ORIGINAL ARTICLE

Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal

Paul Wilmes¹, Anders F Andersson¹, Mark G Lefsrud², Margaret Wexler³, Manesh Shah⁴, Bing Zhang⁵, Robert L Hettich⁶, Philip L Bond⁷, Nathan C VerBerkmoes⁶ and Jillian F Banfield^{1,8}

¹Department of Earth and Planetary Sciences, University of California at Berkeley, Berkeley, USA;

²Bioresource Engineering, McGill University, Ste-Anne-de-Bellevue, Canada; ³School of Biological Sciences, University of East Anglia, Norwich, UK; ⁴Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, USA; ⁵Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, USA;

⁶Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, USA; ⁷Advanced Wastewater Management Center, University of Queensland, St Lucia, Queensland, Australia and ⁸Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, USA

Enhanced biological phosphorus removal (EBPR) selects for polyphosphate accumulating microorganisms to achieve phosphate removal from wastewater. We used high-resolution community proteomics to identify key metabolic pathways in ‘*Candidatus Accumulibacter phosphatis*’ (*A. phosphatis*)-mediated EBPR and to evaluate the contributions of co-existing strains within the dominant population. Overall, 702 proteins from the *A. phosphatis* population were identified. Results highlight the importance of denitrification, fatty acid cycling and the glyoxylate bypass in EBPR. Strong similarity in protein profiles under anaerobic and aerobic conditions was uncovered (only 3% of *A. phosphatis*-associated proteins exhibited statistically significant abundance differences). By comprehensive genome-wide alignment of 13 930 orthologous proteins, we uncovered substantial differences in protein abundance for enzyme variants involved in both core-metabolism and EBPR-specific pathways among the *A. phosphatis* population. These findings suggest an essential role for genetic diversity in maintaining the stable performance of EBPR systems and, hence, demonstrate the power of integrated cultivation-independent genomics and proteomics for the analysis of complex biotechnological systems.

The ISME Journal (2008) 2, 853–864; doi:10.1038/ismej.2008.38; published online 1 May 2008

Subject Category: integrated genomics and post-genomics approaches in microbial ecology

Keywords: ‘*Candidatus Accumulibacter phosphatis*’; community proteogenomics; enhanced biological phosphorus removal; metaproteomics; proteomics; strain variation

Introduction

Inorganic phosphate (P_i) is a principal limiting nutrient in freshwater ecosystems. High nutrient

loads, mainly from anthropogenic sources, impact on the quality of surface waters and may promote the deleterious effect of eutrophication (Correll, 1998).

Rapid human population growth and global climate change are placing considerable pressures on the global water cycle (United Nations World Water Assessment Programme, 2006). Aquatic P_i levels can be reduced through biological wastewater treatment. Activated sludge undergoing alternating anaerobic and aerobic regimes favors the growth of polyphosphate accumulating microorganisms that enable enhanced biological phosphorus removal (EBPR; Seviour *et al.*, 2003). Continuous removal

Correspondence: JF Banfield, Department of Earth and Planetary Sciences, Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA 94720, USA.

E-mail: jbanfield@berkeley.edu and

PL Bond, Advanced Wastewater Management Center, University of Queensland, St Lucia, Queensland 4072, Australia.

E-mail: phil.bond@uq.edu.au

Received 18 January 2008; revised 25 March 2008; accepted 25 March 2008; published online 1 May 2008

of defined amounts of polyphosphate-rich biomass facilitates P_i removal and the treated supernatant can be safely discharged into surface waters.

EBPR installations are used extensively throughout the world because they are considered more cost-effective and sustainable compared to chemical precipitation (Blackall *et al.*, 2002). Although EBPR has been empirically optimized since its fortuitous discovery in the 1950s (Srinath *et al.*, 1959), detailed knowledge of the microbiology and biochemistry is required to improve operational stability.

Since its inception, basic research on EBPR has been plagued by enrichment bias. Several potential polyphosphate accumulating organism (PAO) isolates (predominantly *Acinetobacter* spp.) were obtained from EBPR biomass but their phenotypes were inconsistent with the biochemical transformations observed in full- and laboratory-scale systems (Blackall *et al.*, 2002; Seviour *et al.*, 2003). With the advent of molecular tools (16S rRNA sequences), a dominant phylotype belonging to the order *Rhodocyclales* was identified within EBPR-activated sludge (Bond *et al.*, 1995; Hesselmann *et al.*, 1999; Crocetti *et al.*, 2000). Putatively named '*Candidatus Accumulibacter phosphatis*' (*A. phosphatis*; Hesselmann *et al.*, 1999), it has since been repeatedly detected in activated sludge in various locations on Earth (Blackall *et al.*, 2002).

Although the identity of *A. phosphatis* as a major PAO has been established, details regarding its metabolism have proven controversial (Seviour *et al.*, 2003). Even the recent landmark paper by García Martín *et al.* (2006), describing the genetic inventory of dominant *A. phosphatis* populations, has not abated discussions regarding its metabolic potential. Causes célèbres include the ability of *A. phosphatis* to denitrify (Kong *et al.*, 2004; García Martín *et al.*, 2006), the nature of its glycolytic pathway (Hesselmann *et al.*, 2000; García Martín *et al.*, 2006) and how it generates reducing equivalents during the anaerobic phase (Mino *et al.*, 1998; García Martín *et al.*, 2006). The ability to detect proteins involved in candidate pathways should help shed light on these unresolved issues.

Proteomics employing liquid chromatography (LC) coupled with tandem mass-spectrometry (MS) can provide comprehensive high-resolution analysis of the protein complement of natural microbial communities (Ram *et al.*, 2005; Wilmes and Bond, 2006a; Lo *et al.*, 2007). So far, comprehensive community proteogenomic analyses have provided unprecedented insight into the structure and function of microbial biofilms inhabiting warm, extremely acidic and metal-rich solutions within the Richmond Mine at Iron Mountain, California, USA (Ram *et al.*, 2005; Lo *et al.*, 2007). These biofilms are ideally suited for the development of community genomic and proteomic methodologies due to their limited species richness and abundant biomass. However, the application of proteogenomic techniques to more complex microbial ecosystems pre-

sents additional challenges. Here we apply shotgun proteomics to a complex activated sludge community cultured in a laboratory-scale sequencing batch reactor (SBR) operated in the United Kingdom (UK), enriched for *A. phosphatis* and performing EBPR at high P_i influent levels. Metagenomic sequences obtained from laboratory-grown sludges derived from wastewater treatment plants in the United States (US) and in Australia (OZ; García Martín *et al.*, 2006) were employed to identify abundant proteins. The identified proteins were interpreted within the context of current EBPR metabolic models. Importantly, very high mass accuracy data were obtained using a linear ion trap (LTQ)-Orbitrap mass spectrometer. This allowed confident identification of proteins derived from distinct strains within the *A. phosphatis* population. Differences in the abundances of protein variants associated with different *A. phosphatis* subpopulations may be a reflection of functional partitioning within the population and highlight the apparent importance of genetic diversity in maintaining stable process performance.

Materials and methods

Laboratory-scale SBR

A laboratory-scale SBR with alternating anaerobic/aerobic phases was operated as previously reported (Wilmes and Bond, 2004, 2006b). The reactor was initially seeded with activated sludge from the Whitlingham WWTP, Norwich, UK. Stable EBPR was obtained with complete removal of 55 mg l^{-1} of P_i from the synthetic wastewater feed with acetate as the main carbon source (Wilmes *et al.*, 2008).

Clone library construction and sequence analysis

DNA was extracted as described earlier (Bond *et al.*, 1995). PCR primers specific for '*Candidatus Accumulibacter*' polyphosphate kinase 1 (Acc-ppk1-254f and Acc-ppk1-1367r) were used for amplification as previously described (He *et al.*, 2007). Primers specific for sludge malate synthase were designed using the metagenomic sequences and were as follows: MS3F (5'-ATTCCGATCAAGAACGATCC-3') and MS3R (5'-CTCAGCGACATCTTGTCGAA-3'). Clone libraries were prepared using the TOPO TA Cloning kit (Invitrogen, Paisley, UK). Nine clones were analyzed per gene, grouped by restriction enzyme analysis and representatives sequenced. Evolutionary analyses of sequence data were performed by distance methods using ARB (Ludwig *et al.*, 2004). DNA base sequence was converted to amino acid sequence, aligned and phylogenetic trees were assembled by neighbor joining and maximum likelihood analyses. The GenBank accession numbers for the nucleotide sequences determined in this study are EU563082 to EU563085.

Proteomics

Proteins were extracted from the SBR biomass at the end of the anaerobic and aerobic phases as described earlier (Wilmes and Bond, 2004, 2006b). Trichloroacetic acid precipitated protein pellets were resuspended in 6 M Guanidine/10 mM DTT for denaturation and reduction, and then digested into peptides with sequencing grade trypsin (Promega, Madison, WI, USA). Samples were de-salted, filtered and concentrated. The anaerobic and aerobic samples were analyzed via 24-h, 12-step 2-D LC (SCX-RP)-MS/MS runs on an LTQ and a hybrid LTQ-Orbitrap with technical duplication (Ram *et al.*, 2005; Lo *et al.*, 2007). Both instruments were operated in data-dependent MS/MS mode. For the LTQ-Orbitrap, full scans were acquired at 30 000 resolution in the Orbitrap, while MS/MS scans were acquired in the LTQ. Separate and combined database searches were performed with SEQUEST (Eng *et al.*, 1994) against the three available sludge metagenomic databases (García Martín *et al.*, 2006; Supplementary Table 1). Detailed information on the metagenomic sequences was retrieved using the integrated microbial genomes with microbiome samples (IMG/M) system (Markowitz *et al.*, 2006). In addition to the metagenomic protein sequences, common contaminants (trypsin, keratin, etc.) were also included in the search database. MS data were filtered with DTASelect (Tabb *et al.*, 2002) at the peptide level [Xcorr of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)]. For positive protein identification, a minimum of two fully tryptic peptides was required per protein in at least one replicate run.

False-positive levels were estimated by reverse database searching and were about 1% for both instruments (deltCN filter of 0.08). High mass accuracy monoisotopic peaks were extracted from the LTQ-Orbitrap full scans and compared to theoretical monoisotopic masses for all identified peptides. Overall, 70% of proteins were identified using both instruments.

To test the uniqueness of identifications against more extended database space, a large database was created and MS data from two runs were searched against it. This database contained: the Phrap assembly of the US sludge (USP), AMD metagenome (12 148 proteins), 249 microbial isolates from IMG (924 816 proteins). From this search, 90% of the unique peptides matched the USP database and only 10% matched with the other database entries. All datasets, databases and supplementary files (spreadsheets in.xls format) can be viewed at http://compbio.ornl.gov/ebpr_sludge.

Protein abundances were determined from sequence coverage (percentage of protein covered by detected peptides), total spectral count (number of spectra detected per protein), unique peptide count (number of times any unique peptide is detected from a protein) and normalized spectral abundance factor (Florens *et al.*, 2006). We compared the differences in protein abundances between the

anaerobic and aerobic phases using a Poisson regression model with total spectral count as the outcome and phase as the main independent variable. In addition, we included instrument as an independent variable to adjust for a potential confounding instrument effect. As we were comparing differential expression for 4688 proteins (combined database), *P*-values generated by the model were further adjusted using the Benjamini and Hochberg correction to account for multiple comparisons (Benjamini and Hochberg, 1995). An adjusted *P*-value of 0.10 (that is 10% False Discovery Rate) was used to select proteins differentially expressed between the anaerobic and aerobic phases.

Tetranucleotide frequency calculation and self-organizing maps

A Perl script was written to calculate tetranucleotide frequencies of the US sludge's Jazz assembly scaffold sequences, serving as input for self-organizing map (SOM) training. We employed the Databionics ESOM program package (Ultsch and Moerchen, 2005) for training and visualizing of self-organizing maps (Kohonen, 1990). To reduce coding strand biases, frequencies of pairs of reverse complementary tetranucleotides were summed (Abe *et al.*, 2005). A 328 × 100 torroid SOM was trained on scaffolds ≥ 2000 bp. Scaffolds ≥ 10 000 bp were fragmented into subsequences by splitting after each 5 kb, such that the last subsequence contained ≥ 5000 and < 10 000 kb of sequence. Weights were initialized by principal component analysis, the order of the input data was randomized and 40 training epochs of 15%-batch training were performed. Onto the resulting map, the tetranucleotide frequencies of all scaffolds ≥ 1000 bp were projected. Scaffolds ≥ 10 000 bp were fragmented as above. Scaffold fragments were highlighted according to previous binning information (García Martín *et al.*, 2006; McHardy *et al.*, 2007). For this, information from the different binning methods (US/OZ overlap, PhyloPythia > 85% confidence and PhyloPythia < 85% confidence) was concatenated according to highest taxonomic affiliation. A scaffold sequence fragment was considered expressed if it overlapped with > 50% of an open reading frame (ORF) that was detected in any of the mass-spectrometry runs. The background was colored according to the U-matrix (Ultsch and Moerchen, 2005).

Ortholog alignment

A programme was written in Perl for aligning contigs and scaffolds against a predefined set of backbone scaffolds, based on sequence similarity of the encoded proteins. Protein sequence similarity was deduced by BLASTP (Altschul *et al.*, 1997); the sequence identity for a pair of proteins was defined

as the number of identical amino acids divided by the length of the smallest protein. Starting with the largest backbone scaffold and progressing in decremental order, all previously non-aligned, non-backbone contigs and scaffolds were aligned fulfilling the following criteria: (a) At least 50% of the proteins in the non-backbone contig or scaffold were at least 90% identical with proteins in the backbone scaffold, (b) The matching proteins were syntenous in the aligned contigs and/or scaffolds, allowing a maximum of 10 ORF insertions in either of the contigs or scaffolds, between any two matching proteins.

The resulting alignment file can be viewed in a spreadsheet format with the backbone scaffolds displayed in the first column, with one IMG/M gene object identifier per row and aligned contigs and scaffolds in succeeding columns. The Perl programme is available from the authors upon request.

Results

SBR performance and community structure

A laboratory-scale SBR was successfully operated for complete removal of around 55 mg l^{-1} of P_i from synthetic wastewater feed (Wilmes *et al.*, 2008).

Fluorescence *in situ* hybridization employing 16S rRNA oligonucleotide probes revealed dominance of *A. phosphatis* (69% of bacterial cells, s.d. 10.8; Wilmes *et al.*, 2008). As the 16S rRNA locus is not ideally suited for discriminating between closely related organisms, we sequenced fragments of the polyphosphate kinase 1 (*ppk1*) gene, a more sensitive phylogenetic marker for distinguishing *Accumulibacter* species and *A. phosphatis* strains (He *et al.*, 2007). We retrieved three distinct *ppk1* gene sequences from the UK sludge (representing nine clones) and compared them to previously reported *ppk1* sequences from EBPR-activated sludges (Supplementary Figure 1 online). This revealed phylogenotypes in the UK sludge that were distinct from the US and OZ phylogenotypes. Furthermore, the UK sequences clustered into two specific clades (I and IID), suggesting greater phylogenetic diversity in the UK sludge compared to the US and OZ sludges.

Protein identifications

Proteins were extracted from the UK sludge at the end of the anaerobic ($t = 120 \text{ min}$) and the aerobic phases ($t = 330 \text{ min}$), proteolytically digested with trypsin and analyzed in duplicate via two-dimensional nano-LC followed by MS/MS analysis using either an LTQ or a hybrid LTQ-Orbitrap mass spectrometer. Peptide tandem mass spectra were matched *in silico* to predicted peptides and, thus, proteins in the sludge metagenomic databases (García Martín *et al.*, 2006; Supplementary Table 1 online). Three distinct metagenomic sequence databases were used for protein identification: USJazz

(Jazz assembly of the US sludge; USJ), USPhrap (US Phrap assembly; USP) and OZPhrap (OZ Phrap assembly; OZP). For positive protein identification, a minimum of two peptides were required per protein in at least one replicate run (Ram *et al.*, 2005). The MS data were searched separately against the metagenomic databases and against a combined database (Supplementary Table 2). A total of 5029 proteins were identified using both instruments and database combinations (Supplementary Table 3). However, this number is inflated due to database redundancy (mainly between USJ and USP). When searching the mass spectra solely against the OZP and USP databases (of similar size), 916 and 980 (LTQ and LTQ-Orbitrap) versus 1857 and 1728 non-redundant positive identifications were obtained, respectively. The results suggest that the UK sludge contained genes present in both the OZ and US sludges. However, based on protein identifications, the genetic composition of the UK sludge appears more similar to the US sludge than the OZ sludge.

Assignment of identified proteins to organisms

Overall, 36% of identified proteins were encoded by genes located on genomic contigs or scaffolds assigned to *A. phosphatis* (García Martín *et al.*, 2006; McHardy *et al.*, 2007; Figure 1a). In addition, several identified proteins were inferred to derive from other *Accumulibacter* species or other β -*Proteobacteria* (2 and 3%, respectively). Numerous proteins affiliated with γ -*Proteobacteria* were also identified (3%).

In all, 56% of identified proteins were encoded on contigs and scaffolds whose organismal affiliation was uncertain based on the previous metagenomic analysis (García Martín *et al.*, 2006; McHardy *et al.*, 2007). Many of these proteins were found to be highly abundant, based on protein sequence coverages, total spectral counts and normalized spectral abundance factors (NSAFs; Florens *et al.*, 2006) and likely derive from other *A. phosphatis* strains, other *Accumulibacter* species, or active members of the flanking populations.

In total, 14% of all detected proteins were of unknown function, 18% of which exhibited uncharacterized conserved domains. Interestingly, the largest number of assigned mass spectra was obtained for an unclassified protein of unknown function encoded on a USP contig (IMG/M gene object identifier: 2000289750). This protein has a conserved domain homologous to zinc peptidases found in several *Proteobacteria*, *Chlorobiaceae* and *Rhodopirellula baltica* and was more abundant in the aerobic protein extract. The specific role of this protein in the context of EBPR has yet to be determined.

To constrain the affiliation of sequences related to, but distinct from, the composite *A. phosphatis* genome (García Martín *et al.*, 2006), we calculated tetranucleotide frequencies for all USJ scaffolds and

Figure 1 Taxonomic affiliation of identified proteins. (a) Proportions of all identified proteins that belong to taxonomic groups based on concatenated genomic contig and scaffold binning (García Martín *et al.*, 2006; McHardy *et al.*, 2007). (b) Self-organizing map based on tetranucleotide frequency of USJ scaffold fragments with genomic fragments encoding identified proteins highlighted according to taxonomic groups (the *A. phosphatis* fragment cluster is delimited by a yellow line). Information in both panels is color-coded according to the concatenated binning information: ■ *A. phosphatis*, ■ *Accumulibacter*, ■ *Betaproteobacteria*, ■ *Gammaproteobacteria*, ■ *Thiothrix*, ■ *Actinobacteria*, ■ *Alphaproteobacteria*, ■ *Sphingobacteria*, ■ *Deinococci*, ■ *Deltaproteobacteria*, ■ *Clostridia*, ■ *Planctomycetacia*, ■ *Spirochaetes*, ■ *Methanomicrobia*, ■ *Chromadorea*, ■ *Insecta*, ■ *Mollicutes*, ■ unclassified.

visualized the results in two-dimensional space using a self-organizing map (SOM; Kohonen, 1990; Abe *et al.*, 2005). To identify and evaluate the SOM clusters, we overlaid the corresponding sludge binning information (Supplementary Figure 2a). A notable feature is the large fraction of unclassified scaffold fragments that cluster with *A. phosphatis*, *Accumulibacter* or β -*Proteobacteria* (64% of scaffold fragments) and, hence, these likely derive from closely related organisms.

We overlaid protein expression data on the SOM. A total of 89% of the genomic fragments encoding identified proteins fall within the *A. phosphatis* region (Figure 1b and Supplementary Figure 2b), and 41% of these were encoded on previously unclassified scaffolds. These genomic fragments encoding identified proteins of possible *A. phosphatis* strains may have marked functional significance in the UK sludge.

Analysis of *A. phosphatis* strain variation

Strain variation among closely related organisms in complex microbial communities may result from single nucleotide polymorphisms, genome rearrangements, as well as gene insertions and deletions. These events can fragment genomic assemblies. Resulting small contigs and scaffolds may be placed in genome context by mapping them against the assembled composite *A. phosphatis* sequence (García Martín *et al.*, 2006) using syntenous and shared genes. We developed an automated approach for aligning variant contigs and scaffolds that required 50% of the ORFs on a fragment to share >90% amino acid sequence identity with the composite backbone genome and allowing a maximum of 10 ORF insertions (Supplementary Table 4). The 50% aligned ORF criterion was empirically determined to account for differences in genetic variation among neighboring genes, and the 90% amino acid identity cutoff was chosen so that orthologous proteins would be identifiable by both unique and non-unique peptides in most cases (Denef *et al.*, 2007). Unique peptides are specific to a given protein in the database, whereas non-unique peptides may be derived from multiple proteins. Gaps in the alignment of orthologous contigs and scaffolds highlight the importance of gene insertions and deletions within the *A. phosphatis* population (Supplementary Table 4).

A. phosphatis protein expression

The composite *A. phosphatis* genomic fragments from USJ and aligned contigs and scaffolds from USJ, USP and OZP were used to interpret *A. phosphatis* protein expression. Overall, 559 proteins were identified using the composite genome backbone (MS data searched against separate databases; average protein coverage: 13%). This represents around 10% of the predicted *A. phosphatis* genes (and more than half of all non-redundant proteins identified using USJ). 1534 proteins with orthologs on the backbone were identified on aligned contigs and scaffolds, representing a non-redundant count of 556 proteins. Among these, 78 were strain-specific variants identified based on peptides specific to proteins encoded by aligned contigs and scaffolds and, hence, not identified using the backbone sequences. In addition, 1338 proteins encoded by aligned contigs and scaffolds but not by the backbone sequence were recruited and 68 of these were identified by proteomics. These aligned proteins are encoded by inserted genes within the variant subpopulations.

Many orthologs present on the backbone and aligned contigs and scaffolds were identified based on unique and non-unique peptides. Overall, proteins encoded by aligned scaffolds and contigs were found to be more abundant than the backbone variants in 414 cases (based on total spectral counts; Supplementary Figure 3 and Supplementary Table 5).

Proteomics-enabled metabolic insights

The abundances of proteins identified in the anaerobic and aerobic phases were very similar for both instruments (Supplementary Figure 4). Four percent of all identified proteins exhibited statistically significant differences in abundance between the anaerobic and aerobic phases (see Methods; Supplementary Table 6). Only 3% of identified proteins from the *A. phosphatis* population and 4% of proteins not associated with the *A. phosphatis* population showed different abundance levels between the two phases. These results were unexpected, given distinct differences in chemical transformations performed under each condition. However, some differentially expressed proteins could be directly linked to the metabolic transformations characteristic of EBPR, including poly(3-hydroxyalkanoate) synthase [anaerobic polyhydroxyalkanoate (PHA) synthesis], glycosidase and glucan phosphorylase (anaerobic glycogenolysis), glyceraldehyde-3-phosphate dehydrogenase (anaerobic glycolysis) and glycogen synthase and 1,4- α -glucan branching enzyme (aerobic glycogen synthesis).

Protein abundance information was placed into metabolic context and the extent to which proteins may play dual roles under anaerobic and aerobic conditions was evaluated. Proteins were assigned to processes involved in P_i transformations, PHA cycling, fatty acid cycling, tricarboxylic acid (TCA)/glyoxylate cycle, carbohydrate transformations, energy production and conversion, denitrification and exopolysaccharide (EPS) synthesis (Figure 2).

Many of the proteins involved in the metabolic pathways exhibited sequence variation compared to the *A. phosphatis* composite genome. Notably, proteins encoded by aligned genome fragments and involved in key metabolic transformations were found to be more abundant than proteins encoded by the backbone reference sequence (Figure 3a, Table 1, Supplementary Figure 3 and Supplementary Table 5). In particular, a large proportion of variant proteins involved in fatty acid metabolism, the TCA/glyoxylate cycle and sugar metabolism exceeded backbone proteins in terms of abundance (Table 1 and Supplementary Figure 3). For example, the proteomic results indicated that the malate synthase protein variant (USP_MALS_1) encoded by an aligned contig was more abundant than that encoded by the backbone (Figure 3a). We confirmed this deduction using the dominant malate synthase gene sequence amplified from our bioreactor (Figure 3b).

Discussion

Using proteogenomics, we identified 702 proteins from the *A. phosphatis* population present in the UK sludge. This provided sufficient information to conduct a detailed analysis of metabolic activity during EBPR. Apart from proteins affiliated with *A. phosphatis*, numerous proteins associated with γ -*Proteobacteria* were identified. The presence of γ -*Proteobacteria* was previously reported in failing EBPR systems (Crocetti *et al.*, 2002), and they have been shown to actively compete with PAOs over carbon substrates (Kong *et al.*, 2005). Indeed, the abundant γ -*proteobacterial* proteins identified suggest a metabolism analogous to *A. phosphatis*' carbon metabolism.

A. phosphatis and closely related organisms are cycled through alternating anaerobic and aerobic phases, both characterized by specific metabolic transformations. During the anaerobic phase, polyphosphate (polyP) and glycogen are degraded, and PHAs are synthesized from short chain volatile fatty acids. This process is reversed during the aerobic phase with the replenishment of both polyP and glycogen in conjunction with the hydrolysis of PHAs. Owing to the distinct metabolic transformations apparent in the two contrasting phases, protein expression differences were anticipated between both conditions. However, protein abundances were similar for the vast majority of metabolic enzymes, as highlighted in the metabolic model (Figure 2). Only 3% of identified proteins from the *A. phosphatis* population versus 4% of proteins not associated with the *A. phosphatis* population exhibited expression differences between the two phases. This may result in part from carryover of proteins from the anaerobic (115 min long) to the aerobic (210 min) phase and vice versa. However, an interesting possibility is that the majority of enzymes are able to catalyze reactions in both forward and reverse directions. For example, in *E. coli*, the enzymes involved in glycolysis and gluconeogenesis are shared except for the two key enzymes that regulate the directionality of the carbon flow, that is phosphofructokinase (glycolysis) and fructose-1,6-bisphosphate phosphatase (gluconeogenesis) and these are regulated allosterically by intracellular levels of ADP and AMP, respectively (White, 2007). Similarly, allosteric regulation of key metabolic steps in *A. phosphatis* would enable rapid switching in response to alternating anaerobic and aerobic conditions and provide a bioenergetic advantage to the

Figure 2 Proposed model for EBPR-specific metabolism carried out by the *A. phosphatis* population during the (a) anaerobic and (b) aerobic phases. Proteins involved in P_i transformations, polyhydroxyalkanoate (PHA) cycling, fatty acid cycling, tricarboxylic acid/glyoxylate cycle, carbohydrate transformations, energy production and conversion, denitrification and exopolysaccharide synthesis are highlighted according to their relative abundances. Full arrows indicate the presence of the coding gene on the *A. phosphatis* composite genome, dashed arrows refer to orthologous identified proteins. For more information on identified proteins, please see Supplementary Table 7 online.

A. phosphatis population, since protein synthesis and degradation are costly processes. A similar suggestion was made in a previous low-resolution

proteomic study (Wilmes and Bond, 2006b). Alternatively, similarities in protein profiles between the two alternating growth conditions may indicate a role for post-translational modifications. For example, in *Streptomyces coelicolor*, proteins involved in both primary and secondary metabolic pathways exhibit substantial post-translational modifications (Hesketh *et al.*, 2002). Future quantitative analyses of protein abundances (for example, enabled by stable isotope labeling; Eng and Mann, 2005) in both

Table 1 Numbers of aligned variant proteins determined to be more abundant than orthologous proteins encoded by the backbone composite *A. phosphatis* genome according to functional categories

Functional category	Number (percentage)
Phosphate transport and metabolism	5 (1.2)
Polyhydroxyalkanoate metabolism	4 (1.0)
Fatty acid metabolism	30 (7.2)
TCA/glyoxylate cycle	30 (7.2)
Carbohydrate metabolism	33 (8.0)
Nitrogen metabolism	6 (1.4)
Energy production and conversion	28 (6.8)
Transport	30 (7.2)
Signal transduction mechanisms	23 (5.6)
Amino acid metabolism	59 (14.3)
Translation, ribosomal structure and biogenesis	17 (4.1)
Post-translational modification, protein turnover, chaperones	23 (5.6)
Transcription	6 (1.4)
DNA/RNA processing, replication, recombination and repair	25 (6.0)
Nucleotide metabolism	1 (0.2)
Cell cycle control, cell division, chromosome partitioning	8 (1.9)
Intracellular trafficking, secretion and vesicular transport	2 (0.5)
Flagellum structure and biosynthesis	2 (0.5)
Outer membrane and extracellular structures	4 (1.0)
Defense mechanisms	9 (2.2)
General function	39 (9.4)
Unknown function	30 (7.2)

Percentages of more abundant variant proteins belonging to the different functional categories provided in brackets.

Figure 3 Identification of sludge protein variants. **(a)** Examples of relative protein abundances for key metabolic enzymes arranged in genome order (from top to bottom) that exhibit substantial differences between the backbone (BB) composite *A. phosphatis* genome and aligned orthologous sequences from USJ, USP and OZP. Black elements indicate aligned orthologous proteins that were not identified and grey elements indicate absence of aligned orthologous proteins. Malate synthase highlighted in red. **(b)** Alignment of amino acid and corresponding nucleotide sequences (arranged by codon sequence from top to bottom) of orthologous malate synthases from the US and OZ sludges. Predicted amino acid sequence and amplified nucleotide sequence of the dominant UK malate synthase included in the bottom row for comparison. Peptide sequences detected using the LTQ-Orbitrap highlighted in green, predicted UK peptides are shaded in grey. Amino acid substitutions are highlighted in red and non-synonymous single nucleotide polymorphisms highlighted in blue.

phases and detection of post-translationally modified proteins (Tyers and Mann, 2003) will more definitively resolve these issues.

P_i transformations underpin the pre-eminent role of the *A. phosphatis* population in EBPR. Multiple mechanisms for polyP degradation and energy generation may be employed during the anaerobic phase (Figure 2a). Exopolyphosphatase (PPX) and polyphosphate kinase 2 (PPK2) both degrade polyP. PPX produces P_i , which in turn allows the maintenance of the proton motive force either through proton transport pyrophosphatase or by export via a low affinity P_i transporter and, hence, production of ATP via F_0F_1 -type ATP synthases. In contrast, PPK2 produces GTP. A high abundance of identified GTPases within the proteomic data suggests an essential role for GTP as an energy transfer molecule within the *A. phosphatis* population. However, this correlation requires future experimental validation. Although polyphosphate kinase 1 (PPK1) obtained from an *A. phosphatis*-dominated community was previously found to be unable to hydrolyze polyP (McMahon *et al.*, 2002), an *A. phosphatis* ortholog was only identified in the anaerobic protein extract (Supplementary Table 6), suggesting a possible involvement of PPK1 in anaerobic polyP degradation.

In the EBPR metabolic model, it is proposed that polyphosphate AMP phosphotransferase hydrolyses polyP to produce ADP, which in turn is converted to ATP by adenylate kinase (Figure 2a). This cycle can be reversed during the aerobic phase, allowing the replenishment of polyP (Figure 2b). Simultaneously, PPK1 may produce polyP at the expense of ATP in the aerobic phase and PPK2 may produce polyP at the expense of GTP. Excess GDP may inhibit PPX via production of guanosine bisdiphosphate (ppGpp) and guanosine pentaphosphate (pppGpp), resulting in excess polyP accumulation (Kornberg *et al.*, 1999). PolyP synthesis requires uptake of extracellular P_i via low (PIT) and high (PST) affinity transporters. Given the relatively high levels of P_i in EBPR, it may be assumed that most P_i translocation occurs through PIT, as described earlier (García Martín *et al.*, 2006). Interestingly, a potential PIT was more abundant in the anaerobic phase, whereas another was solely identified in the aerobic phase (Figure 2 and Supplementary Table 6). This raises the possibility that different specialized PIT transporters may be active in the anaerobic and aerobic phases. Furthermore, the unexpected high abundance of PST in the anaerobic phase may indicate P_i efflux via PST during the anaerobic phase, possibly coupled to ATP generation (Figure 2a).

The anaerobic carbon metabolism of EBPR has attracted considerable attention. Short-chain fatty acids (acetate in our case) are rapidly assimilated during the anaerobic phase. According to protein abundances, acetate is preferentially activated via the high affinity enzyme acetyl-CoA synthase (Figure 2a). Glycogenolysis and glycolysis provide

reducing equivalents [NAD(P)H] for the subsequent polymerization of acetyl-CoA to PHAs (Figure 2a). The PHA content of EBPR biomass generally exceeds theoretical levels obtained through provision of NAD(P)H via glycogen catabolism alone (Schuler and Jenkins, 2003; Seviour *et al.*, 2003). Additional sources have been suggested and include the anaerobic operation of the TCA (Pereira *et al.*, 1996) and/or glyoxylate cycle (Louie *et al.*, 2000). However, García Martín *et al.* (2006) highlighted the requirement for anaerobic reoxidation of reduced quinones produced by succinate dehydrogenase in the absence of electron acceptors, and proposed that this may be achieved by a novel cytochrome *b/b6*. From our proteomic data, we are unable to verify this hypothesis, as this cytochrome was not identified. Instead, we suggest the following four alternative metabolic processes for the flux of reducing equivalents during the anaerobic phase:

- (a) *Fatty acid β oxidation*: Proteins involved in fatty acid degradation and synthesis were identified and found to be highly abundant. An *A. phosphatis* acyl-CoA dehydrogenase (involved in the first enzymatic step of fatty acid β oxidation) was only identified in the anaerobic protein extract whereas an acetyl-CoA carboxylase (first enzymatic step of fatty acid biosynthesis) was only abundant in the aerobic phase (Supplementary Table 6). Furthermore, the identified enzyme enoyl-CoA hydratase (PhaJ) directly links fatty acid β oxidation to PHA formation (Figure 2a). Consequently, apart from the requirement of fatty acids for lipid biosynthesis, β oxidation of fatty acids stored during the aerobic phase may provide reducing equivalents for PHA formation in the anaerobic phase as suggested previously (Wilmes *et al.*, 2008).
- (b) *Denitrification and operation of full TCA/glyoxylate cycle*: Laboratory-scale SBRs dominated by *A. phosphatis* have shown ability to denitrify (Zeng *et al.*, 2003). Although the majority of enzymes involved in denitrification were encoded by the composite *A. phosphatis* genome, García Martín *et al.* (2006) lamented the absence of a respiratory nitrate reductase (NAR) due to the lack of quinol reductase subunit (*napC*) in the NAR operon. In addition to NAR subunits (*napDAGHB*), we found a *napC* homologue elsewhere in the composite genome, which was identified by proteomics. A similar disjointed NAR operon structure has been reported previously in *Shewanella putrefaciens*, which is a denitrifier (Richardson *et al.*, 2001). Several subunits of NAR and subunits of other enzymes involved in denitrification were identified by proteomics (Figure 2a). Consequently, we infer an active denitrification pathway in *A. phosphatis*, which enables the operation of the TCA and/or glyoxylate cycle when nitrate is present in the anaerobic phase (Figure 2a).

- (c) *Anaerobic respiration via a split TCA cycle*: Fumarate reductase was identified. This enzyme enables the operation of a split TCA cycle, providing propionyl-CoA moieties for inclusion into PHAs (Figure 2a).
- (d) *Glyoxylate shunt*: Isocitrate lyase and malate synthase were identified possibly facilitating a partial TCA cycle during the anaerobic phase (Figure 2a).

From nuclear magnetic resonance studies, the Entner–Doudoroff (ED) pathway had been considered to be the main glycolytic pathway employed by *A. phosphatis* (Hesselmann *et al.*, 2000). However, all the enzymes for the Embden–Meyerhof (EM) pathway are present on the composite genome (García Martín *et al.*, 2006) and were identified by proteomics (Figure 2). In contrast, only one putative ED specific enzyme, 6-phosphogluconate dehydratase, was identified. Combined metabolomic and proteomic studies are required to conclusively demonstrate the importance of the EM pathway.

EPS produced by sludge-dwelling organisms are essential for floc formation and, hence, retention of cells within the EBPR cycle. EPS biosynthesis cassettes within *A. phosphatis* were found to differ between the US and OZ sludges (García Martín *et al.*, 2006). Based on the anomalously high degree of variability among EPS biosynthesis modules, it is not surprising that we did not identify EPS biosynthesis proteins described by García Martín *et al.* (2006) in the UK sludge. The possible EPS biosynthesis proteins we identified indicate that the extracellular polymers in our reactor would mainly be constituted of glucose, galactose and rhamnose sugars (Figure 2b).

A most striking finding from the proteogenomic analysis is the differential detection of orthologous proteins from closely related *A. phosphatis* variants (Figure 3a and Supplementary Figure 3). Reactors operating under subtly different conditions should enrich for particular protein variants, but overall ecosystem heterogeneity will retain genetic diversity. A large proportion of variant proteins involved in EBPR-specific biochemical processes as well as ‘housekeeping’ functions were more abundant than those encoded by the *A. phosphatis* composite genome (Table 1 and Supplementary Table 5). Such variation likely alters the kinetics and specificity of enzymes and may provide a means for the adaptation of distinct strains to specific environmental conditions. As illustrated by malate synthase, the relatively high abundance of a particular variant in the UK sludge is likely due to conditions that favored growth of an *A. phosphatis* strain that was only a minor constituent in the US sludge (Figure 3b).

Strain variation within the *A. phosphatis* population may be selectively neutral. However, a more likely explanation is that genes specific to one strain convey distinct functionality within the population.

Niche partitioning and genetic adaptation are essential for selection of organisms along specific nutrient gradients (Johnson *et al.*, 2006). Within activated sludge, microbial flocs harbour numerous micro-niches for the dominant population and flanking organisms. Furthermore, significant gradients have been previously described in wastewater bioreactors (Schramm *et al.*, 1999). The proteomic data suggest the coexistence of functionally distinct strain variants, which may be important for community optimization in the heterogeneous EBPR system. Interestingly, these findings are analogous to results of a macroecological study in which P_i removal was enhanced in wetland mesocosms with higher macrophyte diversity (Engelhardt and Ritchie, 2002). Furthermore, increases in microbial diversity have been shown to increase the level of ecosystem functioning (Bell *et al.*, 2005).

Our results suggest that proteins differing by only a few amino acids may be functionally distinct. Hence, proteogenomic investigations of complex microbial systems require resolution at the strain level (Lo *et al.*, 2007). Unfortunately, genomic coverage of all constituent members within complex microbial communities is unrealistic. However, the large-scale implementation of MS-based top-down and bottom-up *de novo* sequencing (Steen and Mann, 2004) will diminish the requirement for comprehensive genomic foundations, and allow strain-resolved analysis of even more complex microbial ecosystems in the future.

The percentage of proteins encoded by *A. phosphatis* and identified in the current study was lower than that obtained in the only previous comprehensive community proteogenomic study, which focused on an acid mine drainage (AMD) biofilm dominated by *Leptospirillum* group II (Ram *et al.*, 2005). However, wastewater microbial communities are more complex than AMD biofilms, and the composite *A. phosphatis* genome (5842 ORFs) is about double the size of the *Leptospirillum* group II genome (2862 ORFs). In addition, in the current study, the two genomic datasets and the proteomics samples came from wastewater reactors from three different continents (in the AMD study, the proteomics samples and genomic datasets came from the same general site). Furthermore, no distinct cellular fractions were prepared (the AMD study included extensive cellular fractionation). Despite these additional challenges, the results were comprehensive enough to provide new insights into the biochemistry of EBPR. Importantly, such insights may provide clues to optimize consortia and maintain stable EBPR operation. For example, the apparent importance of fatty acid synthesis and degradation may provide an *A. phosphatis* enrichment strategy based on pre-fermentation of exogenous fatty acids prior to the anaerobic/aerobic sludge cycling. This fermentation process could be engineered to produce volatile fatty acids with longer chain lengths than acetate. These in turn would require less

energy expenditure for the accumulation of intracellular macromolecules (polyhydroxyalkanoate and fatty acids) and provide additional selective pressure for PAOs. Indeed, propionate has been found to be a more favourable substrate for EBPR compared to acetate in laboratory-scale reactors (Oehmen *et al.*, 2005). Furthermore, the apparent ability of the *A. phosphatis* population to denitrify may lead to the addition of anoxic zones to current EBPR biological wastewater treatment plants that solely employ anaerobic/aerobic phases. Again, this should lead to an enrichment of the *A. phosphatis* population and result in better phosphate removal performance.

The present study demonstrates the power of proteogenomic methodologies when applied to complex microbial consortia that are of fundamental biotechnological and environmental importance. The results reveal an intriguing level of fine-scale genetic variation among closely related organisms and, consequently, highlight the likely importance of strain variation for overall community homeostasis.

Acknowledgements

We thank Phil Hugenholtz, Héctor García Martín and Ernest Szeto for provision of the genomic information. Furthermore, we thank Vincent Deneff, Greg Dick and Sheri Simmons for helpful comments and discussions. Funding was provided by the United States Department of Energy Genomics:GTL Program (Office of Science). AA was supported by a grant from the Swedish Research Council.

References

- Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. (2005). Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.* **12**: 281–290.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.
- Bell T, Newman JA, Silverman BW, Turner SL, Lilley AK. (2005). The contribution of species richness and composition to bacterial services. *Nature* **436**: 1157–1160.
- Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* **57**: 289–300.
- Blackall LL, Crocetti GR, Saunders AM, Bond PL. (2002). A review and update of the microbiology of enhanced biological phosphorus removal in wastewater treatment plants. *Antonie van Leeuwenhoek* **81**: 681–691.
- Bond PL, Hugenholtz P, Keller J, Blackall LL. (1995). Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactor. *Appl Environ Microbiol* **61**: 1910–1916.
- Correll DL. (1998). The role of phosphorus in the eutrophication of receiving waters: a review. *J Environ Qual* **27**: 261–266.
- Crocetti GR, Banfield JF, Keller J, Bond PL, Blackall LL. (2002). Glycogen-accumulating organisms in laboratory-scale and full-scale wastewater treatment process. *Microbiology* **148**: 3353–3364.
- Crocetti GR, Hugenholtz P, Bond PL, Schuler A, Keller J, Jenkins D *et al.* (2000). Identification of polyphosphate-accumulating organisms and design of 16S rRNA-directed probes for their detection and quantification. *Appl Environ Microbiol* **66**: 1175–1182.
- Deneff VJ, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF. (2007). Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *J Proteome Res* **6**: 3152–3161.
- Eng JK, McCormack AL, Yates III JR. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**: 976–989.
- Eng S-E, Mann M. (2005). Mass spectrometry-based proteomics turn quantitative. *Nat Chem Biol* **1**: 252–262.
- Engelhardt KAM, Ritchie ME. (2002). The effect of aquatic plant species richness on wetland ecosystem processes. *Ecology* **83**: 2911–2924.
- Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL *et al.* (2006). Analysing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**: 303–311.
- García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.
- He S, Gall DL, McMahon KD. (2007). "Candidatus Accumulibacter" population structure in enhanced biological phosphorus removal sludges as revealed by polyphosphate kinase genes. *Appl Environ Microbiol* **73**: 5865–5874.
- Hesketh AR, Chandra G, Shaw AD, Rowland JJ, Kell DB, Bibb MJ *et al.* (2002). Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*. *Mol Microbiol* **46**: 917–932.
- Hesselmann RPX, Von Rummell R, Resnick SM, Hany R, Zehnder AJB. (2000). Anaerobic metabolism of bacteria performing enhanced biological phosphate removal. *Water Res* **34**: 3487–3494.
- Hesselmann RPX, Werlen C, Hahn D, van der Meer JR, Zehnder AJB. (1999). Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced biological phosphate removal in activated sludge. *Syst Appl Microbiol* **22**: 454–465.
- Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1739.
- Kohonen T. (1990). The self-organizing map. *Proc IEEE* **78**: 1464–1480.
- Kong Y, Nielsen JL, Nielsen PH. (2004). Microautoradiographic study of *Rhodocyclus*-related polyphosphate-accumulating bacteria in full-scale enhanced biological

- cal phosphorus removal plants. *Appl Environ Microbiol* **70**: 5383–5390.
- Kong Y, Xia Y, Nielsen JL, Nielsen PH. (2005). Ecophysiology of a group of uncultured gammaproteobacterial glycogen-accumulating organisms in full-scale enhanced biological phosphorus removal wastewater treatment plants. *Environ Microbiol* **8**: 479–489.
- Kornberg A, Rao NN, Ault-Riche D. (1999). Inorganic polyphosphate: a molecule of many functions. *Annu Rev of Biochem* **68**: 89–125.
- Lo I, Deneff VJ, VerBerkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al.* (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537–541.
- Louie TM, Mah TJ, Oldham W, Ramey WD. (2000). Use of metabolic inhibitors and gas chromatography/mass spectrometry to study poly- β -hydroxyalkanoates metabolism involving cryptic nutrients in enhanced biological phosphorus removal systems. *Water Res* **34**: 1507–1514.
- Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.
- Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A *et al.* (2006). The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**: D344–D348.
- McHardy AC, García Martín H, Tsirigos A, Hugenholtz P, Rigoutsos I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**: 63–72.
- McMahon KD, Dojka MA, Pace NR, Jenkins D, Keasling JD. (2002). Polyphosphate kinase from activated sludge performing enhanced biological phosphorus removal. *Appl Environ Microbiol* **68**: 4971–4978.
- Mino T, van Loosdrecht MCM, Heijnen JJ. (1998). Microbiology and biochemistry of the enhanced biological phosphate removal process. *Water Res* **32**: 3193–3207.
- Oehmen A, Yuan Z, Blackall LL, Keller J. (2005). Comparison of acetate and propionate uptake by polyphosphate accumulating organisms and glycogen accumulating organisms. *Biotechnol Bioeng* **91**: 162–168.
- Pereira H, Lemos PC, Reis MAM, Crespo JPSG, Carrondo MJT, Santos H. (1996). Model for carbon metabolism in biological phosphorus removal processes based on *in vivo* ^{13}C -NMR labelling experiments. *Water Res* **30**: 2128–2138.
- Ram RJ, VerBerkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake II RC *et al.* (2005). Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–1920.
- Richardson DJ, Berks BC, Russell DA, Spiro S, Taylor CJ. (2001). Functional, biochemical and genetic diversity of prokaryotic nitrate reductases. *Cell Mol Life Sci* **58**: 165–178.
- Schramm A, De Beer D, Van den Heuvel JC, Ottengraf S, Amann R. (1999). Microscale distribution of populations and activities of *Nitrosospira* and *Nitrospira* spp. along a macroscale gradient in a nitrifying bioreactor: quantification by *in situ* hybridization and the use of microensors. *Appl Environ Microbiol* **65**: 3690–3696.
- Schuler AJ, Jenkins D. (2003). 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.
- Seviour RJ, Mino T, Onuki M. (2003). The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* **27**: 99–127.
- Srinath EG, Sastry CA, Pillai SC. (1959). Rapid removal of phosphorus from sewage by activated sludge. *Experientia* **15**: 339–340.
- Steen H, Mann M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**: 699–711.
- Tabb DL, McDonald WH, Yates III JR. (2002). DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1**: 21–26.
- Tyers M, Mann M. (2003). From genomics to proteomics. *Nat Biotechnol* **22**: 193–197.
- Ultsch A, Moerchen F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical Report of the Department of Mathematics and Computer Science, University of Marburg, Germany* **46**: 1–7.
- United Nations World Water Assessment Programme (2006). *The United Nations World Water Development Report 2*. United Nations Educational, Scientific and Cultural Organization (UNESCO) and Berghahn Books: Paris and New York, France and USA.
- White D. (2007). *The Physiology and Biochemistry Of Prokaryotes*. Oxford University Press: New York, USA.
- Wilmes P, Bond PL. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol* **6**: 911–920.
- Wilmes P, Bond PL. (2006a). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* **14**: 92–97.
- Wilmes P, Bond PL. (2006b). Towards exposure of elusive metabolic mixed-culture processes: the application of metaproteomic analyses to activated sludge. *Water Sci Technol* **54**: 217–226.
- Wilmes P, Wexler M, Bond PL. (2008). Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS ONE* **3**: 1–11 e1778. doi:10.1371/journal.pone.0001778.
- Zeng RJ, Saunders AM, Yuan Z, Blackall LL, Keller J. (2003). Identification and comparison of aerobic and denitrifying polyphosphate-accumulating organisms. *Biotechnol Bioeng* **83**: 140–148.

Supplementary Information accompanies the paper on The ISME Journal website (<http://www.nature.com/ismej>)