Metagenomic analysis of phosphorus removing sludge communities.

Héctor García Martín[1], Natalia Ivanova[1], Victor Kunin[1], Falk Warnecke[1], Kerrie Barry[1], Alice C. McHardy[4], Christine Yeates[2], Shaomei He[3], Asaf Salamov[1], Ernest Szeto[1], Eileen Dalin[1], Nik Putnam[1], Harris J. Shapiro[1], Jasmyn L. Pangilinan[1], Isidore Rigoutsos[4], Nikos C. Kyrpides[1], Linda Louise Blackall[2], Katherine D. McMahon[3], and Philip Hugenholtz[1*]

[1] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.

[2] Advanced Wastewater Management Centre, University of Queensland, St Lucia, 4072, Queensland, Australia.

[3] Civil and Environmental Engineering Department, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA.

4 IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA

*corresponding author: phugenholtz@lbl.gov

Abstract

Enhanced Biological Phosphorus Removal (EBPR) is not well understood at the metabolic level despite being one of the best-studied microbially-mediated industrial processes due to its ecological and economic relevance. Here we present a metagenomic analysis of two lab-scale EBPR sludges dominated by the uncultured bacterium, "*Candidatus* Accumulibacter phosphatis". This analysis sheds light on several controversies in EBPR metabolic models and provides hypotheses explaining the dominance of *A. phosphatis* in this habitat, its lifestyle outside EBPR and probable cultivation requirements. Comparison of the same species from different EBPR sludges highlights recent evolutionary dynamics in the *A. phosphatis* genome that could be linked to mechanisms for environmental adaptation. In spite of an apparent lack of phylogenetic overlap in the flanking communities of the two sludges studied, common functional themes were found, at least one of them complementary to the inferred metabolism of the dominant organism. The present study provides a much-needed blueprint for a systems-level understanding of EBPR and illustrates that metagenomics enables detailed, often novel, insights into even well-studied biological systems.

Excessive inorganic phosphate (Pi) supply to freshwater negatively affects water quality and ecosystem balance through a process known as eutrophication[1]. Limitations on allowable Pi discharges from municipal and industrial sources via wastewater treatment have proven effective in reducing Pi levels in many waterways[2]. Increasingly stringent Pi limits for effluent wastewater are expected in the future and hence efficient and reliable Pi removal methods are required. Due to the massive quantity of wastewater treated daily (more than 120 billion liters in the US alone[3]), any improvement in existing methods should have tangible economic and ecological consequences.

Enhanced Biological Phosphorus Removal (EBPR) is a treatment process in which microorganisms remove Pi from wastewater by accumulating it inside their cells as polyphosphate. These polyphosphate-accumulating organisms (PAOs) are then allowed to settle in a separate tank (clarifier), leaving the effluent water largely Pi-depleted. EBPR is more economical in the long term[2] and has a lower environmental impact[4] than traditional (chemical) Pi removal[5], but is prone to unpredictable failures due to loss or reduced activity of microbial populations responsible for Pi removal[6]. This is primarily because the design process is highly empirical due to an incomplete understanding of sludge microbial ecology. Environmental engineers and microbiologists have been studying EBPR since its introduction in municipal wastewater treatment plants over thirty years ago[5] with the goal of making it a more reliable industrial process. Typically, EBPR is studied in lab-scale sequencing batch reactors (SBRs) where the microbial community can be better monitored and perturbed, and PAOs can be enriched to much higher levels than in full scale systems[7].

For thirty years, the bacterial genus *Acinetobacter* was incorrectly assumed to be primarily responsible for EBPR based on cultivation studies[8-10]. Only recently have culture-independent methods pointed to "*Candidatus* Accumulibacter phosphatis", a member of the order *Rhodocyclales*, as the principal agent in acetate-fed EBPR[11-13]. *A. phosphatis* has yet to be obtained in axenic culture despite continuing efforts but can be enriched up to 85% of the community in lab-scale bioreactors[14]. Shotgun sequencing applied directly to environmental samples has recently demonstrated that near complete

genomes can be obtained for dominant populations in a community without the need for cultivation[15-16]. Therefore, it was anticipated that a near complete genome could be obtained for *A. phosphatis* from lab-scale EBPR enrichment cultures, allowing a comprehensive metabolic reconstruction. Comparative analysis of two sludge samples should also lend insights into recent evolutionary dynamics of this important PAO. Moreover, shotgun sequencing gives a metagenomic context for dominant organisms by providing low-level genomic coverage of many other community members suitable for a gene-centric analysis that could highlight habitat-specific metabolic traits[17].

RESULTS

EBPR community structure

Sludge samples were obtained from two lab-scale SBRs that had been performing EBPR successfully for several months; one from Madison, Wisconsin (US) and the other from Brisbane, Australia (OZ). Each SBR was independently seeded from a local wastewater treatment plant. Despite significant differences in operating conditions, including different volatile fatty acid (VFA; US-acetate, OZ-propionate) feeds, sludge volume and solids retention time (Supplementary Table 1), *Accumulibacter* species dominated both sludges, comprising ~80% (US) and ~60% (OZ) of the biomass, as determined by fluorescence in situ hybridization[14,18](FISH).

Approximately 98 and 78 Mbp of shotgun sequence data were obtained from the US and OZ sludge respectively. The US *A. phosphatis* genome is estimated to be 5.6±0.2 Mbp in size (Supplementary Fig. 1) with an average GC content of 63% and average read depth of 8X and 5X respectively for the US and OZ assemblies. A near complete set (97.2%) of essential genes (Supplementary Table 2), typically not co-localized in bacterial genomes, was identified in the higher depth *A. phosphatis* scaffolds of the US JAZZ assembly suggesting that the possibility of incorrectly inferring the absence of pathways in this organism was low. Interestingly, the US and OZ *A. phosphatis* genomes are >95% identical at the nucleotide level over 79% of the reconstructed US genome (Supplementary Fig. 1) indicating that they are closely related strains of the same species.

Low abundance *A. phosphatis* strains were also detected in both metagenomes that were up to 15% divergent at the nucleotide level from the dominant strains. This raises the question as to whether the dominant strains have become EBPR specialists in lab-scale systems or whether the low abundance strains may be dominant at other sampling times or in other lab-scale EBPR sludges.

Thirteen 16S rRNA phylotypes were identified on contigs of two or more reads in each Phrap assembly with the only overlap being *A. phosphatis* (Fig. 1). However, many US and OZ phylotypes clustered into broader phylogenetic groups, e.g. *Xanthomonadales*, *Flavobacteriales* and *Rhizobiales* (Fig. 1), suggesting the possibility of common functional themes in related flanking populations.

Metabolic reconstruction of *Accumulibacter phosphatis*

Several metabolic models for EBPR have been proposed based on gross biochemical measurements of lab-scale systems. The consensus of these models is that Pi is removed from wastewater by uptake into PAOs and conversion into polyphosphate during the aerobic period. These PAOs then break the phosphodiester bonds of the stored polyphosphate to provide an energy source for taking up and storing available VFA (mostly acetate and propionate) as polyhydroxyalkanoates (PHAs, see Fig. 2) during the anaerobic period. Efficiently sequestering VFA during the anaerobic period is thought to give the PAOs a selective advantage over other members of the community for subsequent growth and replication in the aerobic period, allowing it to dominate lab-scale EBPR sludges.

The genome coverage of the dominant *A. phosphatis* populations was sufficiently complete to confidently infer presence and absence of pathways and thereby allow a comprehensive metabolic reconstruction (Supplementary Table 2). Figure 2 highlights the major metabolic pathways likely to be used by *A. phosphatis* during the anaerobic and aerobic phases of the EBPR cycle.

A. phosphatis shuttles Pi across its plasma membrane (out anaerobically, in aerobically) via low affinity transporters encoded by two sets of transporter genes and

high affinity Pi transporters encoded by three sets of transporter genes (Fig. 2). However, due to feedback inhibition[19], we anticipate that the high affinity transporters should only be active at the end of the aerobic period (Fig. 2B) when Pi concentrations are at their lowest. The ability to scavenge relatively low levels of Pi may contribute to superior EBPR performance. The Pi transported into the cell during the aerobic period can be synthesized into polyphosphate via ATP (Fig. 2B). In the anaerobic phase, polyphosphate can be used directly to synthesize ATP or be degraded into Pi for ATP production via V and F type ATPases. The ATP generated is then used in PHA production.

Arguably, the least well understood component of EBPR metabolism is the source of the reducing power (NAD(P)H) required for PHA production in the anaerobic phase. NAD(P)H production via glycogen degradation is insufficient to explain the observed levels of PHA in acetate-fed systems[20-22]. It has been suggested that the tricarboxylic acid (TCA) cycle operates in the anaerobic phase to provide the extra reducing power[23-25]. However, no explanation has been proposed for the necessary re-oxidation of reduced quinones produced by succinate dehydrogenase (Fig. 2A) in the absence of electron acceptors[26]. We propose that quinone is re-oxidized by a novel cytochrome *b/b6*. This protein appears to be a fusion of a cytochrome *b/b6* with 5 transmembrane helices and a soluble NAD(P)- and flavin-binding domain, a domain configuration that is currently unique in public sequence databases (Fig. 3A). Since conventional membrane and soluble NAD(P)H-quinone dehydrogenases are present in the genome, we suggest that this fusion protein functions in reverse as a quinol-NAD(P) reductase, at the expense of the proton gradient (Fig. 3B). A similar uphill electron transfer through a bc_1 and NADH-Q oxidoreductase complex has been experimentally shown in *Thiobacillus ferrooxidans*[27]. Full anaerobic functioning of the TCA cycle enabled by the novel cytochrome would allow *A. phosphatis* to outcompete other species for VFA storage and may explain why *A. phosphatis* dominates EBPR communities.

An alternative scenario to full anaerobic TCA function is the operation of a split TCA cycle, since fumarate reductase is also present (Fig. 2A, dashed line). This would

result in PHV accumulation via methylmalonyl-CoA (Fig. 2A) and may explain the small amount of PHV (5-20% of PHAs) usually observed in acetate-fed EBPR, which is not accounted for in most EBPR models[28]. Gene expression or proteomic data that leverages the metagenomic data will determine which pathway or pathways are being used by *A. phosphatis* to generate the extra reducing power.

Another contentious point in EBPR metabolic models is the pathway used for glycogen degradation, Embden Meyerhof (EM) or Entner Doudoroff (ED). This has a substantial impact on the cellular energy budget because the EM pathway yields more ATP. All EM pathway genes are present in the dominant *A. phosphatis* strains. In contrast, the key genes for the ED pathway were not found as well as enzymes typically feeding into this pathway, indicating that *A. phosphatis* likely only has the EM pathway available to degrade glycogen. NMR studies of EBPR sludges indicate that the ED pathway may be dominant[29,30], suggesting that the sludges analyzed did not contain *A. phosphatis*, or that other *A. phosphatis* strains or other *Accumulibacter* species may have the ED pathway. The latter explanation is less likely as the closest sequenced relatives of *A. phosphatis*, *Dechloromonas aromatica* and *Azoarcus* sp. EbN1, also lack the key ED genes. However, the EBPR sludge studied by Hesselmann *et al.* [30], implicating ED as the dominant pathway, is likely to have been dominated by *Accumulibacter*. This apparent contradiction between the genomic evidence and the NMR data will need to be addressed by further experimental work such as proteomics.

The production of extracellular polymeric substances (EPS) is essential for the survival of *A. phosphatis* in the wastewater treatment environment[31]. EPS bind *A. phosphatis* cells in dense "flocs", which are necessary for settling in the clarifier. Non-settling cells are washed out of the system. Consistent with the vital role of EPS, there are at least two EPS gene clusters (25-38 kbp) in the US *A. phosphatis* genome (Supplementary Fig. 1). The gene complements of the clusters strongly suggest that exopolysaccharide- and glycoprotein- containing EPS types are produced, each with distinct physical and chemical properties. Interestingly, the EPS clusters are conspicuously volatile between the two otherwise closely related dominant strains in the

US and OZ sludges. We speculate that EPS clusters are modular structures that are interchangeable via non-homologous recombination to allow rapid adaptation to local conditions, such as varying influent composition and temperature. Other possible ecological implications of volatile EPS clusters are discussed elsewhere (V. Kunin *et al.*, in preparation).

It is interesting that respiratory nitrate reductase (*nar*) appears to be absent from *A. phosphatis* since experimental evidence indicates that both acetate- and propionate-fed EBPR sludges dominated by *A. phosphatis* can denitrify[32]. The genome does encode the rest of the denitrification pathway from nitrite onwards and a dissimilatory nitrate reductase (*nap*). However, the *nap* appears to lack the subunit that usually functions as a quinol reductase, suggesting that it may not be able to function as part of the electron transport chain. This casts doubt on the ability of the dominant *A. phosphatis* strain to reduce nitrate, although it is possible that other strains encode respiratory nitrate reductase genes. If *A. phosphatis* does not reduce nitrate, flanking EBPR species must perform this essential task under anoxic conditions. Although the EBPR sludges in the present study were not grown on nitrate, *nar* was identified on small contigs (< 2 kbp), derived from low abundance community members in both the US and OZ sludges. We predict that these populations would increase in relative abundance if the sludges were operated under anaerobic/anoxic conditions with nitrate. Nitrate reducing populations would occupy an important ecological niche under these conditions by supplying the dominant *A. phosphatis* population with nitrite for respiration.

Not all of the inferred metabolic capabilities of *A. phosphatis* appear to be related to its EBPR lifestyle. One of the most surprising findings is a full complement of genes for nitrogen fixation, an energetically very expensive process[33]. The key genes for fixing CO_2 are also present, including phosphoribulokinase and the large subunit of rubisco. Since wastewater contains high levels of fixed nitrogen and readily available organic carbon, it is unlikely that these genes will be expressed in EBPR sludge. This suggests that *A. phosphatis* is adapted to carbon- and nitrogen-limited habitats. Furthermore, the presence of the high affinity Pi transporters would allow this bacterium to function in

phosphorus-limited habitats. This implies the existence of *A. phosphatis* reservoirs in nutrient-limited habitats such as freshwater. *A. phosphatis* also has genes for flagella biosynthesis, although no flagella have been observed for this organism in EBPR. If the environmental reservoir is water, then flagella may be expressed in these habitats to facilitate the ability of *A. phosphatis* to move towards sources of limiting nutrients. These reservoirs may serve as sources for reseeding of EBPR communities. Preliminary studies using *Accumulibacter*-specific PCR confirms the hypothesis that this organism is indeed present in freshwater and associated sediments (V. Kunin *et al.*, in preparation).

The inferred ability of *A. phosphatis* to fix nitrogen suggests a selection strategy for isolating this bacterium, as has recently been demonstrated for *Leptospirillum ferrodiazotrophum*[34]. A. phosphatis also appears to have an unusual cobalt dependence: it has only cobalamin-dependent versions of methionine synthase and ribonucleotide reductase and a full complement of genes for cobalamin biosynthesis. A nitrogen-free selective growth medium would therefore need to be supplemented with cobalt or cobalamin to support growth of *A. phosphatis*. Preliminary efforts to isolate *A. phosphatis* in a nitrogen-free, colbalt-rich medium have resulted in enrichment but not isolation (data not shown).

EBPR-related metabolism of dominant flanking populations

In addition to *A. phosphatis*, several flanking species were relatively well represented in the metagenomic datasets resulting in genomic fragments up to 64 kbp in size. The dominant flanking populations in the US and OZ sludges, determined by conserved gene analysis, were a *Xylella*- and *Flexibacter*-like species (0.4% and 2.7% of reads in US phrap contigs containing 16S rRNA genes respectively) and *Thiothrix*-like species (13.8% of reads in OZ phrap contigs containing 16S rRNA genes) (Fig. 1). This level of representation in the datasets allows the presence but not absence of metabolic pathways to be inferred. In the *Thiothrix*-like population, we identified a complete methylcitrate pathway used for propionate degradation. This may explain why *Thiothrix* is the dominant flanking species in the propionate-fed OZ sludge. This hypothesis could be

tested by changing the OZ sludge to an acetate feed and monitoring the effect on the *Thiothrix* population. Conversely, if the US sludge was switched to a propionate feed, we would expect that one or more propionate-utilizing flanking populations would be enriched, although not necessarily *Thiothrix*. This is because we speculate that the composition of the flanking communities is determined by a combination of operating conditions and chance (e.g. presence of a given species in the seeding sludge).

Genes encoding enzymes for carbohydrate polymer hydrolysis and pathways for subsequent monomer degradation were identified in the *Flexibacter*-like (beta-galactose), *Xylella*-like (glucuronic acid) and *Thiothrix*-like (xylose) populations. One of the EPS gene cassettes in the US *A. phosphatis* encodes a gene (UDP-glucose dehydrogenase) for the production of the precursor of glucoronic acid suggesting the presence of this sugar in the EPS. The *Xylella*-like population may therefore be able to degrade this component of the *A. phosphatis* EPS and use it as a food source.

Gene-centric analysis

We performed a gene-centric analysis of the metagenomic data[17] to determine over-represented gene families in EBPR communities relative to other habitats. Genes annotated in the two US sludge assemblies, OZ sludge, acid mine drainage biofilm[15], soil and three whalefall samples[17] were classified in gene families according to sequence similarity and the relative representation of each family was determined. A sizable fraction of the families believed to be important for survival in the EBPR environment from the metabolic reconstruction were over-represented in the sludge datasets. These include genes required for phosphate transport (specific components of both low and high affinity transporters), VFA handling (VFA sodium symporter, and PHA synthetases), anaerobic operation of the TCA cycle (cytochrome *b/b6*) and cobalt uptake (cobalt transporters).

Despite the two EBPR communities having minimal overlap at the species level (apart from *A. phosphatis*), 24 gene families lacking *A. phosphatis* representatives were over-represented in both sludges relative to the other habitats. For instance, a family of

nitrate transporters is overrepresented, supporting the hypothesis that the flanking communities are responsible for nitrate reduction. This suggests operating conditions broadly determine some niches that may be occupied by multiple species. i.e. some functional redundancy appears to be present in the two phylogenetically different flanking communities.

The above examples suggest that other overrepresented gene families with no obvious fit in the present metabolic model merit serious consideration. Most conspicuous amongst these are 11 families annotated as Ca^{2+}-binding proteins related to RTX toxins (repeat toxins) with representatives in *A. phosphatis* and other flanking species. We speculate that they are part of EPS, given the affinity of EPS for cations, and their abundance in overrepresented families suggests an important role. Other overrepresented families with intriguing annotated functions, but no clear role in the present model, include two iron transport families and a family involved in DNA exchange. Unfortunately, around 15% of the overrepresented families in EBPR have no annotation suggesting that many functionally important genes in EBPR remain to be characterized.

DISCUSSION

The determination of the near complete genome of *A. phosphatis* represents a turning point in our understanding of the genetic basis of EBPR. It will enable targeted studies of the enzymes involved in carbon and phosphorus transformation pathways, as well as flux through these pathways. We can now study how gene expression is regulated in response to environmental factors such as concentrations of dissolved oxygen, VFAs, nitrate and phosphate. In short, the metagenome will facilitate investigations of the EBPR transcriptome, proteome and metabolome. We believe this will lead to breakthroughs in metabolic modeling and our ability to predict when and where EBPR will be operating effectively.

METHODS

Both SBRs were seeded from local wastewater treatment plants (Nine Springs Wastewater Treatment Plant in Madison, Wisconsin, and Thornside Sewage Treatment Plant in Queensland, Australia). The SBRs were operated in four cycles of 6h and the hydraulic residence time (HRT) and solids residence time (SRT) were 12 hours and 4 days respectively for the US SBR and 24 hours and 8 days for the OZ SBR. DNA was extracted at the end of the anaerobic period. Three whole genome shotgun libraries, containing inserts of ~3, 8 and 40kb, were created for each of the two sludge DNA samples and sequenced. Both datasets were assembled with Phrap 4 (beta version, http://www.phrap.org/), and JAZZ[35] as a control for assembly and annotation artifacts. A reimplemented version of the JAZZ assembler was subsequently applied to both datasets. Both datasets readily assembled and the largest Phrap contigs obtained from the US and OZ sludges were 170 and 65 kbp respectively. The largest JAZZ scaffold (contigs linked by end pair information) for the US sludge was over 5 Mbp (second JAZZ assembly). However, 26% of US and 32% of OZ reads did not assemble using Phrap, representing low abundance populations in the EBPR sludges. Over 30,000 coding sequences were predicted in each Phrap assembly using *ab initio* gene predictions (fgenesb, http://www.softberry.com). The genomic fragments were binned (classified) using a combination of read depth, % GC content, clade-characteristic features in intrinsic sequence composition (A.K. *et al.*, submitted), sequence similarities to isolate genomes and commonality of *A. phosphatis* between the two sludges. As expected, most of the large fragments, including the 5 Mbp JAZZ scaffold, originated from *A. phosphatis* and a composite genome of this population could be reconstructed based on the JAZZ scaffolds (Supplementary Fig. 1). The assembled metagenomic data was incorporated into the U.S. Department of Energy Joint Genome Institute Integrated Microbial Genomes & Metagenomes (IMG/M) experimental system (www.jgi.doe.gov/m) to facilitate visualization, comparative analyses and metabolic reconstruction of the data in the context of other metagenomic datasets and all publicly available complete microbial genomes. The sequence and annotation of the Phrap assemblies have been deposited in

the NCBI databank under the project accession xxxxxxxx. High quality sequence reads from the project have been deposited in the NCBI trace archive.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgements

We thank Edward Berry for expert advice on cytochromes and Greg Crocetti, Daniel Noguera, Suzan Yilmaz, Paul Wilmes, Phil Bond, Jay Keasling, and Eddy Rubin for helpful discussions. We also thank Aaron Saunders, Huabing Liu, Daniel Gall, Eugene Goltsman, Inna Dubchak, Matt Nolan, Steve Lowry, Alla Lapidus, Bryce Shepherd, Thomas Huber, Khrisna Palaniappan, Frank Korzeniewski and Sam Pitluck for technical assistance and additional analyses, and Chris Detter, Paul Richardson, Tijana Glavina del Rio, Susan Lucas, Alex Copeland, Dan Rokhsar, Igor Grigoriev and Victor Markowitz for facilitating the study. The sequencing for the project was provided by the DOE Community Sequencing Program at JGI (http://www.jgi.doe.gov/CSP/index.html). The National Science Foundation (BES 0332136) supported the efforts of KDM and SH. This work was performed under the auspices of the DOE's Office of Science, Biological and Environmental Research Program; the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract no. DE-AC03-76SF00098; and Los Alamos National Laboratory under contract no. W-7405-ENG-36.

Figures and legends

Fig. 1. Maximum likelihood tree based on partial and complete 16S rRNA genes identified on metagenomic contigs comprising at least two reads. Sludge sequences with >97% identity are clustered into phylotypes shown as circles on the tree. Blue circles indicate

US sludge-derived sequences and red circles, OZ sludge-derived sequences. Circle sizes indicate relative abundance of phylotypes in the metagenomic datasets based on the number of reads comprising each contig on which a 16S rRNA gene was identified. Green shading indicates clusters of US and OZ phylotypes that may be closely enough related to share common metabolic traits. An expanded phylogenetic tree is presented in Supplementary Figure 2 that includes all 16S rRNAs identified in the metagenomic datasets.

Fig 2. EBPR-relevant metabolism inferred from the *A. phosphatis* composite genome. In the **anaerobic phase (A)** acetate and proprionate are stored as four types of PHA; polyhydroxybutyrate (PHB, from acetate only), polyhydroxyvalerate (PHV, from acetate and propionate), polyhydroxy-2-methylbutyrate (PH2MB, from acetate and propionate) and polyhydroxy-2-methylvalerate (PH2MV, from propionate only). PHA production requires energy (ATP) and reducing power (NAD(P)H). ATP (in red) is supplied by polyP degradation and, to a lesser degree, glycogen degradation. NAD(P)H (in blue) is provided by glycogen degradation and the TCA cycle (enabled by a novel quinol reductase). A possible alternative use of the TCA cycle splits it in two branches through the use of fumarate reductase (dashed line). In the **aerobic phase (B)**, when oxygen is available for respiration, acetate is not present in the medium for other species and the PHA reserves of *A. phosphatis* ensures its dominance in the SBR microbial ecosystem. The restoration of polyphosphate reserves via ATP depletes the water of Pi, thus giving rise to Enhanced Biological Phosphorus Removal. The dashed line represents an alternative pathway for PHB degradation[36] for which not all genes have been characterized. The dotted lines leading from the high affinity phosphate transporters (Pst) indicate that these transporters are unlikely to be active for most of the aerobic phase.

Fig 3. Domain structure **(A)** and hypothesized quinol reductase function **(B)** of a novel cytochrome encoded in the *A. phosphatis* genome that would allow anaerobic use of the TCA cycle. It is a fusion of one gene with a cytochrome *b/b6* domain and another gene

with soluble ferrodoxin-, NAD(P)- and flavin-binding domains. We predict that electrons are passed from the reduced quinone via the cytochrome, ferrodoxin and flavin groups to reduce NAD(P)+, at the expense of the proton motive force. Full anaerobic functioning of the TCA cycle enabled by the novel cytochrome would allow *A. phosphatis* to outcompete other species for VFA storage and may explain why *A. phosphatis* dominates EBPR communities.

References

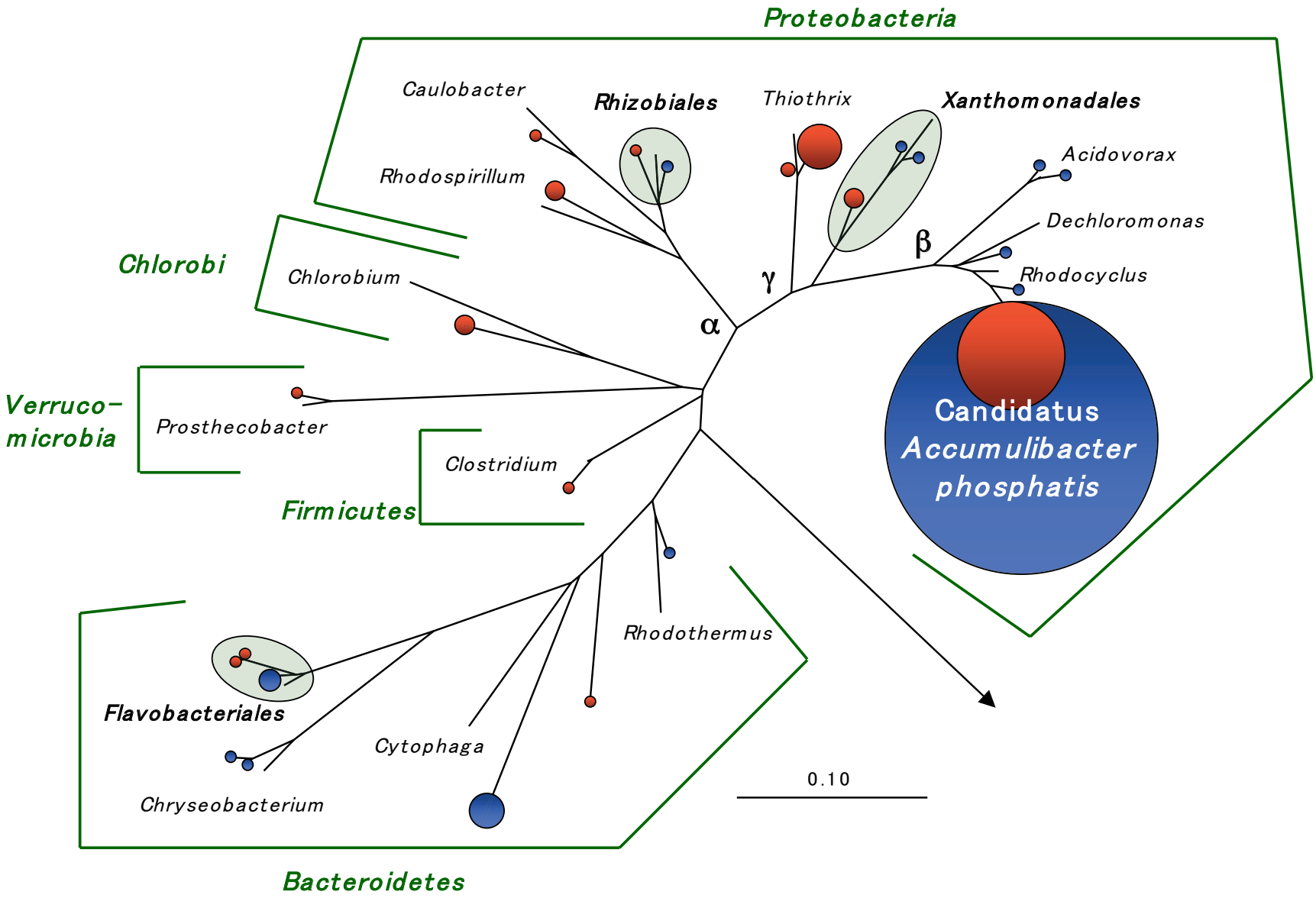
1. Harper, D. Eutrophication of freshwaters. (Chapman and Hall, London; 1991).
2. SCOPE Newsletter, Implementation of the 1991 EU Urban Waste Water Directive and its role in reducing phosphate discharges, (1998).
3. E.P.A., U.S. EPA/OW Clean Water Needs Survey (CWNS) for the United States and U.S. Territories, US EPA/Office of Water, Washington, D.C. (1996).
4. CEEP, in Second International Conference on the recovery of phosphorus from sewage and animal wastes, Noordwijkerhout, Netherlands (2001).
5. Tchobanoglous, G. & Burton, F.L. Wastewater Engineering: Treatment, disposal, and reuse. (McGraw-Hill, New York; 1991).
6. Blackall, L.L., Crocetti, G.R., Saunders, A.M. & Bond, P.L. A review and update of the microbiology of enhanced biological phosphorus removal in wastewater treatment plants. *Antonie Van Leeuwenhoek* **81**, 681-691 (2002).
7. He, S., Gu, A.Z. & McMahon, K.D. Fine-scale differences between Accumulibacter-like bacteria in enhanced biological phosphorus activated sludge. *Water Sci Technol* **54**, In press. (2006).
8. Fuhs, G.W. & Chen, M. Microbiological basis of phosphate removal in the activated sludge process for the treatment of wastewater. *Microb Ecol* **2**, 119-138 (1975).
9. Streichan, M., Golecki, J.R. & Schon, G. Polyphosphate-accumulating bacteria from sewage treatment plants with different processes for biological phosphorus removal. *FEMS Microbiol Ecol* **73**, 113-124 (1990).
10. Deinema, M.H., van Loosdrecht, M.C.M. & Scholten, A. Some physiological characteristics of Acinetobacter spp. accumulating large amounts of phosphate. *Water Sci Technol* **17**, 119-125 (1985).
11. Wagner, M. et al. Development of an rRNA-targeted oligonucleotide probe specific for the genus Acinetobacter and its application for in situ monitoring in activated sludge. *Appl Environ Microbiol* **60**, 792-800 (1994).
12. Hesselmann, R.P., Werlen, C., Hahn, D., van der Meer, J.R. & Zehnder, A.J. Enrichment, phylogenetic analysis and detection of a bacterium that performs

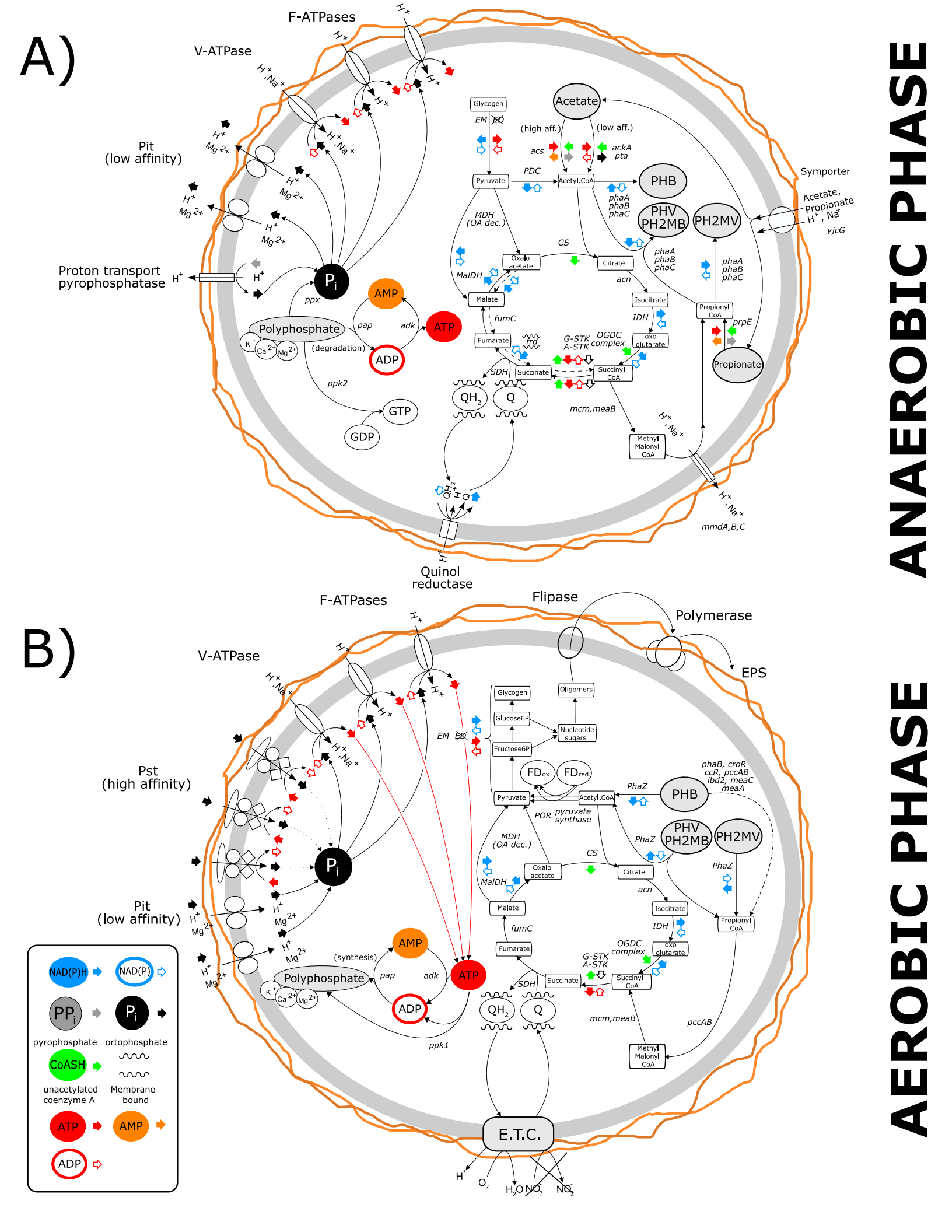
enhanced biological phosphate removal in activated sludge. *Syst Appl Microbiol* **22**, 454-465 (1999).

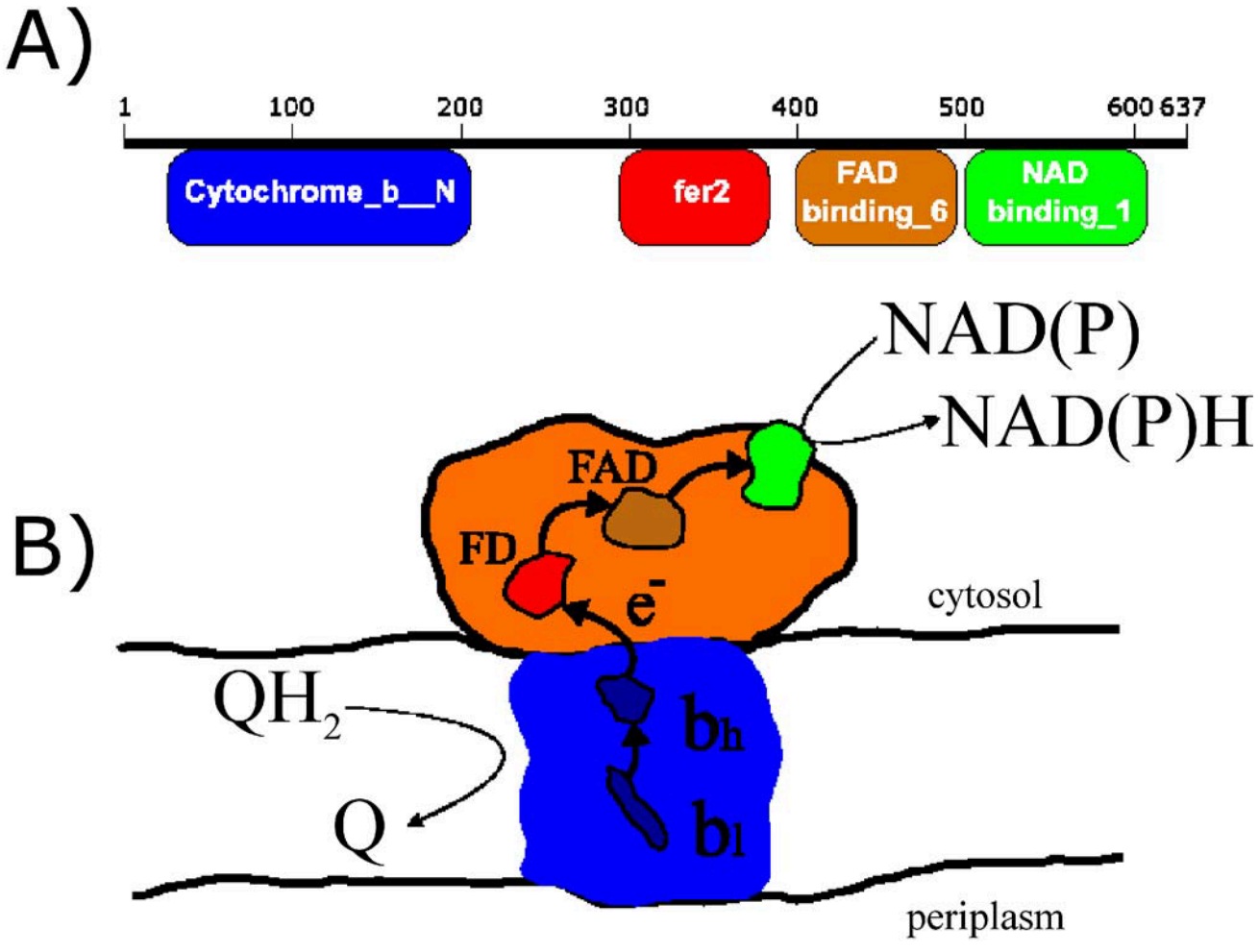
13. Crocetti, G.R. et al. Identification of Polyphosphate-Accumulating Organisms and Design of 16S rRNA-Directed Probes for Their Detection and Quantitation. *Appl Environ Microbiol* **66**, 1175-1182 (2000).
14. McMahon, K.D., Dojka, M.A., Pace, N.R., Jenkins, D. & Keasling, J.D. Polyphosphate kinase from activated sludge performing enhanced biological phosphorus removal. *Appl Environ Microbiol* **68**, 4971-4978 (2002).
15. Tyson, G.W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
16. Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
17. Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
18. Oehmen, A., Saunders, A.M., Vives, M.T., Yuan, Z. & Keller, J. Competition between polyphosphate and glycogen accumulating organisms in enhanced biological phosphorus removal systems with acetate and propionate as carbon sources. *J Biotechnol* (2005).
19. Burns, D.J. & Beever, R.E. Mechanisms controlling the two phosphate uptake systems in Neurospora crassa. *J Bacteriol* **139**, 195-204 (1979).
20. Kortstee, G.J., Appeldoorn, K.J., Bonting, C.F., van Niel, E.W. & van Veen, H.W. Recent developments in the biochemistry and ecology of enhanced biological phosphorus removal. *Biochemistry (Mosc)* **65**, 332-340 (2000).
21. Seviour, R.J., Mino, T. & Onuki, M. The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* **27**, 99-127 (2003).
22. Schuler, A.J. & Jenkins, D. Enhanced biological phosphorus removal from wastewater by biomass with different phosphorus contents, Part III: Anaerobic sources of reducing equivalents. *Water Environ Res* **75**, 512-522 (2003).
23. Pereira, H. et al. Model for carbon metabolism in biological phosphorus removal processes based on in vivo13C-NMR labelling experiments. *Wat Res* **30**, 2128-2138 (1996).

24. Louie, T.M., Mah, T.J., Oldham, W. & Ramey, W.D. Use of Metabolic Inhibitors and Gas Chromatography/Mass Spectrometry to Study Poly-β-Hydroxyalkanoates metabolism involving Cryptic Nutrients in Enhanced Biological Phosphorus Removal Systems. *Wat Res* **34**, 1507-1514 (2000).
25. Lemos, P.C., Serafim, L.S., Santos, M.M., Reis, M.A. & Santos, H. Metabolic pathway for propionate utilization by phosphorus-accumulating organisms in activated sludge: 13C labeling and in vivo nuclear magnetic resonance. *Appl Environ Microbiol* **69**, 241-251 (2003).
26. Mino, T., Van Loosdrecht, M.C.M. & Heijnen, J.J. Microbiology and biochemistry of the enhanced biological phosphate removal process. *Wat Res* **32**, 3193-3207 (1998).
27. Elbehti, A., Brasseur, G. & Lemesle-Meunier, D. First evidence for existence of an uphill electron transfer through the bc(1) and NADH-Q oxidoreductase complexes of the acidophilic obligate chemolithotrophic ferrous ion-oxidizing bacterium Thiobacillus ferrooxidans. *J Bacteriol* **182**, 3602-3606 (2000).
28. Schuler, A.J. & Jenkins, D. Enhanced biological phosphorus removal from wastewater by biomass with different phosphorus contents, Part I: Experimental results and comparison with metabolic models. *Water Environ Res* **75**, 485-498 (2003).
29. Maurer, M., Gujer, W., Hany, R. & Bachmann, S. Intracellular carbon flow in phosphorus accumulating organisms from activated sludge systems. *Wat Res* **31**, 907-917 (1997).
30. Hesselmann, R.P.X., Von Rummell, R., Resnick, S.M., Hany, R. & Zehnder, A.J.B. Anaerobic metabolism of bacteria performing enhanced biological phosphate removal. *Wat Res* **34**, 3487-3494 (2000).
31. Wilen, B.M., Jin, B. & Lant, P. Relationship between flocculation of activated sludge and composition of extracellular polymeric substances. *Water Sci Technol* **47**, 95-103 (2003).
32. Zeng, R.J., Lemaire, R., Yuan, Z. & Keller, J. Simultaneous nitrification, denitrification, and phosphorus removal in a lab-scale sequencing batch reactor. *Biotechnol Bioeng* **84**, 170-178 (2003).

33. White, D. The Physiology and Biochemistry of Prokaryotes. (Oxford University Press, New York; 1995).
34. Tyson, G.W. et al. Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. *Appl Environ Microbiol* **71**, 6319-6324 (2005).
35. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1283-1285 (2002).
36. Korotkova, N., Lidstrom, M.E. & Chistoserdova, L. Identification of genes involved in the glyoxylate regeneration cycle in *Methylobacterium extorquens* AM1, including two new genes, meaC and meaD. *J Bacteriol* **187**, 1523-1526 (2005).


Proteobacteria
Caulobacter
Rhizobiales
Thiothrix
Xanthomonadales
Acidovorax
Rhodospirillum
Dechloromonas
β
γ
Chlorobi
Chlorobium
Rhodocyclus
α
Candidatus
Accumulibacter
phosphatis
Verruco-
microbia
Prosthecobacter
Clostridium
Firmicutes
Rhodothermus
Flavobacteriales
Cytophaga
0.10
Chryseobacterium
Bacteroidetes


A)
ANAEROBIC PHASE
F-ATPases
V-ATPase
Pit
(low affinity)
Proton transport
pyrophosphatase
Symporter
Acetate,
Propionate
H+, Na+
yjcG
Glycogen
EM
Acetate
(high aff.)
(low aff.)
acs
ackA
pta
PDC
Pyruvate
Acetyl.CoA
PHB
phaA
phaB
phaC
PHV
PH2MB
PH2MV
MDH
(OA dec.)
CS
Oxalo
acetate
Citrate
MalDH
acn
Malate
Isocitrate
fumC
IDH
Propionyl
CoA
prpE
Propionate
Fumarate
frd
G-STK
A-STK
OGDC
complex
oxo
glutarate
SDH
Succinate
Succinyl
CoA
QH2
Q
mcm,meaB
Methyl
Malonyl
CoA
mmdA,B,C
Pi
ppx
AMP
ATP
ADP
pap
adk
Polyphosphate
(degradation)
ppk2
GTP
GDP
Quinol
reductase
B)
AEROBIC PHASE
F-ATPases
Flipase
Polymerase
EPS
V-ATPase
Pst
(high affinity)
Pit
(low affinity)
Glycogen
Oligomers
Glucose6P
Nucleotide
sugars
Fructose6P
EM
FDox
FDred
PhaZ
phaB, croR
ccR, pccAB
ibd2, meaC
meaA
Pyruvate
Acetyl.CoA
PHB
POR
pyruvate
synthase
MDH
(OA dec.)
PHV
PH2MB
PH2MV
CS
Oxalo
acetate
Citrate
MalDH
acn
Malate
Isocitrate
fumC
IDH
Propionyl
CoA
Fumarate
OGDC
complex
oxo
glutarate
G-STK
A-STK
Succinate
Succinyl
CoA
SDH
QH2
Q
mcm,meaB
pccAB
Methyl
Malonyl
CoA
Pi
AMP
ATP
ADP
(synthesis)
pap
adk
Polyphosphate
ppk1
E.T.C.
O2
H2O
NO3
NO2
NAD(P)H
NAD(P)
PPi
pyrophosphate
Pi
ortophosphate
CoASH
unacetylated
coenzyme A
Membrane
bound
ATP
AMP
ADP


A)
1
100
200
300
400
500
600 637
Cytochrome_b__N
fer2
FAD binding_6
NAD binding_1
B)
NAD(P)
NAD(P)H
FAD
FD
e-
cytosol
QH2
Q
bh
bl
periplasm

Supplementary Online Material

Metagenomic analysis of phosphorus removing sludge communities.

Héctor García Martín, Natalia Ivanova, Victor Kunin, Falk Warnecke, Kerrie Barry, Alice C. McHardy, Christine Yeates, Shaomei He, Asaf Salamov, Ernest Szeto, Eileen Dalin, Nik Putnam, Harris J. Shapiro, Jasmyn L. Pangilinan, Isidore Rigoutsos, Nikos C. Kyrpides, Linda Louise Blackall, Katherine D. McMahon, and Philip Hugenholtz

Materials and methods

Sequencing Batch Reactor (SBR) operation

The US SBR was inoculated with activated sludge mixed liquor from the Madison, WI, USA Nine Springs Wastewater Treatment Plant (University of Cape Town process), which has been the subject of several previous studies (e.g. Zilles *et al* [1]). The reactor, with a working volume of two liters, was operated in four cycles of 6 h per day, including 130 min. anaerobic phase, 190 min. aerobic phase, 30 min. settling and 10 min effluent withdrawing (1000 mL). The anaerobic and aerobic phases were maintained by sparging nitrogen gas and air into the reactor, respectively. The hydraulic residence time (HRT) was 12 h, and the solids retention time (SRT) was maintained at four days by wasting 500 mL once per day during the aerobic phase. The pH was controlled in the range of 7.0-7.3. The SBR was fed with acetate, casamino acids, and a mineral salts medium with sodium phosphate to achieve a COD:P of 14 (mg COD: mg P)[2,3] and to achieve a sludge non-soluble phosphorus content of 18% (mg P per mg of volatile suspended solids). The SBR had been operating for 11 months at the time of sampling for metagenomic analysis.

The OZ SBR was inoculated with activated sudge from Thornside Sewage Treatment Plant in Queensland, Australia. It had a working volume of 8 liters, was operated in four cycles of 6 h. per day. Each cycle consisted of 150 min anaerobic period, 180 min anerobic period, 30 min. settling and, in 5 min of effluent withdrawing (900 ml). The HRT was 24 h. and SRT was kept at 8 days by sludge wastage during the aerobic

phase. The pH was kept at 7.0 and the SBR operating temperature was 23-24^{o}C during all stages. To achieve anaerobic and aerobic conditions, nitrogen gas and air was bubbled through the liquid respectively. The SBR was fed with propionate and a synthetic feed[4].

The major operating differences between the US and OZ SBRs are summarized in Supplementary Table 1.

DNA extraction

Sludge biomass was collected directly from the US and OZ SBRs at the end of the anaerobic period when polyphosphate inclusions were at their lowest, since polyphosphate interferes with DNA extraction.

US sample: Frozen sludge was thawed on ice, and 0.25 g (wet weight) aliquots were transferred to 2 mL tubes. Bulk genomic DNA was extracted from the aliquots using a series of enzymatic digestions, followed by phenol-chloroform extraction, essentially as described previously[5]. Extracts were then suspended in TE (10 mM Tris, 1 mM EDTA, pH 7.6) and digested with RNase A (10 mg L^{-1} final concentration) at 37 ^{o}C for 30 min. The final concentration and purity of DNA were estimated by spectrophotometry at 260 nm and 280 nm. The integrity of extracted DNA was evaluated by agarose gel electrophoresis using standard methods[6].

OZ sample: DNA was extracted from frozen sludge aliquots (2 ml) using the FastDNA SPIN kit (Qbiogene), following the manufacturer's instructions, and quantified by spectrophotometric analysis at 260 nm.

Library construction and sequencing

Three whole genome shotgun libraries, containing inserts of ~3, 8 and 40kb, were created for each of the two sludge DNA samples. For the library creation of the 3 and 8kb insert libraries, DNA was randomly sheared by a hydroshear, size selected on an agarose gel, extracted and purified. The insert was then end-repaired for blunt-end cloning, size selected on an agarose gel, extracted, and purified a second time. Inserts of approximately 3kb were cloned into pUC18, and 8kb fragments were cloned into pMCL200. The plasmids were then transformed into *E. coli* and plated on the

appropriate vector antibiotic. PCR was then used to determine the percentage of clones with inserts for the 3 and 8kb libraries before sequencing occurred. For 40kb libraries, the DNA was randomly sheared using a hydroshear, end-repaired for blunt-end cloning, size selected on a pulse field agarose gel, extracted and purified. The 40kb insert was ligated into pCC1FOS, packaged and infected by phage into *E. coli.* The infection was plated on the appropriate antibiotic and titered. Ten 384 well plates for each library were initially sequenced and the quality of each of the libraries was assessed.

A breakdown of the total amount of sequence data obtained for each sample is as follows:

1) US sample

- 98,147 3 KB reads, containing 90.8 MB of raw sequence.
- 46,843 8 KB reads, containing 47.6 MB of raw sequence.
- 10,752 35 KB reads, containing 11.1 MB of raw sequence.

The reads were screened for vector using cross_match, then trimmed for vector and quality[7]. Reads shorter than 100 bases after trimming were then excluded. This reduced the data set to:

- 91,596 3 KB reads, containing 60.5 MB of sequence.
- 42,922 8 KB reads, containing 32.4 MB of sequence.
- 9,071 35 KB reads, containing 5.6 MB of sequence.

Total: 98.5 Mbp

2) Oz sample

- 58,251 3 KB reads, containing 64.1 Mbp of raw sequence.
- 56,064 8 KB reads, containing 56.4 Mbp of raw sequence.
- 5,376 35 KB reads, containing 5.4 Mbp of raw sequence.

The reads were screened for vector using cross_match, then trimmed for vector and quality[7]. Reads shorter than 100 bases after trimming were then excluded. This reduced the data set to:

- 54,980 3 KB reads, containing 41.6 Mbp of sequence.
- 47,393 8 KB reads, containing 33.1 Mbp of sequence.
- 4,592 35 KB reads, containing 2.9 Mbp of sequence.

Total: 77.6 Mbp

Assembly

Both datasets were assembled using parallel phrap version SPS 4.18 (www.phrap.org) compiled for SUN version 4.18 with the following arguments: minmatch 30, maxmatch 55, minscore 55, vector bound 20, revise_greedy. The US sludge was assembled with the JGI JAZZ assembler[8] as previously described[9]. A reimplemented version of the JAZZ assembler was subsequently applied to both datasets:

Jazz Assembly Parameters: US sample

The data was assembled using release 2.9.3 of JAZZ, a WGS assembler developed at the JGI[8-10]. A word size of 13 was used to compare reads for alignment, with a minimum of 14 such words needed to seed an alignment. The unhashability threshold was set to 50, preventing words present in more than 50 copies in the data set from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. As the different organisms in the data set were expected to be present at different sequence depths, the usual depth-based bonus/penalty system was turned off.

Post-Assembly Analysis: US sample

The initial assembly contained 4,339 scaffolds, with 15.0 MB of sequence, of which 30.7% was gap. The scaffold N/L50 was 52/35.4 KB, while the contig N/L50 was 124/4.5 KB. Redundant scaffolds were identified by BLAT-aligning all scaffolds with less than 5 KB of contig sequence against those with more than 5 KB of contig sequence. Any scaffolds from the former set that matched any of the larger over more than 80% of their length were excluded. Short scaffolds (< 1 KB of contig sequence) were also excluded. This filtering left 1,511 scaffolds, with 12.3 MB of sequence. The scaffold N/L50 was 14/42.2 KB, while the contig N/L50 was 48/40.8 KB. This filtered scaffold set served as the starting point of all downstream analysis.

Jazz Assembly Parameters: OZ sample

The data was assembled using release 2.9.3 of JAZZ, a WGS assembler developed at the JGI[8-10]. A word size of 13 was used to compare reads for alignment, with a minimum of 14 such words needed to seed an alignment. The unhashability threshold was set to 50, preventing words present in more than 50 copies in the data set from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together

sequences that are more than about 97% identical. As the different organisms in the data set were expected to be present at different sequence depths, the usual depth-based bonus/penalty system was turned off.

Post-Assembly Analysis: OZ sample

The initial assembly contained 4,097 scaffolds, with 18.0 MB of sequence, of which 29.5% was gap. The scaffold N/L50 was 39/43.8 KB, while the contig N/L50 was 536/4.5 KB. Redundant scaffolds were identified by BLAT-aligning all scaffolds with less than 5 KB of contig sequence against those with more than 5 KB of contig sequence. Any scaffolds from the former set that matched any of the larger over more than 80% of their length were excluded. Short scaffolds (< 1 KB of contig sequence) were also excluded. This filtering left 1,937 scaffolds, with 16.1 MB of sequence, of which 33.0% was gap. The scaffold N/L50 was 23/80.4 KB, while the contig N/L50 was 362/6.8 KB. This filtered scaffold set served as the starting point of all downstream analysis.

We tested the possibility of coassembly of different species using simulated metagenomic datasets produced from isolate genomes (unpublished data) and found that only closely related strains (>96% nucleotide identity) of the same species could be co-assembled.

A. phosphatis genome size and completeness estimates

We used two independent methods to estimate the genome size of the dominant US *A. phosphatis* strain, and an inventory of conserved gene sets to estimate completeness. All methods were used on the second JAZZ assembly (version 2.9.3).

To start with, we estimated the genome size using the Lander-Watermann equation. Since this is a metagenomic dataset, all variables in the Lander-Waterman equation pertain to the dominant *A. phosphatis* strain and not the whole community. To achieve this, we based the calculation on scaffolds in the US JAZZ assembly binned as *A. phosphatis* with high confidence using phylopythia (see *Binning*).

As is well known[11], for a shotgun assembly the probability of a genome nucleotide being covered by *n* reads is a poisson distribution:

$$P(c,n) = c^n e^{-c} / n!$$

where c is the coverage $c=LN/G$, with L being the read length, N the total number of

reads and *G* the genome length. A histogram of the coverage for each read was fit to a poisson distribution in order to find *c* (Supplementary Fig. 4). The best fit was obtained for *c=7.64*. The genome size is, hence, *G=LN/c*.

Since we know the total number of reads that were assembled into *A. phosphatis* scaffolds and assuming a random distribution of reads we can calculate what is the fraction of unassembled *A. phosphatis* reads (1 read contigs[11]) $f_u = exp(-2c)$, where is the one minus the detectable overlap divided by the read length[11]. We used this estimation because the binning of individual reads is unreliable. Using c=7.64 and $= 1-100/704$, $f_u = 2.03\ 10^{-6}$; and the total number of reads is $N = N_a + Nf_u$, where N_a is the number of assembled reads. The number of assembled reads is known: $N_a = 55{,}904$ and therefore: $N = N_a/(1-f_u) = 55{,}904.11$. The number of unassembled reads is negligible ($Nf_u = 0.11$) since the Lander-Waterman equation assumes a purely random read distribution and for a coverage of $c = 7.64$, no unassembled reads are expected. The fact that the genome is not complete attests to the fact that coverage is not random and there are areas of no or low coverage due to several factors, such as insert toxicity to *E. coli* host cells. The estimation of the genome size through this method is $G = c/NL = 5.151\ Mb$ and is likely to be an underestimate due to artificially under-represented areas.

A better estimate may be offered by summing the lengths of the high confidence *A. phosphatis* scaffolds. The assembly of these scaffolds used read pair information to estimate gap sizes and does not assume a purely random distribution : $G = 5.651\ Mb$. A correction to this estimate and an estimation of its accuracy can be obtained as follows. The binning method outlined below, gives a probability P_i (p-value, McHardy *et al*., in preparation) that a scaffold *i* belongs to *A. phosphatis*. The total genome length is, thus:

$$G = \sum x_i$$

where $x_i = l_i$ with probability P_i and $x_i = 0$ with probability $1-P_i$, where l_i is the length of scaffold *i* and the sum is over all the *A. phosphatis* scaffolds (i.e. x_i only contributes to the genome length if it belongs to *A. phosphatis*). The estimated value of the genome length is then:

$$\langle L \rangle = \sum \langle x_i \rangle = \sum l_i P_i = 5.580 Mb$$

The accuracy can calculated as the variance:

$$var(L) = \left\langle \left(\sum x_i - \langle L \rangle \right)^2 \right\rangle = \left\langle \left(\sum x_i \right)^2 \right\rangle - 2\langle L \rangle \langle \sum x_i \rangle + \langle L \rangle^2 = \left\langle \left(\sum x_i \right)^2 \right\rangle - \langle L \rangle^2$$
$$= \left\langle \left(\sum x_i \right)^2 \right\rangle + \sum_{i \neq j} x_i x_j - \langle L \rangle^2 = \sum l_i^2 P_i + \sum_{i \neq j} P_i l_i P_j l_j - \langle L \rangle^2 = 205kb$$

to yield the estimate of genome size:

$G = 5.580 \pm 0.205 Mb$

Using this size estimate, we can calculate genome completeness as:

(sum of contiguous *A. phosphatis* sequence / genome size estimate) x 100%

5.300 / (5.580±0.205) = 91.6 - 98.6%

The completeness of the inferred *A. phosphatis* genome also was estimated using presence or absence of 182 core genes (Supplementary Table 2). Absence of 4 genes (2 tRNA synthetases, homoserine kinase and panthothenate kinase) were dismissed on evolutionary grounds since the closest sequenced phylogenetic neighbors of *A. phosphatis, Dechloromonas aromatica* and *Azoarcus* sp. EbN1, were also deficient in these genes. In these cases, it is expected that either a non-orthologous gene or a different instance of a pathway substitutes for the function. Of the remaining 178 genes, 5 ribosomal proteins and ketopantoate reductase were not found (97.2% of core genes present).

Gene calling

All sludge assemblies were annotated using the *ab initio* gene calling program, fgenesb (www.softberry.com). Normally *ab initio* gene calling of isolate genomes trains on the dataset being annotated, however, since metagenomes are multi-genomic datasets, self training generates low quality results (data not shown). Instead parameters were obtained from training on multiple bacterial isolate genomes to provide an “average” bacterial coding preference and other sequence features such as Shine-Dalgarno sequences. The command string used was bactg_ann.pl mixr_paths_newcog.list1 <sequence_file> 60, where the sequence file is the fasta output of the assembled contigs (e.g. in Phrap this is the fasta.screen.contigs file), and 60 is the minimal length of predicted ORFs in bp, and mixr_paths_newcog.list1 is a config file that contains information about used programs, databases, etc. This file contains reference to gener.par which provides generalized 'bacterial' gene parameters.

Binning

Phylogenetic clades (available in IMG/M, www.jgi.doe.gov/m) were assigned based on clade-characteristic features in intrinsic sequence composition for the ranks of the domain, phylum, class and order (McHardy *et al*, in preparation). The method uses SVM-based multi-class classifiers to assign a sequence to one of the known phylogenetic clades, or classify it as "origin unknown", in case there is too little evidence for assignment to one of these classes (which might be the case for very short sequences or sequences from poorly explored clades that are not part of the model). The applied models include the three domains, 14 phyla, 22 classes, and 30 clades at the order level. The order-level model includes a clade for the *Rhodocyclales* that was constructed from the genomic sequences of *Dechloromonas aromatica* and *A. phosphatis* contigs identified by phylogenetic marker genes and read coverage. *A. phosphatis* genomic fragments were also identified by "overlap binning", i.e since the only species-level overlap between the two sludges was *A. phosphatis*, contigs from the US and OZ assemblies with ≥95% nucleotide identity over at least 1 kb were assumed to be *A. phosphatis*.

In addition, sample-specific *A. phosphatis*-identifying classifiers were constructed from the known *A. phosphatis* sequence (Phrap contigs with GC between 0.6 and 0.65 and coverage >8X for US and >7X for OZ), and applied for the identification of additional fragments in both sequence collections.

These methods were complemented via a gene similarity method working as follows: for each gene in each scaffold the top 10 highest BLAST score matches in any of the 337 genomes available in the internal version of the IMG database[12] were retrieved. Taxa for each of these 10 hits were given a score linearly dependent on the ranking (100 for best hit, 90 for second best hit, etc) and multiplied by the identity percentage. The score for each taxon was summed for each of the genes in the scaffold. Typically one or two taxa obtained much greater scores than the rest, providing the closest sequenced organisms. For the case of *A. phosphatis* highest scoring taxa were *Dechloromonas aromatica RCB* and *Azoarcus sp. (strain EbN1)*. Since accuracy for this method relied on having large contigs/scaffolds only the JAZZ assembly was used. Additionally, these assignments were confirmed using the sequence overlap between both samples: since the

phylogenetic commonality between the communities was minimal apart from *A. phosphatis*, scaffolds with a majority of their sequence having >95% identity with any of the contigs of the OZ assembly (Supplementary Fig 1) were classified as *A. phosphatis*. Appendix 1 contains the scaffolds and contigs for the *A. phosphatis* binning.

Metabolic reconstruction

The annotated sequences were loaded into IMG/M (www.jgi.doe.gov/m), a data management and analysis platform for genomic and metagenomic data based on IMG[12]. IMG/M provides pre-computed sequence similarity relationships, functional annotations and pathway information, reducing the time required for a metabolic reconstruction. KEGG pathways present in the metagenome were automatically assigned based on EC numbers in the annotation and pathways not included in the KEGG database were inferred from comparative analysis with other genomes, based on sequence similarity and gene cluster structure conservation. The resulting metabolic model is summarized in figure 2. An exact reproduction of this figure with gene object identifiers (oids) for tracking in IMG/M is shown in Supplementary Figure 3. The metabolic reconstruction was largely based on the JAZZ assembly, for which the binning is most reliable.

Phylogenetic inference

Partial and complete 16S rRNA gene sequences identified in the metagenomic datasets were aligned using greengenes[13] and imported into an ARB database[14] for comparative analysis with reference sequences. 99 full length reference sequences were used to construct a maximum likelihood tree (axML, Lanemask filter) and partial sequences < 1300 bp were added to the tree using the parsimony insertion tool without allowing changes in the overall tree topology. For clarity, some reference sequences were trimmed from the topology to produce Fig. 1.

Gene-centric comparative analysis of EBPR sludges to other metagenomes

A gene-centric analysis similar to that described by Tringe *et al*. [15] was used to determine relative representation of gene families between metagenomic datasets. We used the following data sources: US sludge (both Phrap and JAZZ assemblies), OZ sludge, acid

mine drainage biofilm[9], soil and whalefall[15]. All samples were sequenced at the JGI, and thus had consistent methods of sample preparation, sequencing methods and assembly. All assemblies were derived using Phrap, with the exception of the US JAZZ assembly and the AMD data set. The same *ab initio* ORF caller (fgenesb) was applied to all assemblies, thus data sets had reasonable consistency facilitating comparisons. Unassembled reads from the sludge and AMD datasets were not included in the analysis.

To derive gene families, we used protein translations and excluded short genes (less than 60aa), pre-masked both the query and the database using the CAST algorithm[16] and applied blastp[17] with no default filtering, an e-value threshold of e-10 and effective database length set to 5 million. As a measure of sequence similarity we used 'conservation score' which is derived as a pairwise comparison score divided by the smallest self-score of the two proteins[18]. The resulting similarity matrix was clustered using MCL[19], with default parameters and an inflation value of 2.0. The contribution of each metagenome to each of the clusters was normalized by the total number of genes in each project. The resulting protein families were annotated using a total consensus annotation program[20]. Only protein families with at least 10 members were considered further.

To sort protein families according to their representation in metagenomic data sets, we represented each environmental project as a vector and computed euclidian distances between these vectors. In order to meaningfully sort the data, we constructed neighbor-joining trees with quicktree[21], and performed leaf sorting with independent implementation of the optimal leaf ordering algorithm[22]. Since the leaf ordering algorithm is limited to an input size of several hundred leaves, we used MCL clustering to break large initial data clusters to a number of smaller clusters, and applied the combination of tree construction and leaf ordering on each of the resulting data subsets. One such MCL cluster clearly contained families over-represented in the sludge, with other communities either missing completely or having a minor representation in the cluster.

Supplementary details of the *A. phosphatis* metabolic reconstruction

Acetate/propionate uptake and activation

A. phosphatis appears to take acetate up through the expression of a gene cluster including yjcG (*actP*) and the acetyl-CoA synthase gene (*acs*). A very similar cluster has been shown to function as an operon in *E. coli* and be involved in acetate transport[23]. The *acs* gene is involved in irreversible high affinity acetyl-CoA synthesis and, co-regulated with *actP*, is part of a transport system used in low acetate concentrations[24]. *A. phosphatis* possesses another copy of *actP*, lacking any nearby *acs* genes, which seems to be part of a low affinity acetate consumption system in conjunction with genes coding for acetate kinase (*ackA*) and phosphotransacetylase (*pta*) (also present separately). These genes encode a low affinity pathway to acetyl-CoA generation from acetate and are found clustered in this genome.

We expect that propionate is taken up through the same transporter as acetate and activated via a propionyl-coA synthase (prpE in Figure 2).

Polyphosphate metabolism

A variety of genes for basic polyphosphate manipulations and storage are present in the *A. phosphatis* genome, as commonly observed in many other organisms such as *E. coli* and *Pseudomonas* sp. Polyphosphate AMP phosphotransferase (*pap*) degrades polyphosphate by cleaving one of the phosphates in the chain away from the rest (see left bottom cycle in Figure 2A). In doing so, it uses AMP and produces ADP. This resulting ADP yields AMP and ATP in a reaction catalyzed by adenylate kinase (*adk*). ATP is used for energetic purposes while the resultant AMP can be reincorporated in the cycle to degrade more polyphosphate. It is interesting to note that adenylate kinase and the membrane bound pyrophosphatase genes occur next to each other on the genome suggesting that ATP production via *pap* and *adk* is linked to maintenance of the proton motive force.

The polyphosphate degradation cycle is reversible and ATP and AMP can be used to add phosphate groups to the polyphosphate chains. Genes are also present that facilitate non-reversible processes to break polyphosphates through exopolyphosphatase

(*ppx*) and polyphosphate kinase 2 (*ppk2*), the latter producing GTP from GDP.

The organization of the *ppx-ppk1-relA* gene cluster is particularly intriguing. A similar cluster with the same gene orientation and order was recovered from EBPR sludge cultivated in Berkeley, CA[25], suggesting this gene order is conserved across other geographically isolated *Accumulibacter* strains. While *ppx* and *ppk1* are often found next to each other in bacterial genomes, the unique cluster of *ppx-ppk-relA* has been found in only one other sequenced genome to date (*Azoarcus* sp. EbN1). The synthetase encoded by RelA produces the intracellular "alarmone" guanosine 3',5'-bisdiphosphate (ppGpp) which initiates global changes in RNA expression via the stringent response[26]. *E. coli* mutants lacking *relA* could not accumulate polyphosphate under certain conditions[27] and *E. coli* engineered to produce large quantities of (p)ppGpp produced massive amounts of polyphosphate[28]. It is thought that (p)ppGpp inhibits polyphosphatase (PPX) by binding to the enzyme, preventing polyphosphate breakdown. Additional links between polyphosphate metabolism, amino acid starvation, and ribosomal protein degradation have also been reported[29]. The proximity of *relA* to *ppk* and *ppx* in the *A. phosphatis* genome suggests a similar role for a "magic spot" in EBPR metabolism, although the exact mechanism probably involves a complex regulatory network also involving RpoS, as proposed by Kornberg *et al.* [30].

Glycogen degradation

Glycogen degradation to pyruvate can be carried out via the EM or ED pathway. Which one is actually used was contentious[3], with proponents for both ED[31] and EM[32]. The controversy is significant since it has a substantial impact on the cellular energy budget, with the EM pathway yielding more ATP[24].

All genes for the EM pathway are present in the *A. phosphatis* genome. In contrast, the key genes for ED (only present in this pathway), encoding 6-phosphogluconate dehydratase and 2-keto-3-deoxy-6-phosphogluconate aldolase[33] are notably absent. These genes are also absent in the two closest sequenced relatives of *A. phosphatis*, *Dechloromonas aromatica* and *Azoarcus* sp. EbN1. Furthermore, other enzymes typically feeding into the ED pathway are not present in *A. phosphatis* (e.g. glucokinase and gluco-6-phosphate dehydrogenase). It is possible that the ED pathway

was missed in *A. phosphatis* because the genome is incomplete, however, the combination of high recovery of essential gene sets (99.7%, Supplementary Table 2) and absense of ED genes in related bacteria suggests that the pathway is indeed absent. We therefore conclude that the EM pathway is used instead of the ED pathway.

Empirical evidence for the functioning of EM versus ED remains inconclusive. Maurer et al.[34] and Hesselmann et al.[24] determined through NMR studies with ^{13}C-labeled acetate that the ED pathway is dominant in some sludges. Pereira et al.[31] also used NMR, but could not distinguish between the two pathways. Thus, part of the EBPR research community still uses the EM pathway while another uses the ED pathway for stochiometric calculations. Both choices yield models fitting some stochiometic ratios properly and others less optimally, depending on the system (a summary can be found in Schuler et al. [3]). As pointed out by Seviour et al.[35], the main weak point of most EBPR studies is that the structure and functional relationships of the populations involved are mainly unknown. For example, in the case of Maurer et al.[34] the sludge was obtained from a pilot scale treatment plant fed municipal wastewater (as opposed to acetate) and was subsequently exposed to only two EBPR cycles. Therefore it is not clear if the population was comprised mainly *A. phosphatis*, or even polyphosphate accumulating organisms (PAOs) in general. However, the lab-scale sludge studied by Hesselmann *et al.*[24], implicating ED as the dominant pathway, is likely to have been dominated by *Accumulibacter* since the authors identified this organism in their sludge using molecular methods[36]. This apparent paradox will require further experiments to be resolved.

The subsequent reduction of pyruvate to acetyl-CoA provides additional reducing power through the pyruvate dehydrogenase complex. Of the three genes encoding this complex, two of them (encoding pyruvate dehydrogenase and dihydrolipoyl transacetylase) are adjacent and found in the same gene neighborhoods as oserved in e.g. *E. coli* and *Buchnera*, with US JAZZ scaffold 3 ending where the third one (dihydrolipoyl dehydrogenase) should be. The dihydrolipoyl dehydrogenase gene can be found by itself in the beginning of a smaller scaffold suggesting these scaffolds are linked.

PHA synthesis and degradation

Genes for biosynthesis of poly(3-hydroxybutyrate) (PHB – from acetate only), poly(3-hydroxyvalerate) (PHV from acetate and propionate), poly-beta-hydroxy-2-methylbutyrate (PH2MB from acetate and propionate) and poly-beta-hydroxy-2-methylvalerate (PH2MV – from propionate only) are found in a cluster formed by *phaA* (beta-ketothiolase) and *phaC* (PHA synthase), and a genomically remote copy of *phaB* (acetoacetyl-CoA reductase). Although the usual configuration clusters all these genes next to each other, there are several known cases in which they are found divided in two clusters[37]. In particular, the same *phaA* and *phaC* group structure is found in *C. acidovorans*[38]. The PHA synthase unit is homologous to poly-beta-hydroxybutyrate synthase, a type I synthase. Since PHA synthases are very versatile and not specific to only one type of hydroxyalkanoic acid[38], we expect the same gene to be used for PHB, PHV, PH2MB and PH2MV synthesis.

The depolymerase gene phaZ is found in the vicinity of the *phaA* and *phaC* cluster. An additional pathway for PHB depolymerization is present in the form of the glyoxylate reduction cycle[39]. Although not all genes in this pathway are characterized in the literature (hence the dashed line in Fig. 2B), the signature genes (*croR*, *ibd2*, *meaC*, *meaA*, *ccR*) are found in a single cluster, suggesting that the whole pathway is present.

Phosphate transporters

The continuous shuttling of Pi through the membrane (in for anaerobic and out for aerobic phases) that *A. phosphatis* performs during the EBPR cycle requires that it is well equipped with phosphate transporters.

Both low and high affinity phosphate transport systems (Pit and Pst respectively[40]) are present. The low affinity, more difficult to saturate, transport system is assumed to be more readily used in the phosphate abundant bioreactor environment. This low affinity system is encoded in a cluster containing two phosphate permeases alternated with two transport regulators. Although this suggests a recent in-site duplication of a single permease/regulator pair, a phylogenetic tree of these genes and its *Dechloromonas* orthologs rules this possibility out: the Pit genes have higher similarities with other organisms than to each other. This tandem set of Pit genes is very uncommon

in other organisms and is likely to be a recent adaptation to the EBPR lifestyle, i.e. increasing the organisms ability to transport Pi across the cell membrane. The proton motive force needed to energize Pit[40] is provided under aerobic conditions by the respiratory chain. In the absence of oxygen, the proton gradient is likely to be sustained in part by a proton transporting pyrophosphatase with pyrophosphate being provided by the high affinity acetyl-CoA conversion pathway.

The high affinity Pi transport system is composed of three highly conserved clusters, each composed of one ATPase, one to two permeases and one periplasmic component system of the ABC phosphate transport system. Such a high number of Pst transporters has only been found in three organisms sequenced to date (*Anabeana variabilis* ATCC29413, *Nostoc punctiforme* PCC73102 and *Symbiobacterium thermophilum* IAM14863), and is surprising considering that *A. phosphatis* is unlikely to be able to use these genes for most of the EBPR cycle (see below).

One of the Pst clusters seems to be regulated by a nearby two component system formed by a histidine kinase (*phoR*) and a CheY-like response regulator (*phoB*). A promoter similar to the pho box encountered in *E. coli*[41] is found upstream of the periplasmic component of another of the Pst clusters. This promoter is located 68 bp upstream of one of the periplasmic components of the ABC component. It is composed of the motif TGTCA repeated twice and separated by 6 bps of equal GC/AT content: TCAAGC, which is very similar to the pho box reported in *E. coli*[41]. This system is known to strongly regulate the expression of Pst, derepressing it only in cases of low Pi concentrations[40,42]. Therefore, it is likely that the Pst system is repressed for most of the EBPR cycle, with the possible exception of the end of the aerobic period when Pi concentrations drop to uM levels.

Exopolysaccharide formation

Generally, extracellular polymeric substances (EPS) production is known to enhance survival of bacteria under conditions of environmental stress (oxidative, osmotic, acid and even temperature). Furthermore, the EPS layer can create a low-O_2 environment, which could be necessary to maintain the activity of oxygen-sensitive enzymes, such as Rubisco and nitrogenase. However, in the EBPR system, a critical role of EPS is to bind

the *A. phosphatis* cells together in dense clusters necessary for settling. Since the clusters appear to be comprised exclusively of *A. phosphatis* cells, and probably originated via replication from single cells, the EPS encoded by this organism is likely to be the principle glue holding the cells together. Flanking EBPR populations also have EPS genes and this may serve to hold together larger multispecies aggregates.

A. phosphatis has two chromosomal clusters coding for biosynthesis of exopolysaccharide, which is rather unusual, but not unique: *Sinorhizobium meliloti*, for instance, makes two acidic exopolysaccharides, succinoglucan and galactanoglucan. The reason for the presence of two exopolysaccharide biosynthesis clusters in *A. phosphatis* is not clear. If EPS1 and EPS2 have different physical properties, such as net electrostatic charge, they would have different affinity for phosphate and metal ions.

Both EPS biosynthesis clusters in *A. phosphatis* belong to the Wzy-dependent type. This type biosynthesizes the repeat unit inside the cell. The oligosaccharide is then exported by a flippase (Wzx family) and finally EPS is polymerized outside the cell by a Wzy-family polymerase. Both clusters in *A. phosphatis* encode some enzymes for biosynthesis of nucleotide-sugar precursors, several glycosyltransferase enzymes, and regulators of EPS biosynthesis. Both clusters also include several membrane proteins with multiple transmembrane helices; apparently, they code for a flippase and EPS polymerase, but it is impossible to determine their exact functions, because these proteins are highly specific for each particular type of EPS and very poorly conserved.

Cluster 1 might code for biosynthesis of a “group-specific” EPS, since several genes in this cluster have orthologs in *Dechloromonas*; however, the structure of the EPS produced by *A. phosphatis* and *Dechloromonas* is different, because the EPS cluster in *Dechloromonas* does not include GDP-mannose dehydrogenase or a fusion protein polysaccharide deacetylase/formyltransferase. Cluster 2 is located next to Rubisco; the genes in this cluster have no orthologs in *Dechloromonas*. These two clusters most likely encode biosynthesis of EPS with different physical and chemical properties: while both EPS contain mannuronate, only EPS2 also contains an aminosugar (indicated by the presence of aminotransferase and acetyltransferase similar to the enzymes for biosynthesis of viosamine), so EPS1 may be more acidic than EPS2. In addition, EPS1, but not EPS2, may be modified by an acyl residue due to the presence of a CoA ligase

family protein.

Nitrogen fixation

One of the most surprising discoveries made during the metabolic reconstruction of *A. phosphatis* is the presence of nitrogen fixation (*nif*) genes. Nitrogen fixation is a very energy expensive process generating ammonia as a final product[43]. Ammonia is present in high quantities in the wastewater environment, so there seems to be little incentive for the *A. phosphatis* to invest energy in fixing nitrogen.

The organization of *nif* genes resembles that found in *Dechloromonas aromatica* and *Azotobacter vinelandii,* namely a *nifTKDH* cluster. Cofactors for the MoFe protein are encoded in a *nifENXQ* cluster, similar in structure to that found in *A. vinelandii.* A nearby cluster contains *nifU* and *nifS*, needed for nitrogenase maturation[44]. Gene *nifB* (FeMo cofactor biosynthesis) along with ferrodoxin and flavodoxin synthases are found in a cluster with identical gene order to *D. aromatica. NifJ,* encoding flavodoxin oxireductase and a cluster composed of *nifA* (transcription activator) and *nifL* (*nifA* regulator) are also present.

Other genes involved in the elaborate nitrogen fixation regulatory network are present including *nifZ,* suggested to encode a chaperon in the stepwise assembly of the nitrogenase MoFe protein[44], *nifW*, shown to interact with the MoFe protein[45], *nifR3*, involved in nitrogen regulation[46].

Carbon fixation

The signature enzyme genes for carbon fixation are present, including the large subunit of rubisco, phosphoribulokinase, fructose-bisphosphatase and sedoheptulose-bisphosphatase. Whereas the presence of the large unit of rubisco by itself does not necessarily mean that *A. phosphatis* fixes carbon, the concommitant presence of phosphoribulokinase indicates that this is the case. The absence of the small subunit of rubisco means that carbon fixation can only take place anaerobically.

A. phosphatis may be capable of CO_2 fixation and chemolithoautotrophic growth using hydrogen as a sole energy source, similar to *Ralstonia eutropha* or *Rhodobacter capsulatus*[47]. We identified three gene clusters coding for [NiFe]

hydrogenases, one of them similar to *Ralstonia eutropha* sensor hydrogenase[48] and two others similar to energy-generating isozymes of *Ralstonia*, membrane-bound, periplasm-facing hydrogenase and cytosolic soluble NAD-reducing hydrogenase. We also found the genes necessary for insertion of Ni cofactor and maturation of [NiFe] hydrogenases and a gene coding for HypX protein, which is necessary for protection of cytosolic hydrogenase against inactivation by oxygen.

Either oxygen or nitrite/nitrate may serve as electron acceptors and that the cytosolic soluble hydrogenase is most likely used aerobically, while the periplasmic oxygen-sensitive enzyme probably serves as a part of anaerobic respiratory chain. It is also possible that *A. phosphatis* is capable of using other energy sources and electron acceptors for chemoautotrophic growth, such as sulfur compounds and dimethyl sulfoxide or trimethylamine N-oxide, due to the presence of several cytochromes and a periplasmic molybdopterin-dependent dehydrogenase of unknown specificity.

However, unlike *Ralstonia* and *Rhodobacter* possessing form I RubisCO, *A. phosphatis* has a form II enzyme lacking the small subunit (87% identity to the CbbM protein of *Thiobacillus denitrificans*[49]). This form requires higher CO_2/O_2 ratios to function as an efficient carboxylase, so it is improbable that the *A. phosphatis* RubisCO can function aerobically. The most likely scenario is that CO_2 fixation happens in anoxic conditions, with hydrogen as a sole energy source and nitrate or nitrite as electron acceptor.

Note that the entire respiratory chain in this case is located in the periplasm and electron transfer is performed through a quinone pool and/or periplasmic pool of cytochromes and ferredoxins, with no connection to the cytosolic pool of NAD(P). However, CO_2 fixation requires NADPH, so a dedicated enzyme capable of using proton-motive force to catalyze the uphill electron transfer from the quinone pool to NAD(P)+, such as an unusual cytochrome b shown in Figure 3, would be beneficial. Upon "domestication" of *A. phosphatis*, this enzyme could be recruited in the anaerobic phase of EBPR to boost the supply of reducing equivalents for polyhydroxyalkanoate biosynthesis.

Appendix: A. phosphatis binnings for JAZZ and Phrap assemblies

US JAZZ scaffolds (scaffold names, not gene OIDs, for first JAZZ assembly) :

1, 2, 3, 4, 6, 7, 10, 15, 22, 26, 34, 35, 44, 48, 52, 53, 55, 64,
71, 84, 97, 115, 123, 127, 128, 130, 131, 152, 153, 177, 200, 206, 246, 256
265, 271, 281, 291, 295, 296, 332, 390, 428, 461, 466, 551, 557, 660, 664,
902, 907

US Phrap contigs (scaffold names, not gene OIDs):

16370,16369,16368,16367,16366,16365,16364,16363,16362,16361,16360,16359,16358
16357,16356,16355,16354,16353,16352,16351,16350,16349,16348,16347,16346,16345
16344,16343,16342,16341,16340,16339,16338,16337,16336,16335,16334,16333,16332
16331,16330,16329,16328,16327,16326,16325,16324,16323,16322,16321,16320,16319
16318,16317,16316,16315,16314,16313,16312,16311,16310,16309,16308,16307,16306
16305,16304,16303,16301,16300,16299,16297,16296,16295,16294,16293,16292,16291
16290,16288,16286,16285,16284,16283,16282,16281,16280,16279,16278,16277,16276
16274,16273,16272,16271,16270,16269,16268,16267,16265,16264,16263,16262,16260
16259,16258,16257,16255,16253,16252,16251,16250,16249,16248,16245,16244,16243
16242,16241,16240,16239,16238,16237,16236,16235,16232,16231,16230,16229,16228
16227,16225,16224,16222,16221,16218,16217,16216,16215,16212,16211,16210,16206
16197,16196,16194,16189,16187,16186,16180,16160,16159,16158,16157,16156,16143
16141,16130,16128,16118,16116,16095,16077,16075,16073,16069,16058,16055,16050
16043,16032,16018,16017,15991,15983,15973,15970,15964,15947,15931,15926,15921
15920,15907,15868,15858,15853,15852,15816,15815,15814,15804,15797,15793,15791
15786,15783,15782,15776,15771,15770,15758,15737,15735,15733,15722,15698,15690
15663,15659,15656,15643,15635,15611,15609,15589,15584,15571,15564,15551,15532
15531,15524,15518,15512,15504,15503,15476,15449,15402,15338,15319,15317,15297
15287,15280,15278,15275,15259,15250,15247,15233,15214,15190,15163,15148,15136
15108,15040,15016,15006,14901,14740,14694,14584,14364,14073,13457,13316,13212
13138,12819

OZ Phrap contigs (scaffold names, not OID):

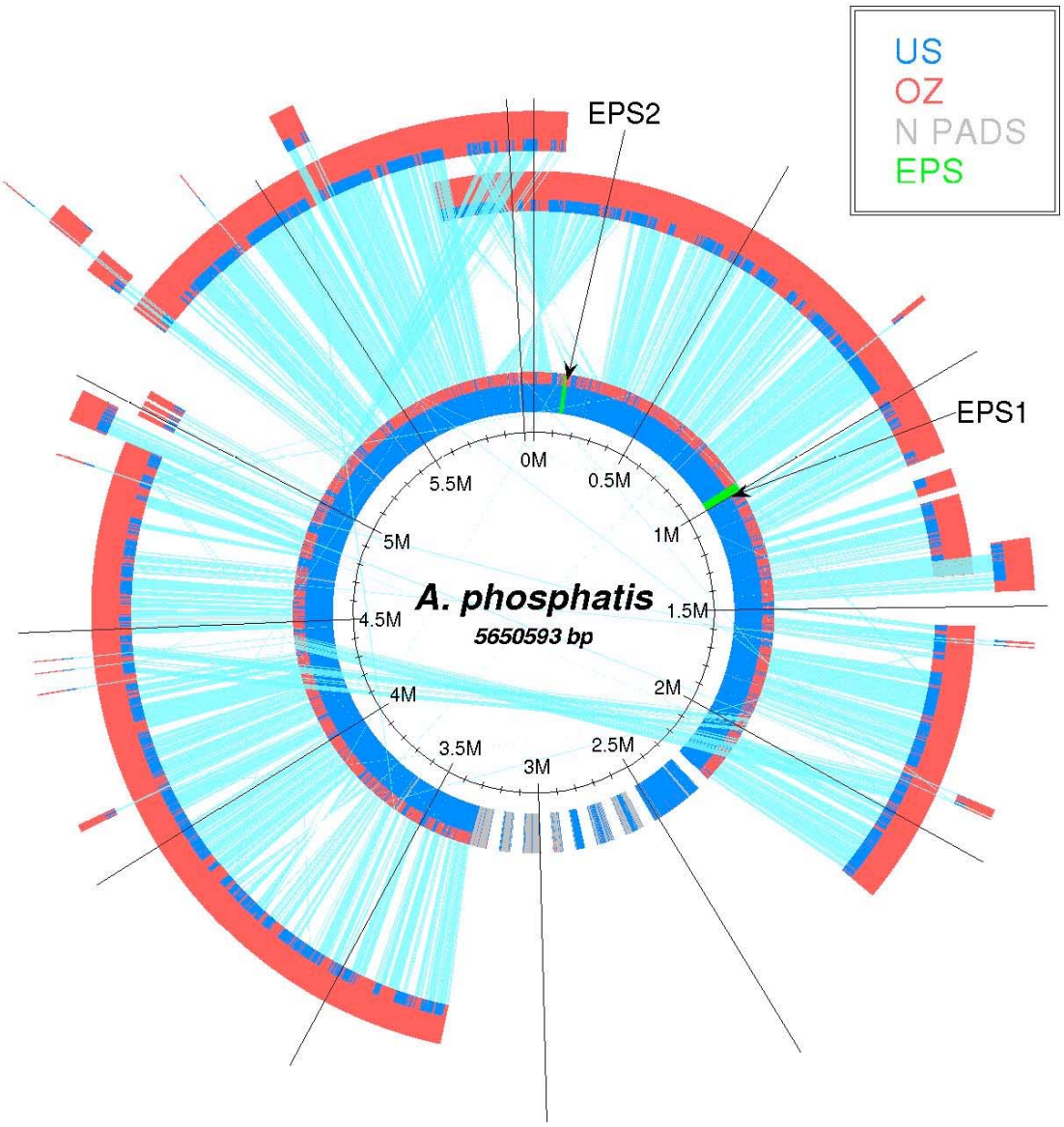
11632,11631,11630,11629,11627,11625,11624,11621,11620,11618,11615,11614,11612
11610,11607,11604,11603,11599,11596,11595,11588,11587,11586,11582,11581,11580
11579,11578,11576,11575,11572,11571,11570,11569,11568,11567,11566,11565,11563
11561,11559,11558,11557,11553,11552,11551,11550,11548,11547,11545,11543,11542
11541,11539,11538,11537,11535,11533,11532,11526,11523,11515,11514,11511,11510
11509,11508,11506,11504,11503,11502,11501,11500,11499,11498,11493,11487,11486
11485,11484,11483,11482,11481,11480,11478,11476,11475,11474,11473,11471,11469
11467,11460,11459,11458,11457,11455,11454,11453,11452,11448,11447,11446,11445
11443,11442,11441,11440,11437,11432,11430,11429,11424,11423,11421,11420,11418
11417,11416,11414,11413,11409,11401,11396,11394,11387,11386,11385,11382,11377
11376,11375,11373,11371,11370,11369,11368,11361,11360,11358,11357,11356,11355
11354,11353,11352,11350,11349,11347,11346,11344,11343,11342,11341,11339,11338
11332,11328,11324,11323,11317,11316,11315,11314,11311,11309,11305,11304,11300
11299,11283,11279,11278,11275,11272,11269,11267,11266,11264,11263,11262,11260
11257,11256,11255,11253,11250,11248,11247,11235,11231,11228,11225,11222,11221
11218,11217,11210,11209,11204,11202,11194,11188,11185,11180,11171,11160,11156
11152,11151,11150,11147,11140,11136,11127,11122,11120,11114,11112,11111,11106
11099,11098,11095,11093,11082,11080,11078,11077,11075,11051,11042,11041,11028
11027,11022,11021,11014,11011,11009,10999,10998,10997,10981,10969,10964,10963
10956,10954,10953,10947,10944,10941,10930,10919,10913,10904,10888,10876,10850
10816,10805,10774,10773,10771,10763,10760,10741,10732,10690,10666,10665,10658
10618,10606,10600,10506,10497,10464,10448,10424,10423,10420,10384,10380,10307
10275,10168,10156, 9509

Supporting References

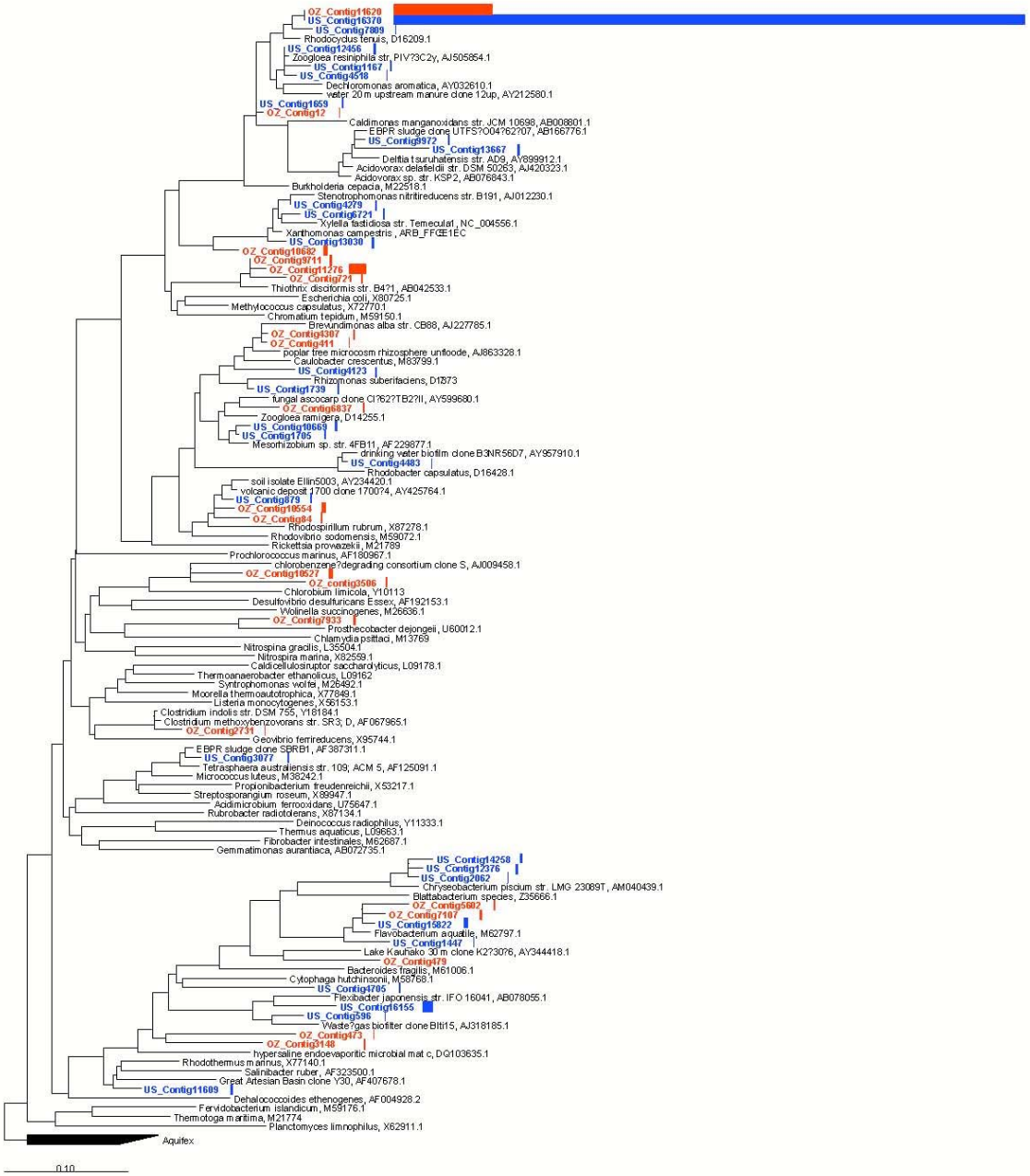
1. Zilles, J.L., Peccia, J., Kim, M.W., Hung, C.H. & Noguera, D.R. Involvement of Rhodocyclus-related organisms in phosphorus removal in full-scale wastewater treatment plants. *Appl Environ Microbiol* **68**, 2763-2769 (2002).
2. McMahon, K.D., Jenkins, D. & Keasling, J.D. Polyphosphate kinase genes from activated sludge carrying out enhanced biological phosphorus removal. *Water Sci Technol* **46**, 155-162 (2002).
3. Schuler, A.J. & Jenkins, D. Enhanced biological phosphorus removal from wastewater by biomass with different phosphorus contents, Part I: Experimental results and comparison with metabolic models. *Water Environ Res* **75**, 485-498 (2003).
4. Oehmen, A., Saunders, A.M., Vives, M.T., Yuan, Z. & Keller, J. Competition between polyphosphate and glycogen accumulating organisms in enhanced biological phosphorus removal systems with acetate and propionate as carbon sources. *J Biotechnol* (2005).
5. Purkhold, U. et al. Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S rRNA and amoA sequence analysis: implications for molecular diversity surveys. *Appl Environ Microbiol* **66**, 5368-5382 (2000).
6. Sambrook, J. & Russell, D.W. Molecular Cloning: A Laboratory Manual. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York; 2001).
7. Chapman, J., Putnam, N., Ho, I. & Rokhsar, D. "JAZZ, a whole genome shotgun assembler", unpublished.
8. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1283-1285 (2002).
9. Tyson, G.W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
10. Putnam, N. "Applications of Statistical Physics to Genome Assembly and Protein Folding", unpublished dissertation, University of California, Berkeley, 2004.
11. Lander, E.S. & Waterman, M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239 (1988).
12. Markowitz, V.M. et al. The Integrated Microbial Genomes (IMG) System. *Nucleic Acids Research* **34** (2006).
13. DeSantis, T.Z., Dubosarskiy, I., Murray, S.R. & Andersen, G.L. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* **19**, 1461-1468 (2003).
14. Ludwig, W. et al. ARB: a software environment for sequence data. *Nucleic Acids Res* **32**, 1363-1371 (2004).
15. Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
16. Promponas, V.J. et al. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* **16**, 915-922 (2000).
17. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
18. Lopez-Bigas, N. & Ouzounis, C.A. Genome-wide identification of genes likely to

be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108-3114 (2004).
19. van Dongen, S. PhD thesis (University of Utrecht, 2000).
20. Darzentas, N. Unpublished.
21. Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18**, 1546-1547 (2002).
22. Bar-Joseph, Z., Gifford, D.K. & Jaakkola, T.S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17 Suppl 1**, S22-29 (2001).
23. Gimenez, R., Nunez, M.F., Badia, J., Aguilar, J. & Baldoma, L. The gene yjcG, cotranscribed with the gene acs, encodes an acetate permease in Escherichia coli. *J Bacteriol* **185**, 6448-6455 (2003).
24. Hesselmann, R.P.X., Von Rummell, R., Resnick, S.M., Hany, R. & Zehnder, A.J.B. Anaerobic metabolism of bacteria performing enhanced biological phosphate removal. *Wat Res* **34**, 3487-3494 (2000).
25. McMahon, K.D. PhD thesis, Civil and Environmental Engineering, (University of California - Berkeley, Berkeley; 2002).
26. Cashel, M., Gentry, D.R., Hernandez, V.J. & Vinella, D. in Escherichia coli and Salmonella: cellular and molecular biology. (eds. F.C. Neidhardt et al.) (American Society for Microbiology, Washington, D.C.; 1996).
27. Ault-Riche, D., Fraley, C.D., Tzeng, C.M. & Kornberg, A. Novel assay reveals multiple pathways regulating stress-induced accumulations of inorganic polyphosphate in Escherichia coli. *J Bacteriol* **180**, 1841-1847 (1998).
28. Kuroda, A., Murphy, H., Cashel, M. & Kornberg, A. Guanosine tetra- and pentaphosphate promote accumulation of inorganic polyphosphate in Escherichia coli. *J Biol Chem* **272**, 21240-21243 (1997).
29. Kuroda, A. et al. Role of inorganic polyphosphate in promoting ribosomal protein degradation by the Lon protease in E. coli. *Science* **293**, 705-708 (2001).
30. Kornberg, A., Rao, N.N. & Ault-Riche, D. Inorganic polyphosphate: a molecule of many functions. *Annu Rev Biochem* **68**, 89-125 (1999).
31. Pereira, H. et al. Model for carbon metabolism in biological phosphorus removal processes based on in vivo13C-NMR labelling experiments. *Wat Res* **30**, 2128-2138 (1996).
32. Filipe, C.D., Daigger, G.T. & Grady, C.P., Jr. Stoichiometry and kinetics of acetate uptake under anaerobic conditions by an enriched culture of phosphorus-accumulating organisms at different pHs. *Biotechnol Bioeng* **76**, 32-43 (2001).
33. Gottschalk, G. Bacterial metabolism. (Springer-Verlag, New York; 1985).
34. Maurer, M., Gujer, W., Hany, R. & Bachmann, S. Intracellular carbon flow in phosphorus accumulating organisms from activated sludge systems. *Wat Res* **31**, 907-917 (1997).
35. Seviour, R.J., Mino, T. & Onuki, M. The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* **27**, 99-127 (2003).
36. Hesselmann, R.P., Werlen, C., Hahn, D., van der Meer, J.R. & Zehnder, A.J. Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced biological phosphate removal in activated sludge. *Syst Appl Microbiol* **22**, 454-465 (1999).
37. Rehm, B.H. & Steinbuchel, A. Biochemical and genetic analysis of PHA synthases and other proteins required for PHA synthesis. *Int J Biol Macromol* **25**,

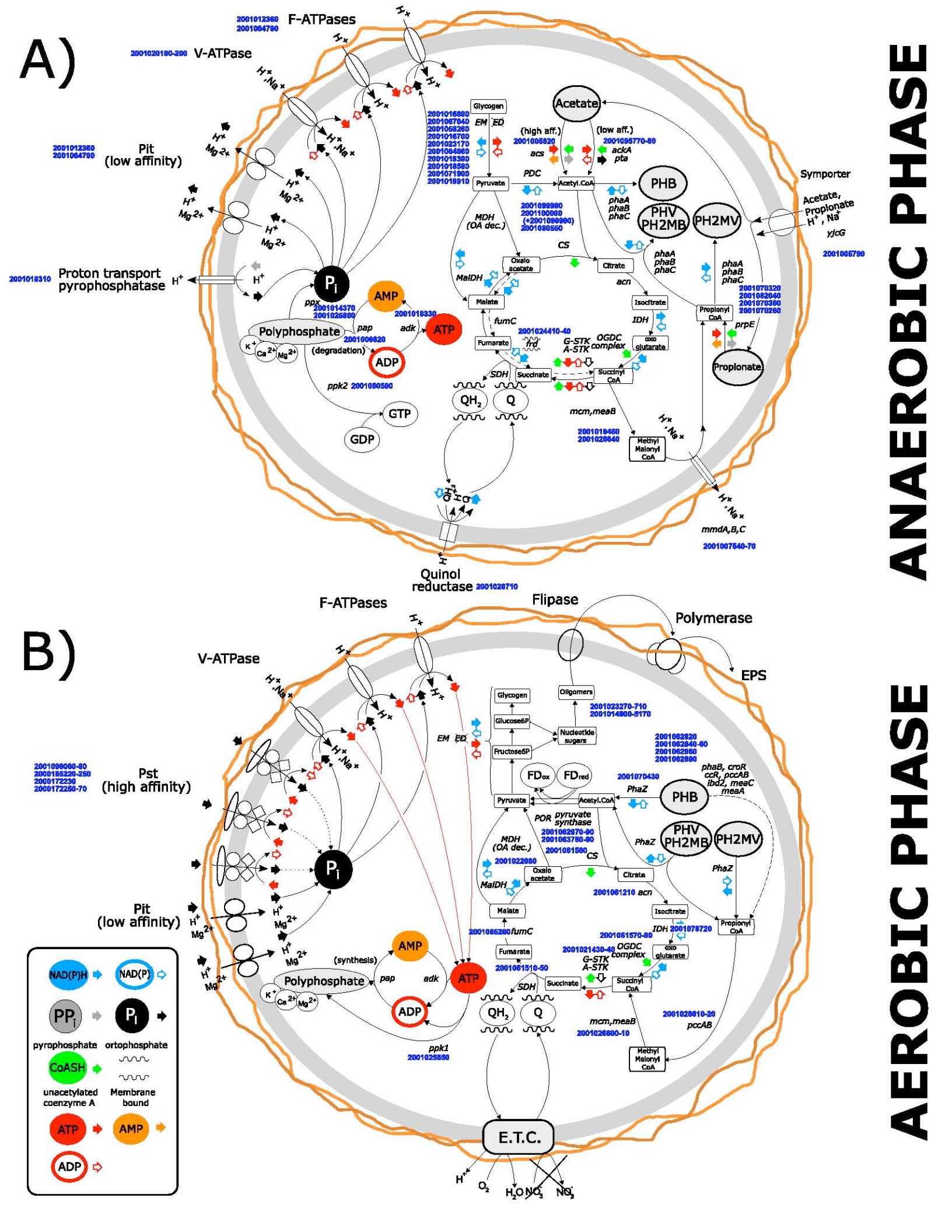
3-19 (1999).
38. Sudesh, K., Fukui, T. & Doi, Y. Genetic analysis of Comamonas acidovorans polyhydroxyalkanoate synthase and factors affecting the incorporation of 4-hydroxybutyrate monomer. *Appl Environ Microbiol* **64**, 3437-3443 (1998).
39. Korotkova, N., Lidstrom, M.E. & Chistoserdova, L. Identification of genes involved in the glyoxylate regeneration cycle in *Methylobacterium extorquens* AM1, including two new genes, meaC and meaD. *J Bacteriol* **187**, 1523-1526 (2005).
40. van Veen, H.W. Phosphate transport in prokaryotes: molecules, mediators and mechanisms. *Antonie van Leeuwenhoek* **72**, 299-315 (1997).
41. Ellison, D.W. & McCleary, W.R. The unphosphorylated receiver domain of PhoB silences the activity of its output domain. *J Bacteriol* **182**, 6592-6597 (2000).
42. Metcalf, W.W. & Wanner, B.L. Involvement of the Escherichia coli phn (psiD) gene cluster in assimilation of phosphorus in the form of phosphonates, phosphite, Pi esters, and Pi. *J Bacteriol* **173**, 587-600 (1991).
43. White, D. The Physiology and Biochemistry of Prokaryotes. (Oxford University Press, New York; 1995).
44. Hu, Y., Fay, A.W., Dos Santos, P.C., Naderi, F. & Ribbe, M.W. Characterization of Azotobacter vinelandii nifZ deletion strains. Indication of stepwise MoFe protein assembly. *J Biol Chem* **279**, 54963-54971 (2004).
45. Kim, S. & Burgess, B.K. Evidence for the direct interaction of the nifW gene product with the MoFe protein. *J Biol Chem* **271**, 9764-9770 (1996).
46. Foster-Hartnett, D., Cullen, P.J., Gabbert, K.K. & Kranz, R.G. Sequence, genetic, and lacZ fusion analyses of a nifR3-ntrB-ntrC operon in Rhodobacter capsulatus. *Mol Microbiol* **8**, 903-914 (1993).
47. Paoli, G.C. & Tabita, F.R. Aerobic chemolithoautotrophic growth and RubisCO function in Rhodobacter capsulatus and a spontaneous gain of function mutant of Rhodobacter sphaeroides. *Arch Microbiol* **170**, 8-17 (1998).
48. Kleihues, L., Lenz, O., Bernhard, M., Buhrke, T. & Friedrich, B. The H(2) sensor of Ralstonia eutropha is a member of the subclass of regulatory [NiFe] hydrogenases. *J Bacteriol* **182**, 2716-2724 (2000).
49. Hernandez, J.M., Baker, S.H., Lorbach, S.C., Shively, J.M. & Tabita, F.R. Deduced amino acid sequence, functional expression, and unique enzymatic properties of the form I and form II ribulose bisphosphate carboxylase/oxygenase from the chemoautotrophic bacterium Thiobacillus denitrificans. *J Bacteriol* **178**, 347-356 (1996).



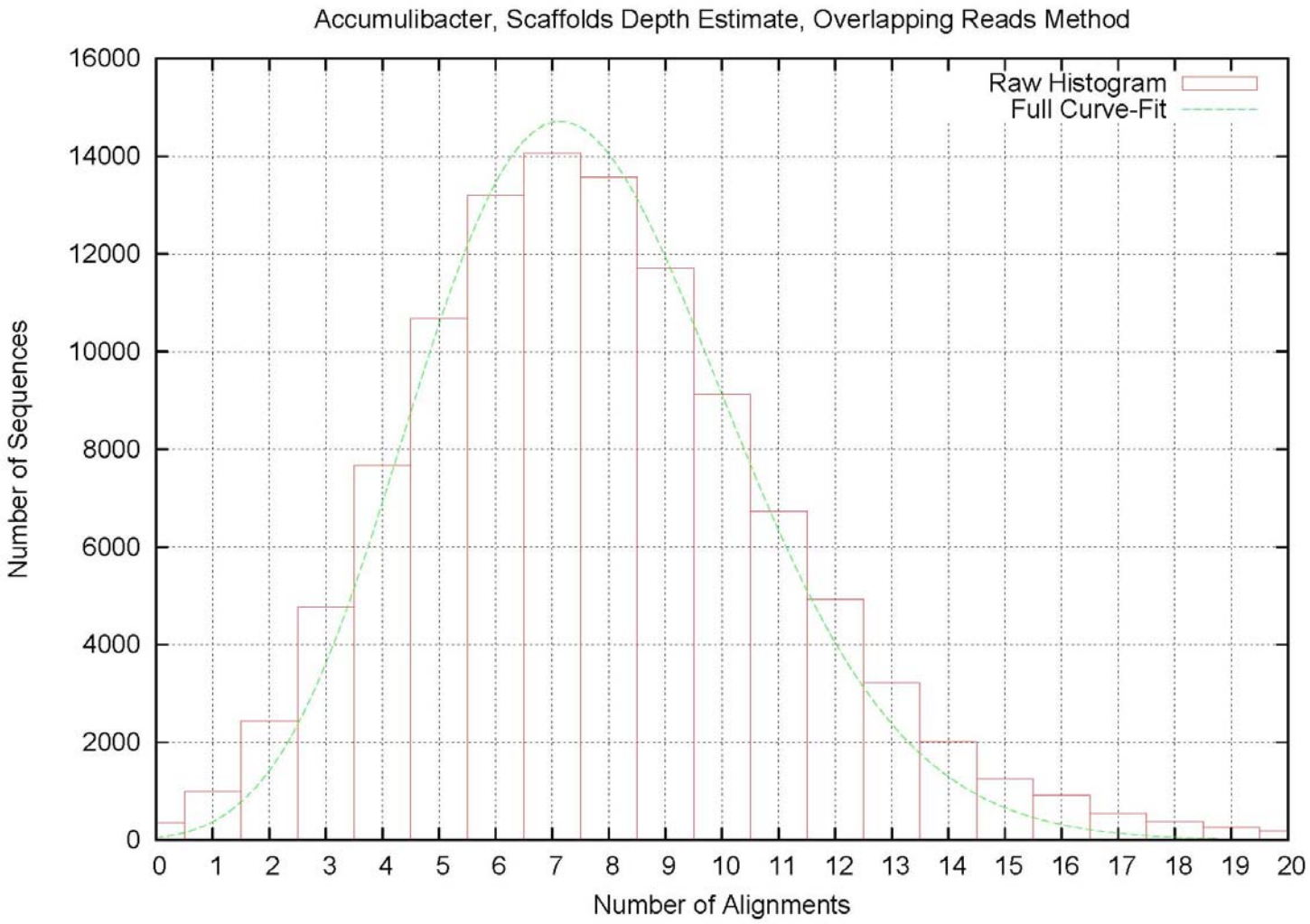
Supplementary Figure 1. Overview of the draft composite genomes of the *A. phosphatis* populations based on the second JAZZ assembly (version 2.9.3) of the US (inner blue ring) and OZ (outer red ring) metagenomic data. Regions of the OZ *A. phosphatis* genome that match the US genome at ≥95% nucleotide identity are shown in red against the blue ring and *vice versa*. The light blue lines connecting these regions indicate putative large scale rearrangements between the two genomes. The two EPS gene cassettes in the US genome are shown in green and N padding (gaps of known size in scaffolds) are shown in grey. The scaffolds shown have been binned as *A. phosphatis* with high confidence ($p>0.85$) by the *A. phosphatis*-specific SVM-based binning method. An estimation of genome length its reliability are given in the Supplementary material.



Supplementary Figure 2. Expanded maximum likelihood tree showing all metagenomic contigs on which 16S rRNA genes were identified, US sequences in blue, OZ sequences in red. Partial sequences <1300 bp were inserted into the tree according to maximum parsimony criteria. Gene object identifiers (gene oids) and contig ids are shown for all metagenomic sequences which can be tracked in IMG/M (www.jgi.doe.gov/m). Reference sequences shown in Figure 1 are bolded to provide complete information including accession numbers. Representatives of the phylum *Aquificae* were used as the outgroup for the analysis. Bars to the right of the dendrogram indicate relative size of contigs based on number of reads, and phylogenetic groups are also indicated on the right hand side.


A)
ANAEROBIC PHASE
F-ATPases
V-ATPase
Pit
(low affinity)
Proton transport
pyrophosphatase
Quinol
reductase
Glycogen
Acetate
(high aff.)
(low aff.)
acs
ackA
pta
PDC
Pyruvate
Acetyl.CoA
PHB
PHV
PH2MB
PH2MV
Symporter
Acetate,
Propionate
H+, Na+
yjcG
phaA
phaB
phaC
MDH
(OA dec.)
CS
Oxalo
acetate
Citrate
MalDH
acn
Malate
Isocitrate
fumC
IDH
Fumarate
frd
G-STK
A-STK
OGDC
complex
oxo
glutarate
Propionyl
CoA
prpE
Propionate
SDH
Succinate
Succinyl
CoA
QH2
Q
mcm,meaB
Methyl
Malonyl
CoA
mmdA,B,C
Pi
AMP
ATP
ADP
GTP
GDP
ppx
pap
adk
ppk2
Polyphosphate
(degradation)
K+
Ca2+
Mg2+
H+
Na+
B)
ANAEROBIC PHASE
F-ATPases
Flipase
Polymerase
EPS
V-ATPase
Pst
(high affinity)
Pit
(low affinity)
Glycogen
Glucose6P
Fructose6P
Nucleotide
sugars
Oligomers
EM
ED
FDox
FDred
Pyruvate
Acetyl.CoA
PHB
PHV
PH2MB
PH2MV
PhaZ
phaB, croR
ccR, pccAB
lbd2, meaC
measA
POR
pyruvate
synthase
MDH
(OA dec.)
CS
Oxalo
acetate
Citrate
MalDH
acn
Malate
Isocitrate
fumC
IDH
Propionyl
CoA
Fumarate
OGDC
complex
oxo
glutarate
G-STK
A-STK
Succinate
Succinyl
CoA
SDH
QH2
Q
mcm,meaB
pccAB
Methyl
Malonyl
CoA
Pi
AMP
ATP
ADP
(synthesis)
pap
adk
ppk1
Polyphosphate
K+
Ca2+
Mg2+
H+
Na+
E.T.C.
O2
H2O
NO3
NO2
NAD(P)H
NAD(P)
PPi
Pi
pyrophosphate
ortophosphate
CoASH
unacetylated
coenzyme A
Membrane
bound
ATP
AMP
ADP

Supplementary Figure 3. Figure 2 with gene oids to facilitate investigation of the genes and pathways in IMG/M. Dashes (-) indicate ranges of gene oids.



Supplementary Figure 4. Histogram of the number of reads (y axis) with a given number of alignments with other reads (x axis). For each read, an alignment is counted every time another aligned read crosses either one of the ends of the read. The fit to the Lander-Waterman equation yields an estimate for the coverage *c*.

Supplementary Table 1. SBR operational differences between the US & OZ EBPR sludges

Differences *	**US**	**OZ**
Feed concentrations:		
NH_4Cl	119	107
Yeast Extract	8.3	None
Casamino acids	30	None
Peptone	None	48
Added carbon source	Acetate	Propionate
In-reactor COD	115	215
COD/P	14	14
		EDTA in trace nutrients
		ATU to initially inhibit nitrification
Other:		
COD concentration at beginning of anaerobic phase	115	200
Feed pH adjustment	Na_2CO_3	NaOH
Sludge operating pH	7.0-7.3	7.5-8
Operating Volume	2 L	8 L
Sludge wastage frequency	1/day	4/day
TSS	~1,000	~3,400
HRT	12 h	24 h
SRT	4 days	8 days
Total biomass in system	2 g	27.2 g

*All values are listed as mg/L, unless otherwise specified

Acronyms

COD = chemical oxygen demand

VFA = volatile fatty acid

ATU = allyl-*N*thiourea

TSS = total suspended solids

EDTA = ethylenediaminetetraacetic acid

HRT = hydraulic residence time

SRT = solids retention time

Sample collection dates: US: July 3rd 2004, OZ: August 18th 2004

Supplementary Table 2. Estimate of completeness of the dominant US *A. phosphatis* genome based on presence/absence of core gene sets typically not co-localized in bacterial genomes.

COG family	conserved gene set	IMG gene object identifier
	Large subunit ribosomal proteins	
COG0080	Ribosomal protein L11	2001003040
COG0081	Ribosomal protein L1	2001003030
COG0087	Ribosomal protein L3	2001002910
COG0088	Ribosomal protein L4	2001002900
COG0089	Ribosomal protein L23	2001002890
COG0090	Ribosomal protein L2	2001002880
COG0091	Ribosomal protein L22	2001002860
COG0093	Ribosomal protein L14	2001002810
COG0094	Ribosomal protein L5	2001002790
COG0097	Ribosomal protein L6P/L9E	2001002760
COG0102	Ribosomal protein L13	absent
COG0197	Ribosomal protein L16/L10E	2001002840
COG0198	Ribosomal protein L24	2001002800
COG0200	Ribosomal protein L15	2001002720
COG0203	Ribosomal protein L17	2001002650
COG0211	Ribosomal protein L27	2001022580
COG0222	Ribosomal protein L7/L12	2001003010
COG0227	Ribosomal protein L28	2001003010
COG0230	Ribosomal protein L34	absent
COG0244	Ribosomal protein L10	2001003020
COG0254	Ribosomal protein L31	2001072010
COG0255	Ribosomal protein L29	2001002830
COG0256	Ribosomal protein L18	2001002750
COG0257	Ribosomal protein L36	2001002690/700*
COG0261	Ribosomal protein L21	2001022590
COG0267	Ribosomal protein L33	2001065740
COG0291	Ribosomal protein L35	2001027450
COG0292	Ribosomal protein L20	2001027460
COG0333	Ribosomal protein L32	absent
COG0335	Ribosomal protein L19	2001009740
COG0359	Ribosomal protein L9	2001059990
COG1825	Ribosomal protein L25 (general stress protein Ctc)	2001065360
COG1841	Ribosomal protein L30/L7E	2001002730
	small subunit ribosomal proteins	
COG0048	Ribosomal protein S12	absent
COG0049	Ribosomal protein S7	2001002950
COG0051	Ribosomal protein S10	2001002920**
COG0052	Ribosomal protein S2	2001066940
COG0092	Ribosomal protein S3	2001002850
COG0096	Ribosomal protein S8	2001002770
COG0098	Ribosomal protein S5	2001002740
COG0099	Ribosomal protein S13	2001002690
COG0100	Ribosomal protein S11	2001002680

COG0103	Ribosomal protein S9	absent
COG0184	Ribosomal protein S15P/S13E	2001072180
COG0185	Ribosomal protein S19	2001002870
COG0186	Ribosomal protein S17	2001002820
COG0199	Ribosomal protein S14	2001002780
COG0228	Ribosomal protein S16	2001009770
COG0238	Ribosomal protein S18	2001059980
COG0268	Ribosomal protein S20	2001026760
COG0360	Ribosomal protein S6	2001059960
COG0522	Ribosomal protein S4 and related proteins	2001002670
COG0539	Ribosomal protein S1	2001067180
COG0828	Ribosomal protein S21	2001019110
	tRNA synthetases	
COG0008	Glutaminyl-tRNA synthetase	2001011240
COG0008	Glutamyl-tRNA synthetase	2001019500
COG0013	Alanyl-tRNA synthetase	2001030750
COG0016	Phenylalanyl-tRNA synthetase alpha subunit	2001027470
COG0017	Aspartyl/asparaginyl-tRNA synthetases	missing from neighbors
COG0018	Arginyl-tRNA synthetase	2001025010
COG0060	Isoleucyl-tRNA synthetase	2001063290
COG0072	Phenylalanyl-tRNA synthetase beta subunit	2001027480
COG0124	Histidyl-tRNA synthetase	2001058120
COG0162	Tyrosyl-tRNA synthetase	2001008660
COG0172	Seryl-tRNA synthetase	2001069610
COG0173	Aspartyl-tRNA synthetase	2001010350
COG0180	Tryptophanyl-tRNA synthetase	2001067940
COG0215	Cysteinyl-tRNA synthetase, class Ia	2001011450
COG0215	Cysteinyl-tRNA synthetase	2001019650
COG0423	Glycyl-tRNA synthetase (class II)	missing from neighbors
COG0441	Threonyl-tRNA synthetase, class IIa	2001027430
COG0441	Threonyl-tRNA synthetase	2001011430
COG0442	Prolyl-tRNA synthetase	2001068310
COG0495	Leucyl-tRNA synthetase	2001068920
COG0525	Valyl-tRNA synthetase	2001009980
COG0751	Glycyl-tRNA synthetase, beta subunit	2001005200
COG0752	Glycyl-tRNA synthetase, alpha subunit	2001005190
COG1190	Lysyl-tRNA synthetase (class II)	2001025520
	Translation Initiation	
COG0290	Translation initiation factor 3 (IF-3)	2001027440
COG0361	Translation initiation factor 1 (IF-1)	2001002700
COG0532	Translation initiation factor 2 (IF-2; GTPase)	2001067880
	Histidine biosynthesis	
COG0040	ATP phosphoribosyltransferase	2001008940
COG0079	Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	2001008960
COG0106	Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase	2001008990
COG0107	Imidazoleglycerol-phosphate synthase	2001009000

COG0118	Glutamine amidotransferase	2001008980
COG0131	Imidazoleglycerol-phosphate dehydratase	2001008970
COG0139	Phosphoribosyl-AMP cyclohydrolase	2001009010
COG0140	Phosphoribosyl-ATP pyrophosphohydrolase	2001009020
COG0141	Histidinol dehydrogenase	2001008950
COG0241	Histidinol phosphatase and related phosphatases	2001005220
COG0462	Phosphoribosylpyrophosphate synthetase	2001065350
	Chorismate biosynthesis	
COG0082	Chorismate synthase	2001027580
COG0128	5-enolpyruvylshikimate-3-phosphate synthase	2001067190
COG0169	Shikimate 5-dehydrogenase	2001017230
COG0337	3-dehydroquinate synthetase	2001010560
COG0703	Shikimate kinase	2001010570
COG0710	3-dehydroquinate dehydratase	2001013840
COG0722	3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	2001065730
COG1605	Chorismate mutase	2001067220
	Threonine biosynthesis	
COG0083	Homoserine kinase	missing from neighbors
COG0136	Aspartate-semialdehyde dehydrogenase	2001027630
COG0460	Homoserine dehydrogenase	2001024060
COG0498	Threonine synthase	2001019730
COG0527	Aspartokinases	2001009620
	Tryptophan biosynthesis	
COG0133	Tryptophan synthase beta chain	2001027690
COG0134	Indole-3-glycerol phosphate synthase	2001026800
COG0135	Phosphoribosylanthranilate isomerase	2001027680
COG0147	Anthranilate/para-aminobenzoate synthases component I	2001027010
COG0159	Tryptophan synthase alpha chain	2001027700
COG0512	Anthranilate/para-aminobenzoate synthases component II	2001026820
COG0547	Anthranilate phosphoribosyltransferase	2001026810
	CoA biosynthesis	
COG0237	Dephospho-CoA kinase	2001022500
COG0413	Ketopantoate hydroxymethyltransferase	2001062280
COG0414	Panthothenate synthetase	2001062270
COG0452	Phosphopantothenoylcysteine synthetase/decarboxylase	2001063400
COG0669	Phosphopantetheine adenylyltransferase	2001008710
COG0853	Aspartate 1-decarboxylase	2001062260
COG1072	Panthothenate kinase	missing from neighbors
COG1893	Ketopantoate reductase	absent
	FAD biosynthesis	
COG0054	Riboflavin synthase beta-chain	2001011980
COG0108	3,4-dihydroxy-2-butanone 4-phosphate synthase	2001011990
COG0117	Pyrimidine deaminase	2001095610
COG0196	FAD synthase	2001063280
COG0307	Riboflavin synthase alpha chain	2001020030
COG0807	GTP cyclohydrolase II	2001011990
COG1985	Pyrimidine reductase, riboflavin biosynthesis	2001095610

	Isoprenoid biosynthesis	
COG0245	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	2001064200
COG0743	1-deoxy-D-xylulose 5-phosphate reductoisomerase	2001066880
COG1154	Deoxyxylulose-5-phosphate synthase	2001009720
COG1211	4-diphosphocytidyl-2-methyl-D-erithritol synthase	2001064210
COG1947	4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase	2001065330
COG0761	Penicillin tolerance protein	2001063320
COG0821	Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis	2001058110
COG0020	Undecaprenyl pyrophosphate synthase	2001066900
	Purine biosynthesis	
COG0015	Adenylosuccinate lyase	2001063100
COG0026	Phosphoribosylaminoimidazole carboxylase (NCAIR synthetase)	2001099840
COG0027	Formate-dependent phosphoribosylglycinamide formyltransferase (GAR transformylase)	2001071750
COG0034	Glutamine phosphoribosylpyrophosphate amidotransferase	2001027750
COG0041	Phosphoribosylcarboxyaminoimidazole (NCAIR) mutase	2001099850
COG0046	Phosphoribosylformylglycinamidine (FGAM) synthase, synthetase domain	2001071650
COG0047	Phosphoribosylformylglycinamidine (FGAM) synthase, glutamine amidotransferase domain	2001071650
COG0104	Adenylosuccinate synthase	2001067290
COG0138	AICAR transformylase/IMP cyclohydrolase PurH (only IMP cyclohydrolase domain in Aful)	2001022940
COG0150	Phosphoribosylaminoimidazole (AIR) synthetase	2001062360
COG0151	Phosphoribosylamine-glycine ligase	2001022930
COG0152	Phosphoribosylaminoimidazolesuccinocarboxamide (SAICAR) synthase	2001023150
COG0299	Folate-dependent phosphoribosylglycinamide formyltransferase PurN	2001066200
COG0516	IMP dehydrogenase/GMP reductase	2001121130
COG0518	GMP synthase - Glutamine amidotransferase domain	2001121110
COG0519	GMP synthase, PP-ATPase domain/subunit	2001121110
COG0563	Adenylate kinase and related kinases	2001018330
	Pyrimidine biosynthesis	
COG0105	Nucleoside diphosphate kinase	2001058070
COG0167	Dihydroorotate dehydrogenase	2001067800
COG0283	Cytidylate kinase	2001067190
COG0284	Orotidine-5-phosphate decarboxylase	2001013760
COG0418	Dihydroorotase	2001020580
COG0458	Carbamoylphosphate synthase large subunit (split gene in MJ)	2001005690
COG0461	Orotate phosphoribosyltransferase	2001021740
COG0504	CTP synthase (UTP-ammonia lyase)	2001071920
COG0505	Carbamoylphosphate synthase small subunit	2001005680
COG0528	Uridylate kinase	2001066920
COG0540	Aspartate carbamoyltransferase, catalytic chain	2001004040
COG0125	Thymidylate kinase	2001062470**
COG0207	Thymidylate synthase	2001020350

COG0717	Deoxycytidine deaminase	2001008910
COG0756	dUTPase	2001063410
	Protein translocase Sec	
COG0201	Preprotein translocase subunit SecY	2001002710
COG0341	Preprotein translocase subunit SecF	2001063050
COG0342	Preprotein translocase subunit SecD	2001063040
COG0653	Preprotein translocase subunit SecA (ATPase, RNA helicase)	2001027070
COG0690	Preprotein translocase subunit SecE	2001003060
COG0706	Preprotein translocase subunit YidC	2001013690
COG1314	Preprotein translocase subunit SecG	2001006130
COG1862	Preprotein translocase subunit YajC	2001063030**
COG1952	Preprotein translocase subunit SecB	2001095700
	RNA polymerase subunits	
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit	2001002980
COG0086	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	2001002960
COG0202	DNA-directed RNA polymerase, alpha subunit/40 kD subunit	2001002660
COG0568	DNA-directed RNA polymerase, sigma subunit RpoD	2001019140
COG0568	DNA-directed RNA polymerase, sigma subunit RpoS	2001029300
COG1758	DNA-directed RNA polymerase, subunit K/omega	2001024490

* gene present but not called
** gene called on wrong strand