RNA Isolation Kit (Omega Bio-Tek, GA, USA). Fastp [67] and SortMeRNA [68]
11. removed adaptation sequences and ribosomal ribonucleic acids (rRNAs). Filtered reads
12. were mapped to the corresponding *Ca.* Accumulibacter draft genome (i.e., SCUT-2)
13. using BBMap version 38.96 [69] and were normalized to transcript per million (TPM).
14. Genes with TPM > 100 were considered to be highly transcribed. Details on the reactor
15. operation, full-cycle study, sample collection, metagenomic analysis, and
16. metatranscriptomic analysis are found in the Supplementary Materials. Raw reads and
17. draft genomes obtained were submitted to NCBI under BioProject No. PRJNA807832
18. and No. PRJNA771771.

# 3 Results

1. **3.1 Identification of orthologous gene clusters**
2. A total of 60722 pan Rhodocyclaceae orthologous gene clusters were identified,
3. including 25080 homologous genes in the *Ca.* Accumulibacter pangenome
4. (Supplementary Materials Spreadsheet 2, Sheets 1 and 3). Large proportions (63.8%
5. and 54.7%) of gene families in the pan Rhodocyclaceae and pan *Ca.* Accumulibacter
6. genomes were present in only single genomes (Fig. 2a,b). Approximately 1% (626) of
7. gene families were present in ≥ 37 of the 43 genomes, which were used to define the
8. core pan Rhodocyclaceae genome (Fig. 2c). In the pan *Ca.* Accumulibacter genome,
9. 6.9% of genes were shared in ≥ 18 genomes (Fig. 2d). Non-paralogous genes (average
10. gene copy per genome = 1) account for high proportions of pan Rhodocyclaceae and
11. pan *Ca.* Accumulibacter genomes (95.6% and 93.8%, respectively) (Fig. 2e,f). The
12. orthologous gene cluster identification results include the number of representative
13. genes in each genome and summary statistics of pan Rhodocyclaceae and pan *Ca.*
14. Accumulibacter gene clusters are provided in Supplementary Materials Spreadsheet 2
15. (Sheets 2 and 4).

# 3.2 Gene flux analysis

1. Among the 25080 gene clusters in the pan *Ca.* Accumulibacter genome, 2499 (9.96%)
2. were inferred to occur in the genome of the LCA, and 1668 (6.73%) occurred before
3. the LCA. Eight hundred eighteen (3.26%) were acquired at the node of LCA. Gene
4. occurrence possibility calculation suggested that with a genome-number cutoff of 18,
5. 99.94% of core genes could be identified (Fig. 3a). At this cutoff value, 1725 (6.88%)
6. core genes were identified in the pan *Ca.* Accumulibacter genome (Fig. 3b,c). By
7. further reducing the cutoff value to 17, the number of core genes increased from 1725
8. to 1829, and those with known functions increased from 298 to 318. As this study
9. mainly focused on the changes in the genetic content, i.e., new core-derived genes and
10. horizontally transferred genes, looser cutoff values did not seem to bring new gains.
11. Thus, a relatively stricter cut-off value (i.e., 18) was used to ensure the accuracy of the
12. results. The gene gain or loss of a pangenome needs to be characterized in specific
13. lineage member genomes. To facilitate a subsequent combination with the
14. transcriptome data, SCUT-2 and UW1 were used as representative genomes for gene
15. flux analysis. Each gene in Clade IIC SCUT-2 and Clade IIA UW1 genomes was
16. classified as ancestral, derived, lineage-specific, or flexible genes. There were no
17. significant differences in the numbers and proportions of ancestral and flexible genes
18. in these two genomes (ancestral genes accounted for 32.6% and 34.7%; flexible genes
19. accounted for 43.8% and 43.6% in SCUT-2 and UW-1, respectively). Six hundred thirty
20. eight and eight hundred and two derived genes were found in the SCUT-2 and UW1
21. genomes (17.6% and 14.0%, respectively). One hundred eighty nine lineage-specific
22. genes (genes occurred only in UW1) were observed in UW1, which was slightly less
23. than those (i.e., 275) in the SCUT-2 genome (Fig. 3d). Figure 4 and Supplementary
24. Materials Spreadsheet 3 provided additional details about the presence, gain and loss
25. of genes, and the discrete categories to which they were assigned.

# 3.3 Evolution of *Ca.* Accumulibacter metabolic pathways

1. The collections of genes identified as ancestral, derived, flexible, and lineage-specific
2. genes were annotated using KEGG [64] and were grouped into different metabolic
3. pathways. In SCUT-2, 2293 genes were annotated to various metabolic pathways. The
4. translation metabolic pathway had the highest proportion of ancestral genes (77,
5. accounting for 96%). The largest number of ancestral genes (224) and derived genes
6. (63) was observed in the carbohydrate metabolism pathway, accounting for 63% and
7. 18%, respectively. The highest proportion (15 out of 53, 28.0%) of derived genes was
8. observed in the cell growth and death metabolic pathway (Fig. 5a). Within each primary
9. pathway, ancestral and derived genes also showed distinct proportions in different
10. secondary pathways. For instance, within the carbohydrate metabolism, the galactose
11. metabolism pathways had the highest proportion (4 out of 5, 80%) of derived genes.
12. Whereas ancestral genes dominated the citric acid cycle (TCA cycle) (25 out of 30,
13. 83%) and the glyoxylate and dicarboxylate metabolism pathways (33 out of 45, 73%).
14. In signal transduction, the two-component system contained the highest proportion of
15. derived genes (27 out of 182, 15%). In membrane transport, among the 122 ABC-
16. transporter encoding genes, 18 were derived (15%) (Fig. 5). Similar number and
17. proportion of genes assigning to various metabolic pathways were observed in the *Ca*.
18. Accumulibacter clade IIA UW1 genome with only two metabolic pathways (transport
19. and catabolism, cell growth and death) showing significant differences in the
20. proportions of derived genes (28% and 40% in SCUT-2 and 14% and 23% in UW1,
21. respectively) (Fig. 5 and Supplementary Materials Fig. S1). These results indicated that
22. different strains of *Ca.* Accumulibacter underwent comparable developmental changes
23. during evolution but, at the same time, preserved a certain degree of gene diversity.
24. Detailed annotation of each gene in SCUT-2 and UW1 can be viewed in Supplementary
25. Spreadsheet 4.

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

1. Relatively strict parameters (i.e., 70% identity and 75% coverage) were used to identify
2. homologous gene clusters. The derived genes were manually classified into those
3. derived from accumulative mutations and HGT. Phylogenetic analysis was further
4. performed to confirm that *ppk*2 and the homolog of *pho*U are horizontally derived
5. (Supplementary Materials Fig. S4). Among 298 core-derived genes successfully
6. annotated in KEGG, 124 were shown to have been acquired via HGT. Among the 124
7. genes, 67 were involved in KEGG pathways. The carbohydrate metabolism pathway
8. harbors the highest numbers (25) of derived genes via HGT, including these in
9. glycolysis/gluconeogenesis (e.g., genes encoding the phosphoglucomutase, the
10. glucokinase, and the phosphoglycerate kinase), starch and sucrose metabolism (e.g., the
11. starch synthase, and the glycogen phosphorylase genes), and in butanoate metabolism
12. (genes encoding the poly[(R)-3-hydroxyalkanoate polymerase subunits). In signal
13. transduction, the two-component system contained ten laterally derived genes, such as
14. genes encoding the REDOX signal transduction system proteins RegA/B and the
15. phosphate regulon proteins PhoR-PhoB. Another remarkable set of genes derived via
16. HGT was oxidative phosphorylation in the energy metabolism pathway, including these
17. encoding the NADH-quinone oxidoreductase subunit, the polyphosphate kinase, and
18. the cytochrome C. The inorganic phosphate transporter gene (*pit*) was also acquired via
19. HGT. Similar results were observed for UW1. In the two-component system, genes
20. encoding the REDOX signal transduction system proteins RegA/B and the phosphate
21. regulon proteins PhoR-PhoB were laterally derived. More details about the BLAST
22. comparison results can be found in Supplementary Materials Spreadsheet 5.

# 3.5 Comparison of genetic compositions in PAOs and non-PAOs

1. In the context of our investigation, the presence or absence of specific genes in *Ca.*
2. Accumulibacter, compared to closely related PAOs and non-PAOs, holds significant
3. implications for elucidating the genetic basis of the P cycling phenotype. If a gene was
4. present in *Ca.* Accumulibacter, but absent in other closely related PAOs, may also not
5. be a key to developing the P cycling phenotype. Conversely, if a gene was present in
6. *Ca.* Accumulibacter, or their closely related PAOs but absent in non-PAOs, might be a
7. key gene to the emergence of the PAO phenotype. For a better understanding of the
8. genomic difference between closely related PAOs and non-PAOs, a pan PAO genome
9. (composed of 21 *Ca.* Accumulibacter and two *Dechloromonas* PAOs) [51] analysis was
10. performed. The pan PAO genome was compared to the *Ca.* Propionivibrio aalborgensis
11. (a closely related GAO) [50] genome to identify differential genes (defined as core
12. genes present in the pan PAO genome but absent in the *Ca.* Propionivibrio aalborgensis
13. genome). In the pan PAO genome, 124 differential genes were identified. Alkaline
14. phosphatase synthesis response regulator (PhoP) and polyphosphate kinase 2 (PPK2)
15. genes were both differential genes. Other genes in the operon or the genes regulated by
16. PhoP were not differential genes. Carbohydrate metabolism had the largest differential
17. genes (16), including those encoding the acetyl-CoA C-acetyltransferase and the enoyl-
18. CoA hydratase. The cofactor and vitamin metabolic pathway harbored the second
19. largest number of differential genes (11), followed by energy metabolism (9),
20. replication and repair (6), and signal transduction (5) metabolic pathways. The lowest
21. number (1) of differential genes was observed in the transcription and metabolism of
22. other amino acid pathways. Further analysis of another 21 available *Propionivibrio*
23. genomes confirmed that *ppk*2 and *pho*U are differential genes between *Ca.*
24. Accumulibacter and *Propionivibrio*. HGT analysis was aimed at gene acquisition in *Ca.*
25. Accumulibacter during evolution, based on the hypothesis that the emergence of the P
26. cycling ability by PAOs resulted from the acquisition of certain key genes. However,
27. the hypothesis ignored the possibility that non-PAOs may have lost certain key genes
28. in the process of evolution, leading to their inability to remove P. Differential genes
29. included gene loss in non-PAOs during evolution. The analysis in this part allows us to
30. understand the evolutionary process from a different perspective more comprehensively.
31. More details about the differential genes (metabolic pathway and functional annotation)
32. can be found in Supplementary Materials Spreadsheet 6.

# 3.6 Metatranscriptomic profiles

1. By analysis of the gene transcription levels of *Ca*. Accumulibacter in a typical EBPR
2. cycle, we excluded genes that displayed no remarkable transcription in the comparative
3. genome may be excluded. Thus, the range of genes could be further narrowed down,
4. facilitating the identification of key genes important to the PAO phenotype.
5. Metatranscriptomic analysis was performed on an enrichment culture of *Ca*.
6. Accumulibacter clade IIC strain SCUT-2 (with a relative abundance of 37.1%, as
7. suggested by the metagenomic analysis). In the SCUT-2 genome, out of 5037 annotated
8. genes, 906 were highly transcribed (TPM > 100). There were 298 core-derived genes,
9. 84 of which were highly transcribed (Supplementary Materials Spreadsheet 7). To
10. understand the dynamic patterns and functional relationships of 905 core genes with
11. known function, they were classified into five clusters using the Mfuzz [70] (Fig. 6b).
12. Most genes (e.g., the acetate permease gene *act*P, NOF05\_02545) in Cluster 1 were
13. related to the transporter for carbon uptake and energy utilization. Cluster 2 showed a
14. pattern of increased transcription throughout the anaerobic period, peaking after oxygen
15. exposure. Key members of this cluster included the phosphate transport system
16. substrate-binding protein (*pstS*, NOF05\_04305) and the laterally derived polyphosphate
17. kinase 2 gene (*ppk2*, NOF05\_17285). Cluster 3 genes showed high transcription at the
18. beginning of the anaerobic stage and reduced towards the end of the anaerobic cycle,
19. correlating with the depletion of acetate (Fig. 6a). Their high transcription in the aerobic
20. stage was mostly related to the routing of anaerobically stored carbon to the TCA cycle
21. and glycogenesis [7, 26]. Cluster 4 contained genes encoding the distant homolog of
22. PhoU (NOF05\_17860, NOF05\_12350) and antitoxin CptB (NOF05\_13125), which
23. showed low transcription during the anaerobic stage but were upregulated during the
24. aerobic phase. These genes possibly play a role in sustaining vital activities and
25. controlling homeostatic environments [71]. Finally, genes in Cluster 5 may be
26. associated with the maintenance of stable intracellular environments or cell growth,
27. including genes encoding the ion transporters, such as the magnesium transporter gene
28. (NOF05\_18175) and the low-affinity inorganic phosphate transporter (*pit*,
29. NOF05\_12345). These clustering patterns aligned with the metabolic characteristics of
30. *Ca.* Accumulibacter in EBPR (Fig. 6a).
31. The transcription of horizontally transferred genes in SCUT-2 was further analyzed. 44
32. genes, which were identified to be obtained via HGT, were highly transcribed (Fig. 6c).
33. These genes were involved in pathways, such as glycolysis/gluconeogenesis
34. (phosphoglycerate kinase, and phosphoglucomutase), ABC transporters (branched-
35. chain amino acid transport system substrate-binding protein), butanoate metabolism
36. (poly[(R)-3-hydroxyalkanoate] polymerase subunit), the two-component system (low
37. molecular weight protein-tyrosine phosphatase, polysaccharide biosynthesis/export
38. protein, tyrosine-protein kinase, and serine protease), transporters for inorganic salts
39. (sulfate permease, and magnesium transporter), and showed high transcription
40. throughout the EBPR cycle. Polyphosphate kinase 2 gene (*ppk*2) was also highly
41. transcribed and was significantly upregulated in the anaerobic phase. The transcription
42. of the phosphate transport regulator (a distant homolog of PhoU) was significantly
43. upregulated in the aerobic stage. PHA synthesis-related genes were also highly
44. transcribed. A full list of the SCUT-2 gene transcription data can be found in
45. Supplementary Materials Spreadsheet 7.
46. Comparisons were further made to the gene transcription characteristics of UW1 [26].
47. 35 horizontally derived gene families were highly transcribed in both SCUT-2 and UW1
48. (Supplementary Materials Fig. S2). Apart from the homolog *pho*U genes and *pit*, which
49. are related to phosphate regulation and transport, 42 laterally derived gene families
50. were under-transcribed in SCUT-2 but highly transcribed in UW1, including the acetate
51. kinase gene. These 42 gene families may not play a key role in the evolution of non-
52. PAO to PAO due to their different transcription behaviors in SCUT-2 and UW1.
53. Combined with transcriptomic analysis, the range of key genes can be effectively
54. reduced, and a metabolic model of *Ca*. Accumulibacter can be constructed (Fig. 6d).
55. Most genes in the central carbon metabolic pathway were highly transcribed non-HGT
56. genes, indicating that this pathway is indispensable for *Ca.* Accumulibacter, yet raises
57. doubts about its direct involvement in the evolution from a non-PAO metabolism to a
58. PAO. In the P cycling pathway, several laterally acquired genes were involved,
59. suggesting their potential pivotal role in the evolution of *Ca*. Accumulibacter. Some of
60. them were highly transcribed, further implying their importance in the evolution of *Ca.*
61. Accumulibacter (Fig. 6).

# 4 Discussion

1. Previous research suggested that the transition of PAO from non-PAO may have
2. occurred at the node of *Ca.* Accumulibacter LCA [42]. However, a recent investigation
3. has put forth compelling evidence indicating the presence of PAOs in the
4. *Dechloromonas* genus (i.e., *C*a. Dechloromonas phosphoritropha, *C*a. Dechloromonas
5. phosphorivorans) [51], raising a possibility that the emergence of the PAO phenotype
6. may have occurred before the *Ca.* Accumulibacter LCA. Here, we discuss the function
7. of key laterally derived genes in the context of pangenomics and known PAO
8. metabolism. A metatranscriptomic analysis of an enrichment culture of *Ca*.
9. Accumulibacter Clade IIC member SCUT-2 contrasting those of *Ca*. Accumulibacter
10. Clade IIA UW1 was performed to study the transcriptional dynamics of key genes in
11. *Ca.* Accumulibacter. This approach allowed the exclusion of genes that were not highly
12. transcribed in the large collection of laterally derived genes to narrow down the range
13. of key genes to obtain new insights on key genomic features of the polyphosphate
14. accumulating trait.

# 4.1 Carbon substrate uptake

1. The largest number of genes were annotated to the carbohydrate metabolism pathway
2. in both SCUT-2 and UW1 genomes (354 and 369, respectively). The SCUT-2 genome
3. contained 224 ancestral genes, 63 derived genes, and 49 laterally derived genes.
4. Transcriptomic analysis suggested that when acetate was used as a carbon source, genes
5. directly related to intracellular acetate processing and PHA synthesis were remarkably
6. upregulated in SCUT-2 (Supplementary Materials Spreadsheet 7). The high-affinity
7. acetyl-CoA synthetase (NOF05\_02565) and low-affinity phosphate acetyltransferase
8. (NOF05\_11790) are responsible for acetate activation [11, 72]. Other genes involved
9. in the acetyl-CoA pathway, including the pyruvate kinase gene (NOF05\_14290) and the
10. phosphoenolpyruvate carboxykinase gene (NOF05\_14615), maintained high levels of
11. transcription throughout the anaerobic-aerobic cycle. However, these genes are all
12. ancestral genes. Only one horizontally transferred gene (i.e., the acetate kinase gene,
13. NOF05\_16845) was barely transcribed. Therefore, genes related to acetate processing
14. may not be pivotal factors contributing to the emergence of the PAO phenotype. In
15. addition, in the TCA cycle [73], there were 30 genes. Among them, only the
16. dihydrolipoamide dehydrogenase gene (NOF05\_18520) was laterally derived, whereas
17. transcribed at a low level. This indicates that the gain/loss of genes in the TCA cycle
18. might not have contributed remarkably to the evolution of non-PAOs to PAOs. Four
19. laterally derived genes occurred in the PHA synthesis pathway (*pha*C NOF05\_18015,
20. NOF05\_21650, NOF05\_21620, and *pha*A NOF05\_18020), NOF05\_21650 and
21. NOF05\_21620 were highly transcribed throughout the EBPR cycle (Fig. 6). Whereas
22. *Ca.* Propionivibrio aalborgensis also encoded these genes [50]. Their contribution to
23. the evolution from a non-PAO metabolism to a PAO metabolism was unlikely.

# 4.2 Two-component systems

1. The two-component signal transduction system enables bacteria to sense, respond, and
2. adapt to diverse and dynamic environmental conditions [74]. This system is commonly
3. preserved in the bacterial domain. The number of genes in the two-component system
4. was considered to be closely related to the bacteria’s living environment [75]. Bacteria
5. living in extreme environments tend to encode many signaling proteins for improved
6. adaption [76]. In the SCUT-2 genome, a total of 182 genes were annotated to the two-
7. component system, including 81 ancestral genes and 27 derived genes. Notably, 12 of
8. these genes have been acquired laterally. In both SCUT-2 and UW1, phosphate regulon
9. response regulator gene *pho*B (NOF05\_18105), phosphate regulon sensor histidine
10. kinase gene *pho*R (NOF05\_18105), and redox signaling genes *reg*A and *reg*B
11. (NOF05\_11115, NOF05\_11120) were laterally derived. RegB/RegA was shown to
12. control and regulate a variety of basic metabolic processes in *Rhodobacter*, *Capsulatus*,
13. and *Sphaeroides*, such as photosynthesis, CO2 fixation, N2 assimilation, denitrification,
14. and electron transport [77] via direct or indirect control of respective operons [78, 79].
15. However, both *reg*A and *reg*B were absent in two *Dechloromonas* PAO genomes
16. (GCA\_016722705.1 and GCA\_016721185.1) [51], suggesting that the redox signaling
17. RegA/B were not indispensable for a PAO phenotype. PhoR-PhoB is present in both
18. *Ca.* Accumulibacter and two *Dechloromonas* PAO genomes can potentially play a role
19. in PAO phenotype evolution. Since the PhoR-PhoB system is a part of the Pho regulon,
20. further discussion was provided in the following subsection.

# 4.3 Phosphate regulatory system

1. The phosphate regulator (Pho) is a regulatory mechanism to maintain and manage
2. inorganic phosphate concentrations in bacterial cells. The system typically consists of
3. extracellular enzymes, transporters, and enzymes involved in the intracellular storage
4. of phosphate [80]. Signal transduction of Pho regulators requires seven proteins,
5. including PhoR, PhoB, four components of the ABC transporter Pst (PstS, PstA, PstB,
6. and PstC), and PhoU. An increase in the extracellular phosphate concentration near the
7. PstSCAB transporter would increase phosphate binding to PhoU, inhibiting the PhoR
8. kinase activity and the PstSCAB transporter activity. In the absence of phosphate input,
9. PhoU dissociates with phosphate, allowing the phosphate transport (Pst) to return to a
10. normal working state [81]. The above feedback control enables bacteria to maintain and
11. control a relatively stable intracellular phosphate concentration. Most of the genes in
12. the Pho regulatory system in *Ca.* Accumulibacter are laterally derived, including those
13. encoding PhoR and PhoB. In addition, within the genomes of *Ca*. Accumulibacter, three
14. distant homologs of the phoU gene (designated as NOF05\_17860, NOF05\_09930, and
15. NOF05\_09935) were identified. Distant homologs are protein pairs with similar
16. structures and functions but low gene sequence similarity [82]. The homolog *pho*U is
17. located in the *pit* operon within *Ca.* Accumulibacter genomes. Moreover, PhoR-PhoB
18. is also present in two *Dechloromonas* PAO genomes (*Ca.* Dechloromonas
19. phosphoritropha and *Ca.* Dechloromonas phosphorivorans).
20. In SCUT-2, the transcription of *pho*R (NOF05\_18110) and *pho*B (NOF05\_18105,
21. NOF05\_19100) was negligible. The transcription level of *pho*R (CAP2UW1\_1997) in
22. UW1 was also low. The transcription of *pho*B (CAP2UW1\_1996) in the aerobic phase
23. was slightly upregulated (with TPM values from 12 to 92) but was still at relatively low
24. levels (Supplementary Materials Spreadsheet 5). These results suggest that PhoR-PhoB
25. in *Ca.* Accumulibacter was probably not active in perceiving phosphate concentrations.
26. Similarly, the *pho*U genes were almost not transcribed (with the maximum TPM values
27. < 12, Fig. 6). Although the homolog *pho*U genes showed high transcription, the trend
28. was not in line with *pst*, indicating that the PhoU or their laterally derived homologs
29. were not effectively regulating Pst (Supplementary Materials Fig. S3). The same
30. phenomenon was observed in UW1 [26] and UW6 [45] metatranscriptome
31. (Supplementary Materials Spreadsheet 7 ). In *Staphylococcus aureus*, the absence of
32. *pho*U homolog, located in the *pit* operon, leads to the upregulation of phosphate
33. transporter genes (*pst*), increasing intracellular polyphosphate levels [71]. In
34. *Sinorhizobium meliloti*, the absence of *pho*U resulted in excessive accumulation of
35. phosphate, which inactivates cells due to P poisoning, resulting in poor cell growth [83,
36. 84]. Based on these results, we proposed two hypotheses. (1) PhoU in *Ca.*
37. Accumulibacter was ineffective in regulating Pst even under high intracellular
38. phosphate concentrations (no transcription of the *pho*U, and the unmatched
39. transcription of *pho*U homolog and *pst*, Supplementary Materials Fig. S3). Pst
40. continued to operate (as indicated by the high transcription of *pst* in the transcriptome,
41. Supplementary Materials Fig. S3), resulting in excessive phosphate accumulation in
42. cells (Fig. 6a). The laterally derived PPK2 functioned (as suggested by the high
43. transcription of *ppk*2, Supplementary Materials Fig. S3) to condense excess phosphate
44. into poly-P to avoid P poisoning. The second is that, in *Ca*. Accumulibacter, since *pho*U,
45. the homolog of *pho*U and *ppk*2 were derived from different donor bacteria
46. (Rhodocyclaceae, *Burkholderia*, and Gramaproteobacteria, respectively, as suggested
47. by the BLAST results, Supplementary Materials SpreadSheet 5), their encoding
48. proteins (i.e., PhoU, PhoU homolog, and PPK2) may have incompatible phosphate
49. activation/inactivation thresholds. PPK2 continued to synthesize poly-P by consuming
50. intracellular phosphate transported via Pst, resulting in consistently low intracellular
51. phosphate concentration, which was insufficient to combine with PhoU and/or its
52. homologs to downregulate Pst. In the SCUT-2 and UW1 transcriptomes (Fig. 6), PPK2
53. showed high levels of transcription during the entire EBPR cycle (with TPM values up
54. to 12481 in SCUT-2), which was further up-regulated in the aerobic stage, suggesting
55. that PPK2 worked to synthesize poly-P by consuming phosphate which was imported
56. via Pst, avoiding possible cell inactivation and poisoning due to elevated intracellular
57. phosphate concentrations and achieved poly-P accumulation. In addition, *Ca.*
58. Dechloromonas phosporitropha lacked *pst*, *pho*U, *pho*B, and *pho*R genes in the Pho
59. regulon, which is consistent with our hypothesis that the Pho regulation may not work
60. properly in PAOs. The transcriptomics data of *Microlunatus phosphovorus* (BioProject
61. No. PRJNA984968) and proteomics data of *Tetrasphaera elongate* (obtained from
62. Herbst et al. [85]) were further analyzed to check whether the same hypothesis could
63. apply to other PAOs. In the transcriptome of *Microlunatus phosphovorus*, we found that
64. the transcriptional patterns of *pst* were also inconsistent with those of *pho*U during an
65. anaerobic and aerobic cycle (Supplementary Materials). From the proteome of
66. *Tetrasphaera elongate*, the relative abundances of Pst and PhoU did not vary
67. significantly between anaerobic and oxic conditions; hence, they were not significantly
68. affected by changes in phosphate concentrations [85]. Taken together, these results
69. suggest that in *Microlunatus phosphovorus* and *Tetrasphaera elongate*, the regulation
70. of Pst by PhoU was not effective and that the Pho dysregulation hypothesis may also
71. apply to non-*Ca.* Accumulibacter PAOs. However, additional work is needed to confirm
72. its broad applicability.
73. Despite that, there is limited research on the Pho regulatory system in *Ca.*
74. Accumulibacter, the transcriptomics and gene origination analysis in the Pho regulon
75. suggested that it may represent a key link in the emergence of the PAO phenotype.

# 4.4 Transport of phosphate

1. Phosphorus (organic and/or inorganic) is a typical restricting nutrient. Therefore,
2. microorganisms developed adaptive mechanisms to cope with ordinary P deficiency.
3. Low-affinity inorganic phosphate transport systems (Pit) and high-affinity phosphate
4. transport systems (Pst) are key transporters used for inorganic phosphate transport [86,
5. 87]. In the pan *Ca.* Accumulibacter genomes, genes encoding the Pst transporter, are
6. neither core nor laterally derived. Furthermore, *Ca.* Dechloromonas phosporitropha
7. (PAO) do not encode any *pst* [51]. These results suggested that the Pst transport system
8. may not be indispensable for a PAO phenotype. *Ca.* Dechloromonas phosporitropha
9. encoded a phosphonates/phosphate transport system (Phn), which was shown to be a
10. high-affinity phosphate transporter in *Mycobacterium smegmatis* [88]. This system may
11. serve as a backup for the Pst transport system in *Ca.* Dechloromonas phosporitropha.
12. In the pan PAO genome, the low-affinity inorganic phosphate transporter gene (*pit*,
13. NOF05\_09925, NOF05\_09940) was laterally derived. The efflux of phosphate in
14. symport with H+ via Pit produces proton motive force, which is a key driving force for
15. the uptake of VFAs, lactate, succinate and amino acids by *Ca.* Accumulibacter [7, 8,
16. 89]. Therefore, *pit* is an important feature gene for the PAO phenotype. In SCUT-2
17. transcriptomes, the transcription of the *pit* was upregulated during the transition from
18. anaerobic to aerobic conditions (Supplementary Materials Fig. S3). The confirmed
19. GAO, *Ca.* Propionivibrio aalborgensis, which is closely related to *Ca.* Accumulibacter
20. (Fig. 1), are lack of *pit*. But *pit* is present in the genomes of other GAOs, for example,
21. *Defluviicoccus* GAO-HK [90], *Ca.* Competibacter denitrificans, and *Ca.*
22. Contendobacter odensis [91]. In addition, we analyzed 21 *Propionivibrio* genomes in
23. the NCBI database. Pit transporter was encoded in 13 of 21 *Propionivibrio* genomes
24. (Supplementary Materials Table S4). Anyhow, *pit* may not be a key feature driving the
25. evolution of non-PAO into PAOs and may neither be used as a marker gene for the PAO
26. phenotype, although it is indispensable for the P cycling trait.

# 5. Conclusion

1. In this study, we conducted pangomics with metatranscriptomic analysis on an
2. enrichment culture of *Ca.* Accumulibacter clade IIC member SCUT-2. The primary
3. objectives of this investigation were to understand the genomic transition in the
4. evolution of *Ca.* Accumulibacter and to identify the key genes responsible for the
5. emergence of the P-accumulating traits. Our study has brought forth several noteworthy
6. findings:
7. (1) A total of 298 core genes were identified as novel acquisitions in the ancestral
8. lineage of *Ca.* Accumulibacter, with 124 of them being derived via HGT. Notably, 44
9. of these laterally derived core genes were highly transcribed in a typical EBPR cycle.
10. (2) A high-affinity phosphate transport system (Pst) may not be indispensable for the
11. PAO phenotype. Inorganic phosphate transporter (Pit) may not be a key feature driving
12. non-PAO evolution into PAOs. Consequently, their encoding genes may not be reliable
13. markers for the PAO phenotype.
14. (3) Low transcription of the *pho*R-*pho*B two-component system genes and the
15. unmatched transcription of *pst* and *pho*U implied that the Pho regulon may not function
16. properly in *Ca.* Accumulibacter.
17. (4) A Pho dysregulation hypothesis was proposed. The PhoU and laterally derived PhoU
18. homologs in *Ca.* Accumulibacter were ineffective in regulating Pst, resulting in
19. excessive P uptake. To avoid P poisoning, the laterally derived PPK2 was employed to
20. condense excess phosphate into poly-P. Alternatively, PhoU and PPK2 genes were
21. derived from different donor bacteria, resulting in unmatched activation/inactivation
22. thresholds. PPK2 tends to reduce the intracellular phosphate concentration to levels
23. perceived by PhoU as low-phosphate states, thereby promoting continuous phosphate
24. uptake.
25. This study is expected to provide a new perspective for understanding the development
26. and evolution of the P cycling traits for *Ca.* Accumulibacter.

# CRediT authorship contribution statement

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

# Declaration of Competing Interest

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

# Acknowledgments

1. This research was supported by the National Natural Science Foundation of China
2. (52270035 and 51808297), the Natural Science Foundation of Guangdong Province
3. (2021A1515010494), the Guangzhou Key Research and Development Program
4. (2023B03J1334), and the Pearl River Talent Recruitment Program (2019QN01L125).

# Data available

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

# Reference

1. [1]. Bunce JT, Ndam E, Ofiteru ID, Moore A, Graham DW: **A review of**

# phosphorus removal technologies and their applicability to small-scale

1. **domestic wastewater treatment systems**. *Frontiers in Environmental Science*
2. 2018, **6:**8.
3. [2]. Abdelfattah A, Ali SS, Ramadan H, El-Aswar EI, Eltawab R, Ho S-H, Elsamahy
4. T, Li S, El-Sheekh MM, Schagerl M, Kornaros M, Sun J: **Microalgae-based**

# wastewater treatment: Mechanisms, challenges, recent advances, and

1. **future prospects**. *Environmental Science and Ecotechnology* 2023, **13**:100205.

|  |  |  |
| --- | --- | --- |
| 636 | [3]. | Qiu G, Law Y, Zuniga-Montanez R, Deng X, Lu Y, Roy S, Thi SS, Hoon HY, |
| 637 |  | Nguyen TQN, Eganathan K, Liu X, Nielsen PH, Williams RBH, Wuertz S: |
| 638 |  | **Global warming readiness: Feasibility of enhanced biological phosphorus** |
| 639 |  | **removal at 35 °C**. *Water Research* 2022, **216**:118301. |
| 640 | [4]. | Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates |
| 641 |  | C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan |
| 642 |  | JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P: |
| 643 |  | **Metagenomic analysis of two enhanced biological phosphorus removal** |
| 644 |  | **(EBPR) sludge communities**. *Nature Biotechnology* 2006, **24**(10):1263-1269. |
| 645 | [5]. | Oehmen A, Lemos PC, Carvalho G, Yuan Z, Keller J, Blackall LL, Reis MAM: |
| 646 |  | **Advances in enhanced biological phosphorus removal: From micro to** |
| 647 |  | **macro scale**. *Water Research* 2007, **41**(11):2271-2300. |
| 648 | [6]. | Qiu G, Zuniga-Montanez R, Law Y, Thi SS, Nguyen TQN, Eganathan K, Liu X, |
| 649 |  | Nielsen PH, Williams RBH, Wuertz S: **Polyphosphate-accumulating** |
| 650 |  | **organisms in full-scale tropical wastewater treatment plants use diverse** |
| 651 |  | **carbon sources**. *Water Research* 2019, **149**:496-510. |
| 652 | [7]. | Qiu G, Liu X, Saw NMMT, Law Y, Zuniga-Montanez R, Thi SS, Ngoc Nguyen |
| 653 |  | TQ, Nielsen PH, Williams RBH, Wuertz S: **Metabolic traits of *Candidatus*** |
| 654 |  | **Accumulibacter clade IIF strain SCELSE-1 using amino acids as carbon** |
| 655 |  | **sources for enhanced biological phosphorus removal**. *Environmental Science* |
| 656 |  | *& Technology* 2020, **54**(4):2448-2458. |
| 657 | [8]. | Chen L, Chen H, Hu Z, Tian Y, Wang C, Xie P, Deng X, Zhang Y, Tang X, Lin |
| 658 |  | X, Li B, Wei C, Qiu G: **Carbon uptake bioenergetics of PAOs and GAOs in** |
| 659 |  | **full-scale enhanced biological phosphorus removal systems**. *Water Research* |
| 660 |  | 2022, **216**:118258. |
| 661 | [9]. | Diaz R, Mackey B, Chadalavada S, kainthola J, Heck P, Goel R: **Enhanced Bio-** |
| 662 |  | **P removal: Past, present, and future – A comprehensive review**. |
| 663 |  | *Chemosphere* 2022, **309**:136518. |
| 664 | [10]. | Zhang C, Guisasola A, Baeza JA: **A review on the integration of mainstream** |
| 665 |  | **P-recovery strategies with enhanced biological phosphorus removal**. *Water* |
| 666 |  | *Research* 2022, **212**:118102. |
| 667 | [11]. | He S, McMahon KD: **Microbiology of '*Candidatus* Accumulibacter' in** |
| 668 |  | **activated sludge**. *Microbial Biotechnology* 2011, **4**(5):603-619. |
| 669 | [12]. | Nielsen PH, McIlroy SJ, Albertsen M, Nierychlo M: **Re-evaluating the** |
| 670 |  | **microbiology of the enhanced biological phosphorus removal process**. |
| 671 |  | *Current Opinion in Biotechnology* 2019, **57**:111-118. |
| 672 | [13]. | Dorofeev AG, Nikolaev YA, Mardanov AV, Pimenov NV: **Role of phosphate-** |
| 673 |  | **accumulating bacteria in biological phosphorus removal from wastewater**. |
| 674 |  | *Applied Biochemistry and Microbiology* 2020, **56**(1):1-14. |
| 675 | [14]. | Zhang C, Chen X, Han M, Li X, Chang H, Ren N, Ho S-H: **Revealing the role** |
| 676 |  | **of microalgae-bacteria niche for boosting wastewater treatment and energy** |
| 677 |  | **reclamation in response to temperature**. *Environmental Science and* |
| 678 |  | *Ecotechnology* 2023, **14**:100230. |
| 679 | [15]. | Seviour RJ, Mino T, Onuki M: **The microbiology of biological phosphorus** |

|  |  |  |
| --- | --- | --- |
| 680 |  | **removal in activated sludge systems**. *FEMS Microbiology Reviews* 2003, |
| 681 |  | **27**(1):99-127. |
| 682 | [16]. | Mao Y, Graham DW, Tamaki H, Zhang T: **Dominant and novel clades of** |
| 683 |  | ***Candidatus* Accumulibacter phosphatis in 18 globally distributed full-scale** |
| 684 |  | **wastewater treatment plants**. *Scientific Reports* 2015, **5**(1):11857. |
| 685 | [17]. | Roy S, Guanglei Q, Zuniga-Montanez R, Williams RBH, Wuertz S: **Recent** |
| 686 |  | **advances in understanding the ecophysiology of enhanced biological** |
| 687 |  | **phosphorus removal**. *Current Opinion in Biotechnology* 2021, **67**:166-174. |
| 688 | [18]. | Petriglieri F, Singleton CM, Kondrotaite Z, Dueholm MKD, McDaniel EA, |
| 689 |  | McMahon KD, Nielsen PH: **Reevaluation of the phylogenetic diversity and** |
| 690 |  | **global distribution of the genus *Candidatus* Accumulibacter**. *mSystems* 2022, |
| 691 |  | **7**(3):e00016-00022. |
| 692 | [19]. | Oehmen A, Zeng RJ, Yuan Z, Keller J: **Anaerobic metabolism of propionate** |
| 693 |  | **by polyphosphate-accumulating organisms in enhanced biological** |
| 694 |  | **phosphorus removal systems**. *Biotechnology and Bioengineering* 2005, |
| 695 |  | **91**(1):43-53. |
| 696 | [20]. | Kolakovic S, Freitas EB, Reis MAM, Carvalho G, Oehmen A: **Accumulibacter** |
| 697 |  | **diversity at the sub-clade level impacts enhanced biological phosphorus** |
| 698 |  | **removal performance**. *Water Research* 2021, **199**:117210. |
| 699 | [21]. | Zhao W, Bi X, Peng Y, Bai M: **Research advances of the phosphorus-** |
| 700 |  | **accumulating organisms of *Candidatus* Accumulibacter, *Dechloromonas*** |
| 701 |  | **and *Tetrasphaera*: Metabolic mechanisms, applications and influencing** |
| 702 |  | **factors**. *Chemosphere* 2022, **307**:135675. |
| 703 | [22]. | Páez-Watson T, van Loosdrecht MCM, Wahl SA: **Predicting the impact of** |
| 704 |  | **temperature on metabolic fluxes using resource allocation modelling:** |
| 705 |  | **Application to polyphosphate accumulating organisms**. *Water Research* |
| 706 |  | 2023, **228**:119365. |
| 707 | [23]. | Bessarab I, Maszenan AM, Haryono MAS, Arumugam K, Saw NMMT, Seviour |
| 708 |  | RJ, Williams RBH: **Comparative genomics of members of the genus** |
| 709 |  | ***Defluviicoccus* with insights into their ecophysiological importance**. |
| 710 |  | *Frontiers in Microbiology* 2022, **13:**834906. |
| 711 | [24]. | Maszenan AM, Bessarab I, Williams RBH, Petrovski S, Seviour RJ: **The** |
| 712 |  | **phylogeny, ecology and ecophysiology of the glycogen accumulating** |
| 713 |  | **organism (GAO) *Defluviicoccus* in wastewater treatment plants**. *Water* |
| 714 |  | *Research* 2022, **221**:118729. |
| 715 | [25]. | Turcotte MM, Corrin MSC, Johnson MTJ: **Adaptive evolution in ecological** |
| 716 |  | **communities**. *PLOS Biology* 2012, **10**(5):e1001332. |
| 717 | [26]. | Oyserman BO, Noguera DR, del Rio TG, Tringe SG, McMahon KD: |
| 718 |  | **Metatranscriptomic insights on gene expression and regulatory controls in** |
| 719 |  | ***Candidatus* Accumulibacter phosphatis**. *The ISME Journal* 2016, **10**(4):810- |
| 720 |  | 822. |
| 721 | [27]. | Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, |
| 722 |  | Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora |
| 723 |  | M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson |

|  |  |  |
| --- | --- | --- |
| 724 |  | WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, |
| 725 |  | Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, |
| 726 |  | Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, |
| 727 |  | White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels |
| 728 |  | MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic** |
| 729 |  | **isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-** |
| 730 |  | **genome"**. *Proceedings of the National Academy of Sciences* 2005, |
| 731 |  | **102**(39):13950-13955. |
| 732 | [28]. | Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou |
| 733 |  | R, Xie W-Z, Cheng Y, Zhang Y, Liu K, Yang Q-Y, Chen L-L, Guo L: **Eight** |
| 734 |  | **high-quality genomes reveal pan-genome architecture and ecotype** |
| 735 |  | **differentiation of *Brassica napus***. *Nature Plants* 2020, **6**(1):34-45. |
| 736 | [29]. | Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-** |
| 737 |  | **genome**. *Current Opinion in Genetics & Development* 2005, **15**(6):589-594. |
| 738 | [30]. | Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN: **How the pan-genome** |
| 739 |  | **is changing crop genomics and improvement**. *Genome Biology* 2021, **22**(1):3. |
| 740 | [31]. | Aggarwal SK, Singh A, Choudhary M, Kumar A, Rakshit S, Kumar P, Bohra A, |
| 741 |  | Varshney RK: **Pangenomics in microbial and crop research: progress,** |
| 742 |  | **applications, and perspectives**. *Genes* 2022, **13**(4):598. |
| 743 | [32]. | Golicz AA, Batley J, Edwards D: **Towards plant pangenomics**. *Plant* |
| 744 |  | *Biotechnology Journal* 2016, **14**(4):1099-1105. |
| 745 | [33]. | Flowers JJ, He S, Malfatti S, del Rio TG, Tringe SG, Hugenholtz P, McMahon |
| 746 |  | KD: **Comparative genomics of two '*Candidatus* Accumulibacter' clades** |
| 747 |  | **performing biological phosphorus removal**. *The ISME Journal* 2013, |
| 748 |  | **7**(12):2301-2314. |
| 749 | [34]. | Camejo PY, Oyserman BO, McMahon KD, Noguera DR: **Integrated omic** |
| 750 |  | **analyses provide evidence that a "*Candidatus* Accumulibacter phosphatis"** |
| 751 |  | **strain performs denitrification under microaerobic conditions**. *mSystems* |
| 752 |  | 2019, **4**(1):e00193-00118. |
| 753 | [35]. | El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler |
| 754 |  | E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, |
| 755 |  | Bartholomeu DC, Haas BJ, Tran A-N, Wortman JR, Alsmark UCM, Angiuoli S, |
| 756 |  | Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy |
| 757 |  | T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, |
| 758 |  | Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, |
| 759 |  | Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, |
| 760 |  | Stuart KD, Hall N: **Comparative genomics of trypanosomatid parasitic** |
| 761 |  | **protozoa**. *Science* 2005, **309**(5733):404-409. |
| 762 | [36]. | Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González |
| 763 |  | JM, Pedrós-Alió C: **Ecology of marine *Bacteroidetes*: a comparative** |
| 764 |  | **genomics approach**. *The ISME Journal* 2013, **7**(5):1026-1037. |
| 765 | [37]. | Coghlan A, Tyagi R, Cotton JA, Holroyd N, Rosa BA, Tsai IJ, Laetsch DR, |
| 766 |  | Beech RN, Day TA, Hallsworth-Pepin K, Ke H-M, Kuo T-H, Lee TJ, Martin J, |
| 767 |  | Maizels RM, Mutowo P, Ozersky P, Parkinson J, Reid AJ, Rawlings ND, |

|  |  |  |
| --- | --- | --- |
| 768 |  | Ribeiro DM, Swapna LS, Stanley E, Taylor DW, Wheeler NJ, Zamanian M, |
| 769 |  | Zhang X, Allan F, Allen JE, Asano K, Babayan SA, Bah G, Beasley H, Bennett |
| 770 |  | HM, Bisset SA, Castillo E, Cook J, Cooper PJ, Cruz-Bustos T, Cuéllar C, |
| 771 |  | Devaney E, Doyle SR, Eberhard ML, Emery A, Eom KS, Gilleard JS, Gordon |
| 772 |  | D, Harcus Y, Harsha B, Hawdon JM, Hill DE, Hodgkinson J, Horák P, Howe |
| 773 |  | KL, Huckvale T, Kalbe M, Kaur G, Kikuchi T, Koutsovoulos G, Kumar S, |
| 774 |  | Leach AR, Lomax J, Makepeace B, Matthews JB, Muro A, O’Boyle NM, Olson |
| 775 |  | PD, Osuna A, Partono F, Pfarr K, Rinaldi G, Foronda P, Rollinson D, Samblas |
| 776 |  | MG, Sato H, Schnyder M, Scholz T, Shafie M, Tanya VN, Toledo R, Tracey A, |
| 777 |  | Urban JF, Wang L-C, Zarlenga D, Blaxter ML, Mitreva M, Berriman M, |
| 778 |  | International Helminth Genomes C: **Comparative genomics of the major** |
| 779 |  | **parasitic worms**. *Nature genetics* 2019, **51**(1):163-174. |
| 780 | [38]. | Kjærbølling I, Vesth T, Frisvad JC, Nybo JL, Theobald S, Kildgaard S, Petersen |
| 781 |  | TI, Kuo A, Sato A, Lyhne EK, Kogle ME, Wiebenga A, Kun RS, Lubbers RJM, |
| 782 |  | Mäkelä MR, Barry K, Chovatia M, Clum A, Daum C, Haridas S, He G, LaButti |
| 783 |  | K, Lipzen A, Mondo S, Pangilinan J, Riley R, Salamov A, Simmons BA, |
| 784 |  | Magnuson JK, Henrissat B, Mortensen UH, Larsen TO, de Vries RP, Grigoriev |
| 785 |  | IV, Machida M, Baker SE, Andersen MR: **A comparative genomics study of** |
| 786 |  | **23 *Aspergillus* species from section Flavi**. *Nature Communications* 2020, |
| 787 |  | **11**(1):1106. |
| 788 | [39]. | Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo |
| 789 |  | C, Faircloth BC, Petersen B, Wang Z, Zhou Q, Diekhans M, Chen W, Andreu- |
| 790 |  | Sánchez S, Margaryan A, Howard JT, Parent C, Pacheco G, Sinding M-HS, |
| 791 |  | Puetz L, Cavill E, Ribeiro ÂM, Eckhart L, Fjeldså J, Hosner PA, Brumfield RT, |
| 792 |  | Christidis L, Bertelsen MF, Sicheritz-Ponten T, Tietze DT, Robertson BC, Song |
| 793 |  | G, Borgia G, Claramunt S, Lovette IJ, Cowen SJ, Njoroge P, Dumbacher JP, |
| 794 |  | Ryder OA, Fuchs J, Bunce M, Burt DW, Cracraft J, Meng G, Hackett SJ, Ryan |
| 795 |  | PG, Jønsson KA, Jamieson IG, da Fonseca RR, Braun EL, Houde P, Mirarab S, |
| 796 |  | Suh A, Hansson B, Ponnikas S, Sigeman H, Stervander M, Frandsen PB, van |
| 797 |  | der Zwan H, van der Sluis R, Visser C, Balakrishnan CN, Clark AG, Fitzpatrick |
| 798 |  | JW, Bowman R, Chen N, Cloutier A, Sackton TB, Edwards SV, Foote DJ, |
| 799 |  | Shakya SB, Sheldon FH, Vignal A, Soares AER, Shapiro B, González-Solís J, |
| 800 |  | Ferrer-Obiol J, Rozas J, Riutort M, Tigano A, Friesen V, Dalén L, Urrutia AO, |
| 801 |  | Székely T, Liu Y, Campana MG, Corvelo A, Fleischer RC, Rutherford KM, |
| 802 |  | Gemmell NJ, Dussex N, Mouritsen H, Thiele N, Delmore K, Liedvogel M, |
| 803 |  | Franke A, Hoeppner MP, Krone O, Fudickar AM, Milá B, Ketterson ED, Fidler |
| 804 |  | AE, Friis G, Parody-Merino ÁM, Battley PF, Cox MP, Lima NCB, Prosdocimi |
| 805 |  | F, Parchman TL, Schlinger BA, Loiselle BA, Blake JG, Lim HC, Day LB, |
| 806 |  | Fuxjager MJ, Baldwin MW, Braun MJ, Wirthlin M, Dikow RB, Ryder TB, |
| 807 |  | Camenisch G, Keller LF, DaCosta JM, Hauber ME, Louder MIM, Witt CC, |
| 808 |  | McGuire JA, Mudge J, Megna LC, Carling MD, Wang B, Taylor SA, Del-Rio |
| 809 |  | G, Aleixo A, Vasconcelos ATR, Mello CV, Weir JT, Haussler D, Li Q, Yang H, |
| 810 |  | Wang J, Lei F, Rahbek C, Gilbert MTP, Graves GR, Jarvis ED, Paten B, Zhang |
| 811 |  | G: **Dense sampling of bird diversity increases power of comparative** |

|  |  |  |
| --- | --- | --- |
| 812 |  | **genomics**. *Nature* 2020, **587**(7833):252-257. |
| 813 | [40]. | Zhang Z, Guo Y, Yang F, Li J: **Pan-Genome analysis reveals functional** |
| 814 |  | **divergences in gut-restricted *Gilliamella* and *Snodgrassella***. *Bioengineering* |
| 815 |  | 2022, **9**(10):544. |
| 816 | [41]. | Fernandez-Fueyo E, Ruiz-Dueñas FJ, Ferreira P, Floudas D, Hibbett DS, |
| 817 |  | Canessa P, Larrondo LF, James TY, Seelenfreund D, Lobos S, Polanco R, Tello |
| 818 |  | M, Honda Y, Watanabe T, Watanabe T, Ryu JS, Kubicek CP, Schmoll M, Gaskell |
| 819 |  | J, Hammel KE, St. John FJ, Vanden Wymelenberg A, Sabat G, Splinter |
| 820 |  | BonDurant S, Syed K, Yadav JS, Doddapaneni H, Subramanian V, Lavín JL, |
| 821 |  | Oguiza JA, Perez G, Pisabarro AG, Ramirez L, Santoyo F, Master E, Coutinho |
| 822 |  | PM, Henrissat B, Lombard V, Magnuson JK, Kües U, Hori C, Igarashi K, |
| 823 |  | Samejima M, Held BW, Barry KW, LaButti KM, Lapidus A, Lindquist EA, |
| 824 |  | Lucas SM, Riley R, Salamov AA, Hoffmeister D, Schwenk D, Hadar Y, Yarden |
| 825 |  | O, de Vries RP, Wiebenga A, Stenlid J, Eastwood D, Grigoriev IV, Berka RM, |
| 826 |  | Blanchette RA, Kersten P, Martinez AT, Vicuna R, Cullen D: **Comparative** |
| 827 |  | **genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium*** |
| 828 |  | **provide insight into selective ligninolysis**. *Proceedings of the National* |
| 829 |  | *Academy of Sciences* 2012, **109**(14):5458-5463. |
| 830 | [42]. | Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, Noguera |
| 831 |  | DR, McMahon KD: **Ancestral genome reconstruction identifies the** |
| 832 |  | **evolutionary basis for trait acquisition in polyphosphate accumulating** |
| 833 |  | **bacteria**. *The ISME Journal* 2016, **10**(12):2931-2945. |
| 834 | [43]. | Arumugam K, Bağcı C, Bessarab I, Beier S, Buchfink B, Górska A, Qiu G, |
| 835 |  | Huson DH, Williams RBH: **Annotated bacterial chromosomes from frame-** |
| 836 |  | **shift-corrected long-read metagenomic data**. *Microbiome* 2019, **7**(1):61. |
| 837 | [44]. | Rubio-Rincón FJ, Weissbrodt DG, Lopez-Vazquez CM, Welles L, Abbas B, |
| 838 |  | Albertsen M, Nielsen PH, van Loosdrecht MCM, Brdjanovic D: **"*Candidatus*** |
| 839 |  | **Accumulibacter delftensis": A clade IC novel polyphosphate-accumulating** |
| 840 |  | **organism without denitrifying activity on nitrate**. *Water Research* 2019, |
| 841 |  | **161**:136-151. |
| 842 | [45]. | McDaniel EA, Moya-Flores F, Keene Beach N, Camejo PY, Oyserman BO, |
| 843 |  | Kizaric M, Khor EH, Noguera DR, McMahon KD: **Metabolic differentiation** |
| 844 |  | **of co-occurring Accumulibacter clades revealed through genome-resolved** |
| 845 |  | **metatranscriptomics**. *mSystems* 2021, **6**(4):e0047421. |
| 846 | [46]. | Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, |
| 847 |  | Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen |
| 848 |  | M: **Connecting structure to function with the recovery of over 1000 high-** |
| 849 |  | **quality metagenome-assembled genomes from activated sludge using long-** |
| 850 |  | **read sequencing**. *Nature Communications* 2021, **12**(1):2009. |
| 851 | [47]. | Srinivasan VN, Li G, Wang D, Tooker NB, Dai Z, Onnis-Hayden A, Bott C, |
| 852 |  | Dombrowski P, Schauer P, Pinto A, Gu AZ: **Oligotyping and metagenomics** |
| 853 |  | **reveal distinct *Candidatus* Accumulibacter communities in side-stream** |
| 854 |  | **versus conventional full-scale enhanced biological phosphorus removal** |
| 855 |  | **(EBPR) systems**. *Water Research* 2021, **206**:117725. |

|  |  |  |
| --- | --- | --- |
| 856 | [48]. | Tian Y, Chen H, Chen L, Deng X, Hu Z, Wang C, Wei C, Qiu G, Wuertz S: |
| 857 |  | **Glycine adversely affects enhanced biological phosphorus removal**. *Water* |
| 858 |  | *Research* 2022, **209**:117894. |
| 859 | [49]. | Deng X, Yuan J, Chen L, Chen H, Wei C, Nielsen PH, Wuertz S, Qiu G: |
| 860 |  | **CRISPR-Cas phage defense systems and prophages in *Candidatus*** |
| 861 |  | **Accumulibacter**. *Water Research* 2023, **235**:119906. |
| 862 | [50]. | Albertsen M, McIlroy SJ, Stokholm-Bjerregaard M, Karst SM, Nielsen PH: |
| 863 |  | **"*Candidatus* Propionivibrio aalborgensis": a novel glycogen accumulating** |
| 864 |  | **organism abundant in full-scale enhanced biological phosphorus removal** |
| 865 |  | **plants**. *Frontiers in Microbiology* 2016, **7**:1033. |
| 866 | [51]. | Petriglieri F, Singleton C, Peces M, Petersen JF, Nierychlo M, Nielsen PH: |
| 867 |  | **"*Candidatus* Dechloromonas phosphoritropha" and "*Ca*. D.** |
| 868 |  | **phosphorivorans", novel polyphosphate accumulating organisms** |
| 869 |  | **abundant in wastewater treatment systems**. *The ISME Journal* 2021, |
| 870 |  | **15**(12):3605-3614. |
| 871 | [52]. | Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM:** |
| 872 |  | **assessing the quality of microbial genomes recovered from isolates, single** |
| 873 |  | **cells, and metagenomes**. *Genome Research* 2015, **25**(7):1043-1055. |
| 874 | [53]. | Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for** |
| 875 |  | **comparative genomics**. *Genome Biology* 2019, **20**(1):238. |
| 876 | [54]. | Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-** |
| 877 |  | **scale detection of protein families**. *Nucleic Acids Research* 2002, **30**(7):1575- |
| 878 |  | 1584. |
| 879 | [55]. | Katoh K, Standley DM: **MAFFT multiple sequence alignment software** |
| 880 |  | **version 7: Improvements in performance and usability**. *Molecular Biology* |
| 881 |  | *and Evolution* 2013, **30**(4):772-780. |
| 882 | [56]. | Castresana J: **Selection of conserved blocks from multiple alignments for** |
| 883 |  | **their sse in phylogenetic analysis**. *Molecular Biology and Evolution* 2000, |
| 884 |  | **17**(4):540-552. |
| 885 | [57]. | Shen W, Le S, Li Y, Hu F: **SeqKit: A cross-platform and ultrafast toolkit for** |
| 886 |  | **FASTA/Q file manipulation**. *PLOS ONE* 2016, **11**(10):e0163962. |
| 887 | [58]. | Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von |
| 888 |  | Haeseler A, Lanfear R: **IQ-TREE 2: New models and efficient methods for** |
| 889 |  | **phylogenetic inference in the genomic era**. *Molecular Biology and Evolution* |
| 890 |  | 2020, **37**(5):1530-1534. |
| 891 | [59]. | Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v5: an online tool for** |
| 892 |  | **phylogenetic tree display and annotation**. *Nucleic Acids Research* 2021, |
| 893 |  | **49**(W1):W293-W296. |
| 894 | [60]. | Zhang A-N, Mao Y, Wang Y, Zhang T: **Mining traits for the enrichment and** |
| 895 |  | **isolation of not-yet-cultured populations**. *Microbiome* 2019, **7**(1):96. |
| 896 | [61]. | Csűös M: **Count: evolutionary analysis of phylogenetic profiles with** |
| 897 |  | **parsimony and likelihood**. *Bioinformatics* 2010, **26**(15):1910-1912. |
| 898 | [62]. | Pál C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic** |
| 899 |  | **networks by horizontal gene transfer**. *Nature genetics* 2005, **37**(12):1372- |

900 1375.

1. [63]. Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T,
2. McMahon K, Bertilsson S, Stepanauskas R, Andersson SG: **Single-cell**

# genomics reveal low recombination frequencies in freshwater bacteria of

1. **the SAR11 clade**. *Genome Biology* 2013, **14**(11):1-14.
2. [64]. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data,**

# information, knowledge and principle: back to metabolism in KEGG.

1. *Nucleic Acids Research* 2013, **42**(D1):D199-D205.
2. [65]. Ravenhall M, Škunca N, Lassalle F, Dessimoz C: **Inferring horizontal gene**
3. **transfer**. *PLOS Computational Biology* 2015, **11**(5):e1004095.
4. [66]. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a**

# curated non-redundant sequence database of genomes, transcripts and

1. **proteins**. *Nucleic Acids Research* 2006, **35**(suppl\_1):D61-D65.
2. [67]. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ**
3. **preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.
4. [68]. Kopylova E, Noé L, Touzet H: **SortMeRNA: fast and accurate filtering of**
5. **ribosomal RNAs in metatranscriptomic data**. *Bioinformatics* 2012,
6. **28**(24):3211-3217.
7. [69]. Bushnell B: **BBMap: A fast, accurate, splice-aware aligner**. In: *Conference:*
8. *9th Annual Genomics of Energy &amp; Environment Meeting, Walnut Creek,*
9. *CA, March 17-20, 2014; United States*. DE-AC02-05CH11231 2016-04-08:
10. 2014: Medium: ED.
11. [70]. Kumar L, M EF: **Mfuzz: a software package for soft clustering of**
12. **microarray data**. *Bioinformation* 2007, **2**(1):5-7.
13. [71]. Shang Y, Wang X, Chen Z, Lyu Z, Lin Z, Zheng J, Wu Y, Deng Q, Yu Z, Zhang
14. Y, Qu D: ***Staphylococcus* aureus PhoU homologs regulate persister**
15. **formation and virulence**. *Frontier of Microbiology* 2020, **11**:865.
16. [72]. Chen L, Wei G, Zhang Y, Wang K, Wang C, Deng X, Li Y, Xie X, Chen J, Huang
17. F, Chen H, Zhang B, Wei C, Qiu G: ***Candidatus* Accumulibacter use**

# fermentation products for enhanced biological phosphorus removal. *Water*

1. *Research* 2023, **246**:120713.
2. [73]. Zhou Y, Pijuan M, Zeng RJ, Yuan Z: **Involvement of the TCA cycle in the**

# anaerobic metabolism of polyphosphate accumulating organisms (PAOs).

1. *Water Research* 2009, **43**(5):1330-1340.
2. [74]. Capra EJ, Laub MT: **Evolution of two-component signal transduction**
3. **systems**. *Annual Review of Microbiology* 2012, **66**:325-347.
4. [75]. Alm E, Huang K, Arkin A: **The evolution of two-component systems in**

# bacteria reveals different strategies for niche adaptation. *PLOS*

1. *Computational Biology* 2006, **2**(11):e143.
2. [76]. Ulrich LE, Zhulin IB: **The MiST2 database: a comprehensive genomics**
3. **resource on microbial signal transduction**. *Nucleic Acids Research* 2010,
4. **38**(Database issue):D401-407.
5. [77]. Elsen S, Swem LR, Swem DL, Bauer CE: **RegB/RegA, a highly conserved**
6. **redox-responding global two-component regulatory system**. *Microbiology*
7. *and Molecular Biology Reviews* 2004, **68**(2):263-279.
8. [78]. Elsen S, Dischert W, Colbeau A, Bauer CE: **Expression of uptake hydrogenase**
9. **and molybdenum nitrogenase in *Rhodobacter capsulatus* is coregulated by**
10. **the RegB-RegA two-component regulatory system**. *Journal of Bacteriology*
11. 2000, **182**(10):2831-2837.
12. [79]. Dubbs JM, Bird TH, Bauer CE, Tabita FR: **Interaction of CbbR and RegA\***
13. **transcription regulators with the *Rhodobacter sphaeroides* cbb promoter-**
14. **operator region \***. *Journal of Biological Chemistry* 2000, **275**(25):19224-
15. 19230.
16. [80]. Santos-Beneit F: **The Pho regulon: a huge regulatory network in bacteria**.
17. *Frontiers in Microbiology* 2015, **6**:402.
18. [81]. Choi S, Jeong G, Choi E, Lee E-J: **A dual regulatory role of the PhoU protein**
19. **in *Salmonella Typhimurium***. *mBio* 2022, **13**(3):e00811-00822.
20. [82]. Monzon V, Paysan-Lafosse T, Wood V, Bateman A: **Reciprocal best structure**

# hits: using AlphaFold models to discover distant homologues.

1. *Bioinformatics Advances* 2022, **2**(1):vbac072.
2. [83]. diCenzo GC, Sharthiya H, Nanda A, Zamani M, Finan TM: **PhoU allows rapid**

# adaptation to high phosphate concentrations by modulating PstSCAB

1. **transport rate in *Sinorhizobium meliloti***. *Journal of Bacteriology* 2017,
2. **199**(18).
3. [84]. Li Y, Zhang Y: **PhoU is a persistence switch involved in persister formation**
4. **and tolerance to multiple antibiotics and stresses in *Escherichia coli***.
5. *Antimicrobial Agents and Chemotherapy* 2007, **51**(6):2092-2099.
6. [85]. Herbst FA, Dueholm MS, Wimmer R, Nielsen PH: **The proteome of**
7. ***Tetrasphaera elongata* is adapted to changing conditions in wastewater**
8. **treatment plants**. *Proteomes* 2019, **7**(2).
9. [86]. Willsky GR, Malamy MH: **Characterization of two genetically separable**
10. **inorganic phosphate transport systems in *Escherichia coli***. *Journal of*
11. *Bacteriology* 1980, **144**(1):356-365.
12. [87]. Martín JF, Liras P: **Molecular mechanisms of phosphate sensing, transport**
13. **and signalling in *Streptomyces* and related *Actinobacteria***. *International*
14. *Journal of Molecular Sciences* 2021, **22**(3).
15. [88]. Gebhard S, Tran SL, Cook GM: **The Phn system of *Mycobacterium smegmatis*:**
16. **a second high-affinity ABC-transporter for phosphate**. *Microbiology* 2006,
17. **152**(11):3453-3465.
18. [89]. Saunders AM, Mabbett AN, McEwan AG, Blackall LL: **Proton motive force**

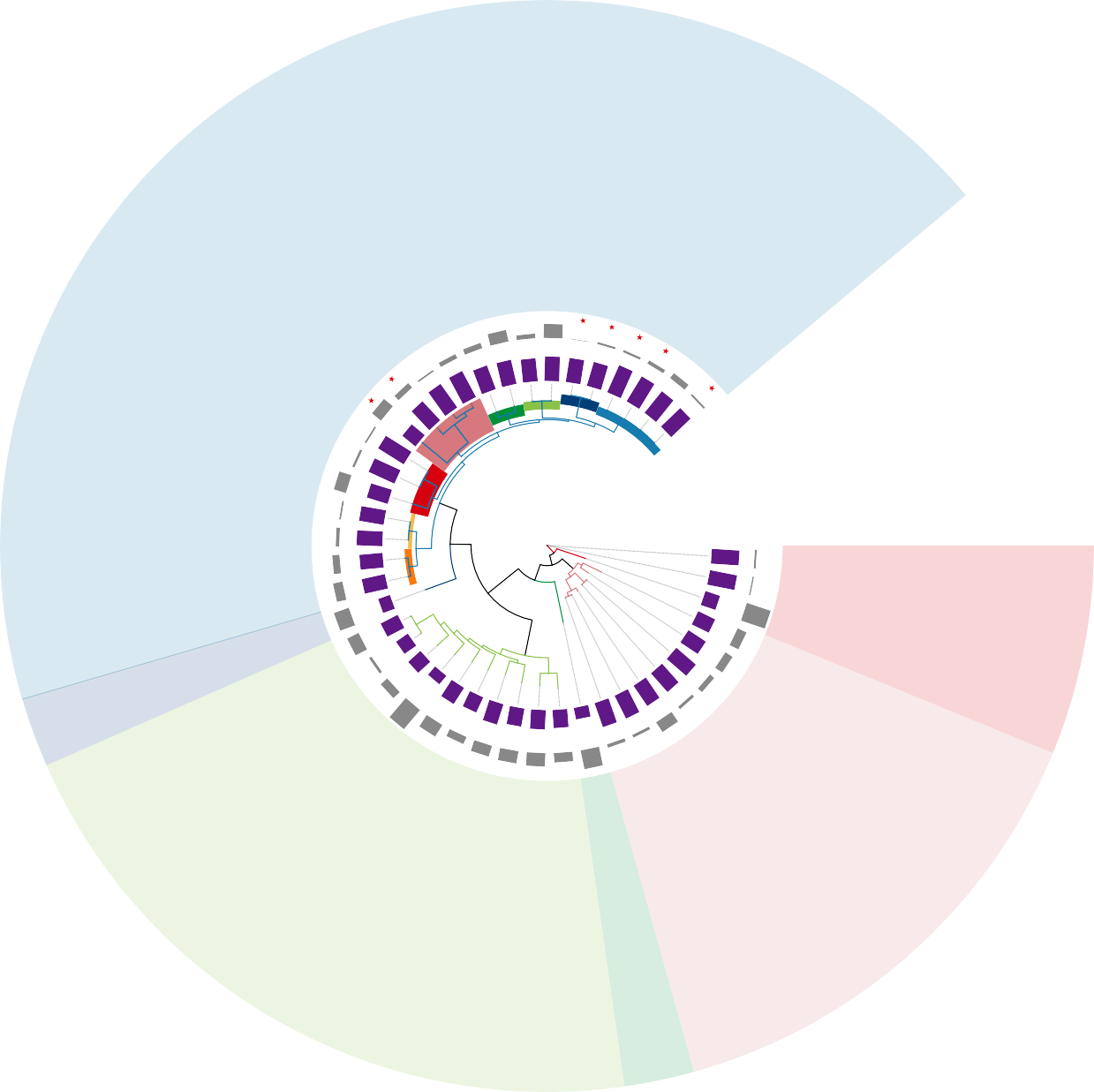
# generation from stored polymers for the uptake of acetate under anaerobic

1. **conditions**. *FEMS Microbiology Letters* 2007, **274**(2):245-251.
2. [90]. Wang Z, Guo F, Mao Y, Xia Y, Zhang T: **Metabolic characteristics of a**
3. **glycogen-accumulating organism in *Defluviicoccus* cluster II revealed by**
4. **comparative genomics**. *Microbial Ecology* 2014, **68**(4):716-728.
5. [91]. McIlroy SJ, Albertsen M, Andresen EK, Saunders AM, Kristiansen R,
6. Stokholm-Bjerregaard M, Nielsen KL, Nielsen PH: **‘*Candidatus***

# *Competibacter*’-lineage genomes retrieved from metagenomes reveal

|  |  |
| --- | --- |
| 988 | **functional metabolic diversity**. *The ISME Journal* 2014, **8**(3):613-624. |
| 989 |  |
| 990 | **Figure 1.** A phylogenetic tree of 43 Rhodocyclaceae members was built based on the |
| 991 | concatenation of 59 single-copy genes. The genomes in red were recovered from our |
| 992 | lab-scale reactors. SSA1, SSB1, and SCUT-1 were recovered in our previous work [43, |
| 993 | 48]. SCELSE-5, SCELSE-7, SCELSE-10, and SCUT-2 were recovered from three of |
| 994 | our lab-scale EBPR reactors (Supplementary Materials). The purple bars represent the |
| 995 | number of shared orthogroups. The gray bars represent the number of unassigned genes. |
| 996 |  |
| 997 | **Figure 2. a**–**b**, The number of gene clusters at different frequencies in the pan |
| 998 | Rhodocyclaceae genome (**a**) and the pan *Ca.* Accumulibacter genome (**b**). **c**–**d**, The |
| 999 | proportion of clusters at different frequencies in the pan Rhodocyclaceae genome (**c**) |
| 1000 | and the pan *Ca.* Accumulibacter genome (**d**). **e**–**f**, The proportion of different average |
| 1001 | gene copies per genome in the pan Rhodocyclaceae genome (**e**) and the pan *Ca.* |
| 1002 | Accumulibacter genome (**f**). In each orthogroup, the average gene copies per genome |
| 1003 | are defined as the number of genes divided by the number of genomes. |
| 1004 |  |
| 1005 | **Figure 3**. **a**, Using the genome integrity estimate, about 99.94% of the core genes could |
| 1006 | be identified with a cut-off value of 18. Only gene families that appear in ≥18 *Ca.* |
| 1007 | Accumulibacter genomes are considered core genes. **b**, A Venn diagram describing the |
| 1008 | numbers of core genes and lineage-specific genes in the pan *Ca.* Accumulibacter |
| 1009 | genome. **c**, The number of core genes observed at different cutoff values. **d**, The number |
| 1010 | 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. |



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Taxa  Tree scale 0.07 *Ca.* Accumulibacter | *Dechloromonas* | *Thauera* | Clades  IIC | IIB | IIF | IA |
| *Propionivibrio* | *Zoogloea* | *Azoarcus* | IIH | IID | IIA | IC |

**a** 40000

38000

35000

9000

8000

Number of clusters

7000

6000

5000

4000

3000

2000

1000

0

1 4 7 10 13 16 19 22 25 28 31 34 37 40 43

**b** 14000

12000



3000

Number of clusters

2000

1000

0

1 3 5 7 9 11 13 15 17 19 21

Number of genomes Number of genomes

**c e f**

**d**

2–5:

31.76%

1: 54.74%

1: 65.32%

2–8:

28.1%

9–16: 3.16%

6–9: 3.38%

10–14: 2.09%

15–17: 1.14%

1: 95.6%

1: 93.84%

17–24: 1.65%

25–36: 0.17%

>36: 1.06%

>17: 6.88% 2: 2.69%

3: 1.24%

>3: 0.46%

2: 4.08%

3: 1.44%

>3: 0.63%

**a** 100

Number of core genes (×103)

95

90

Core genes (%)

85

80

75

70

65

60

20 15 10 5 0

99.94

99.30

93.95

66.29

Genomes

**b**

IIA

IIC

20 15 10 5

Genomes

14 **d**

12

IA

**c**

25

IC

IIH

Number of core genes Number of additions

629

20

1890

328

15

1969

Core 1725

2659 IIF

10

1280 1990

1600

5

IIB IID

0

Number of additions (×103)

Number of genes (×102)

10

8

6

4

2

0

20

15

UW1 SCUT-2

10

5

0

Ancestral Derived Lineage Flexible

specific

Gene classification

*Thauera/Azoarcus*



*Zoogloea ramigera* NBRC 15342

*Dechloromonas*

*Ca.* Propionivibrio aalborgensis

*Ca.* Accumulibacter delftensis SBRS

*Ca.* Accumulibacter meliphilus UWLDOIC

*Ca.* Accumulibacter regalis BA93

*Ca.* Accumulibacter regalis UW8

*Ca.* Accumulibacter aalborgensis AALB *Ca.* Accumulibacter phosphatis UW5 *Ca.* Accumulibacter phosphatis UW1 *Ca.* Accumulibacter SCELSE-10

*Ca.* Accumulibacter similis SSB1 *Ca.* Accumulibacter conexus UW7 *Ca.* Accumulibacter conexus UW13

*Ca.* Accumulibacter necessarius UW12 *Ca.* Accumulibacter proximus BATAC285 *Ca*. Accumulibacter propinquus BAT3C415

*Ca*. Accumulibacter propinquus MAXAC027

*Ca*. Accumulibacter SCELSE-7

*Ca*. Accumulibacter SCELSE-5

*Ca*. Accumulibacter cognatus SCUT-2 *Ca*. Accumulibacter cognatus SSA1 *Ca*. Accumulibacter cognatus Bin19

*Ca*. Accumulibacter cognatus SCUT-1



Tree scale: 0.1

Taxa

IIC IIF

IIH IIB IID

IIA IA

IC

Lost Gained Present

>300 >1000 >4500

100–300 500–1000 4000–4499

<100 <500 <4000

1. Cell motility



Cellular community - prokaryotes

Cell growth and death

Transport and catabolism

Signal transdution Membrane transport Replication and repair

Folding, sorting and degradation

Translation Transcription

Pathway

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

1. 12

11

10

9

8

7

6

Derived

5

4

3

2

1

0

−1

6

5

4

Derived

3

2

1

0

0 5 10 15 20 25 30 35



Amino sugar and nucleotide sugar metabolism

Starch and sucrose metabolism

Glycolysis/Gluconeogenesis

Pyruvate metabolism

Propanoate metabolism

Oxidative phosphorylation

Methane metabolism

Butanoate metabolism

Galactose metabolism Inositol

Glyoxylate and dicarboxylate

metabolism

Carbon fixation pathways in prokaryotes

phosphate Carbon fixation in photosynthetic organisms

metabolism 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

Ancestral

MAP Carbohydrate metabolism Energy metabolism



40 45 50

7

6

Purine metabolism

Glycine, serine, and threonine metabolism

Glycerophospholipid

metabolism

Phenylalanine metabolism

Histidine metabolism

Tyrosine metabolism

Glycerolipid metabolism Biosynthesis of

Alanine, aspartate, and glutamate metabolism

Cysteine and methionine metabolism

unsaturated fatty acids

Primary bile

acid biosynthesis

Tryptophan metabolism Lysine degradation

Arginine biosynthesis

Valine, leucine and isoleucine degradation

Fatty acid degradation

Pyrimidine metabolism

Lysine biosynthesis

Arginine and proline metabolism

Valine, leucine, and

isoleucine biosynthesis Fatty acid

biosynthesis alpha-Linolenic acid metabolism

Phenylalanine, tyrosine, and tryptophan biosynthesis

5

4

Derived

3

2

1

0

0 5 10 15 20 25 30 35

Ancestral

MAP Lipid metabolism Nucleotide metabolism Amino acid metabolism

28



Two-component system

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

ABC transporters

Aminoacyl-tRNA RNA polymerase biosynthesis

Ethylbenzene degradation

Ribosome

26

24

22

20

18

16

Derived

14

12

Styrene degradation

10

8

Peroxisome

6

4

2

0

0 5 10 15 20



Folate biosynthesis O-Antigen nucleotide sugar biosynthesis

Taurine and hypotaurine metabolism

Glutathione metabolism

Porphyrin metabolism

Biotin metabolism

Thiamine metabolism

Selenocompound metabolism

Nicotinate and nicotinamide metabolism Pantothenate and CoA biosynthesis

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

Ancestral

25 30 35

0 10 20 30 40 50 60 70 80

Ancestral



MAP Metabolism of other amino acids Glycan biosynthesis and metabolism Metabolism of cofactors and vitamins

Metabolism of terpenoids and polyketides



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

1. 40

Concentration of PHA or Glycogen (mg per g SS)

32

24

16

8

0

0

**b**

1.0

Expression changes

100 **c**

80

3−-P (mg L−1)

4

60

of PO

40

Concentration

20

0

40 80 120 160 200 240 280

Time of cycle (min)

Cluster 1

Cluster 2

1.0

Group

PYG,glgP; glycogen phosphorylase hns; DNA-binding protein H-NS

gapdh; glyceraldehyde-3-phosphate dehydrogenase TC.SULP; sulfate permease

HSP2-2 Heat shock protein

phaE; multicomponent K+:H+ antiporter subunit E

HSP2-1 Heat shock protein

glgC;glucose-1-phosphate adenylyltransferase GDH2; glutamate dehydrogenase

TC.SULP; sulfate permease

mch, mcd; 2-methylfumaryl-CoA hydratase

poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaC kch, trkA, mthK, pch; voltage-gated potassium channel comEC;competence protein ComEC pgm;phosphoglucomutase

23S rRNA (uridine2552-2'-O)-methyltransferase ppk2; polyphosphate kinase

0.0

1.5

1.0

Expression changes

0.0

1.5

1.0

Cluster 3

alkA; DNA-3-methyladenine glycosylase Il K-7182;CBS domain-containing protein epsF; protein tyrosine kinase modulator glgA; starch synthase

K-78-7; uncharacterized protein

asnB, ASNS;asparagine synthase pqqL pqqL; zinc protease Exopolyphosphatase

ACAT, atoB; acetyl-CoA C-acetyltransferase PYG,glgP; glycogen phosphor ylase

alaC; alanine-synthesizing transaminase

1.5

1.0

0.5

0.0

−0.5

−1.0

−1.5

0.0

1.5

1.0

Expression changes

0.0

0.0

1.5

Cluster 4

Time

mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerasett

corA; magnesium transporter

nemA; N-ethylmaleimide reductase

wzb, etp; low molecular weight protein-tyrosine phosphatase wza, gfcE; polysaccharide biosynthesis/export protein

degP, htrA; serine protease Do ligA, ligB; DNA ligase(NAD+) ald; alanine dehydrogenase

membrane fusion protein, type I secretion system etk-wzc; tyrosine-protein kinase Etk/Wzc

hasF, prtF; outer membrane protein, type I secretion system

branched-chain amino acid transport system substrate-binding protein PGK, pgk; phosphoglycerate kinase

phou homolog

phaE; poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaE poly[(R)-3-hydroxyalkanoate] polymerase subunit PhaC

1.5

Cluster 5

Time

**d**

AHLs

K+ Mg2+

H+ PO 3−

4

**Glycogen**

4

**LuxR**

Pit

Pit

Sulfate

2.4.1.1

2.4.1.21

PO 3−

Phospholipid

Heme Iron(III) Zinc

α-D-glucose-6P

5.3.1.9

β-D-fructose-6P

3.1.3.11 2.7.1.90

β-D-fructose-1,6P2

4.1.2.13

Glyceraldehyde-3P

K+

*PPX* 3.6.1.11

|  |  |  |
| --- | --- | --- |
| Mg2+ |  | PO4  3− |

**Poly-P**

ADP

H+

ATP PO 3−

F1F0-ATPase

4

ADP

1.2.1.12

PO 3− PO 3−

Cobalt

Nickel

Glycerate-1,3P2

2.7.2.3

Glycerate-3P

4

ATP

PhoU

4

PstSABC

5.4.2.12 5.4.2.11

Tungstate

K+ limitation

4.1.1.32

Glycerate-2P 4.2.1.11

Phosphoenol-pyruvate

2.7.1.40

Pyruvate

**PHB**

PhoRB

2.3.1.12

1.2.4.1

**PHV**

**Acetate**

H+

**ActP**

6.2.1.1

**Acetate**

**Acetyl-CoA**

S-CoA

**Oxaloacetate** 2.3.3.1

1.1.1.37

3HB-CoA

3HV-CoA

Citrate

4.2.1.3

Isocitrate

Anaerobic

Malate

**TCA cycle**

2.3.3.9

4.2.1.2

Glyoxylate

4.1.3.1

1.1.1.42

**Ketoglutarate**

1.2.4.2

Aerobic

High transcription non-HGT

**Fumarate**

Succinate

2.3.1.61

Succinyl-CoA

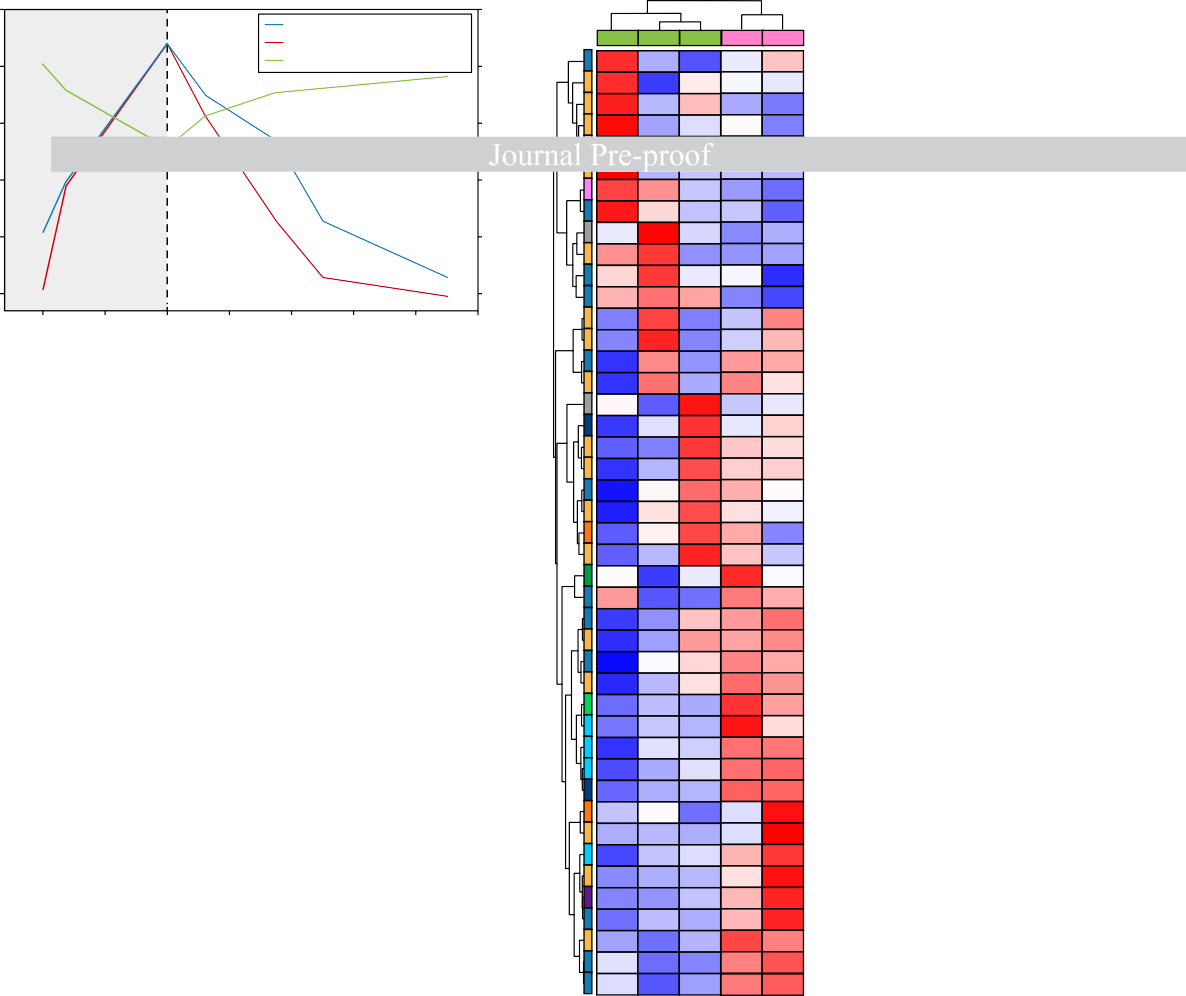
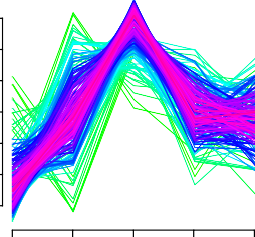
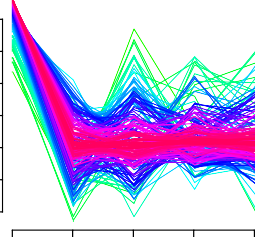
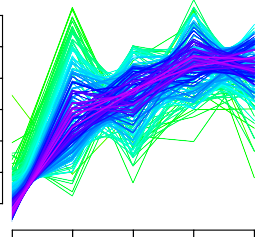
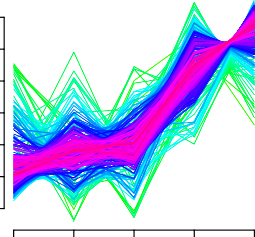
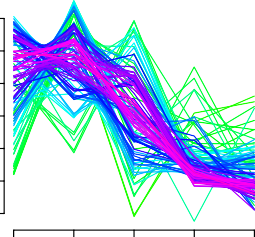
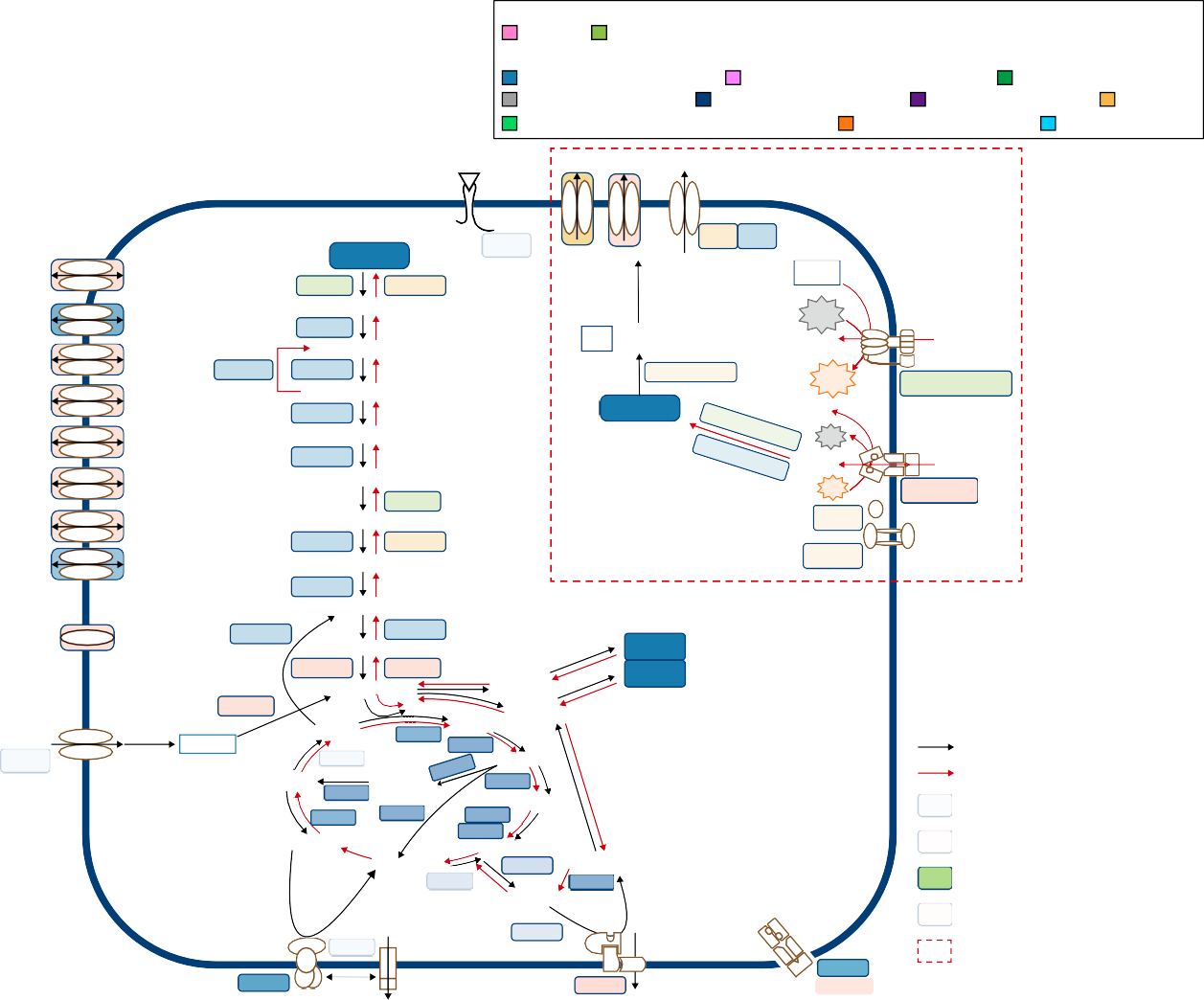
5.4.99.2 **Propionyl-CoA**

Low transcription non-HGT

6.2.1.5

6.4.1.3

High transcription HGT



PHA (mg per g SS) PO 3−-P (mg L−1)

Glycogen (mg per g SS)

4

Group

Aerobic

Pathway

Anaerobic

Carbohydrate metabolism Folding, sorting and degradation

Nucleotide metabolism

Energy metabolism Replication and repair Membrane transport Unknown

Xenobiotics biodegradation and metabolism Amino acid metabolism Signal transduction

Frd

H+

1.3.5.1

e−

Methylmalonyl-CoA

4.1.1.41

Mmd

Na+

GltI GltL GltJ GltK

Low transcription HGT P cycling pathway



# 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: