TbasCO: Trait-based Comparative 'Omics Identifies Ecosystem-Level and Niche-Differentiating Adaptations of an Engineered Microbiome

McDaniel, E.A.[1,2#+*], van Steenbrugge, J.J.M[3,4,5#+], Noguera, D.R.[6], McMahon, K.D.[1,6], Raaijmakers, J.M.[4,7], Medema, M.H.[3,7], Oyserman, B.O.[3,4+]

[1] Department of Bacteriology, University of Wisconsin – Madison, Madison, WI, USA

[2] Microbiology Doctoral Training Program, University of Wisconsin - Madison, Madison, WI, USA

[3] Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands

[4] Microbial Ecology, Netherlands Institute of Ecological Research, Wageningen, The Netherlands

[5] Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands

[6] Department of Civil and Environmental Engineering, University of Wisconsin – Madison, Madison WI USA

[7]Institute of Biology, Leiden University, Leiden, Netherlands

[#] = contributed equally

[+] = corresponding author

* Current address:

Department of Microbiology and Immunology, University of British Columbia, Vancouver, CA

Corresponding authors:

Elizabeth McDaniel elizabethmcd93@gmail.com

Joris van Steenbrugge jorisvansteebrugge@gmail.com

Ben Oyserman BenOyserman@gmail.com

ABSTRACT

A grand challenge in microbial ecology is disentangling the traits of individual populations within complex communities. Various cultivation-independent approaches have been used to infer traits based on the presence of marker genes. However, marker genes are not linked to traits with complete fidelity, nor do they capture important attributes, such as the timing of expression or coordination among traits. To address this, we present an approach for assessing the trait landscape of microbial communities by statistically defining a trait attribute as shared transcriptional pattern across multiple organisms. Leveraging the KEGG pathway database as a trait library and the Enhanced Biological Phosphorus Removal (EBPR) model microbial ecosystem, we demonstrate that a majority (65%) of traits present in 10 or more genomes have niche-differentiating expression attributes. For example, while 14 genomes containing the high-affinity phosphorus transporter *pstABCS* display a canonical attribute (e.g. up-regulation under phosphorus starvation), we identified another attribute shared by 11 genomes where transcription was highest under high phosphorus conditions. Taken together, we provide a novel framework for revealing hidden metabolic versatility when investigating genomic data alone by assigning trait-attributes through genome-resolved time-series metatranscriptomics.

INTRODUCTION

A longstanding cornerstone of deterministic ecological theory is that the environment selects for traits. Traits may be defined as any physiological, morphological, or genomic signature that affects the fitness or function of an individual [1]. Trait-based approaches have become indispensable in macroecological systems to describe fitness trade-offs and the effects of biodiversity on ecosystem functioning [2–5]. Recently, trait-based frameworks have been proposed as an alternative to taxonomy-based methods for describing microbial ecosystem processes [6, 7]. Connecting microbial traits and their phylogenetic distributions to ecosystem performance can provide powerful insights into the ecological and evolutionary dynamics underpinning community assembly, microbial biogeography, and organismal responses to changes in the environment [8–10]. Additionally, pinpointing the organismal distribution of traits and the selective pressures that enrich them may be leveraged to reproducibly and rationally engineer stable, functionally redundant ecosystems [11–15]. However, applying trait-based approaches to microbial communities is challenging due to the difficulty in identifying and measuring relevant ecological traits for a given ecosystem [16].

High-throughput sequencing technologies and multi-omics techniques have been used to describe the diversity, activity, and functional potential of uncultivated microbial lineages [17–20]. Improvements in bioinformatics algorithms, and in particular metagenomic binning methods, have allowed for genome-resolved investigations of microbial communities rather than gene-based analyses of assembled contigs [21]. These (meta) genomes are subsequently leveraged to detect the presence of key genes or pathways and predict specific traits of the whole community [22, 23]. Integrating metatranscriptomics data addresses a key limitation, as expression patterns better reflect the actual functional dynamics of a trait compared to gene presence alone. Here, we present TbasCO, a software package and statistical framework for *T*rait-*bas*ed *C*omparative ‘*O*mics to identify expression attributes. We adopt the terminology *attribute* as a hierarchically structured feature of a trait and assert that statistically similar transcriptional patterns of traits across multiple organisms be treated as *attributes* (Figure 1)*.* In this manner, the identification of expression-based *attributes* provides a high-throughput and intuitive framework for extending trait-based methods to time-series expression patterns in microbial communities. We implement this trait-based approach to classify transcriptional attributes in a microbial community performing Enhanced Biological Phosphorus Removal (EBPR), a globally important biotechnological process implemented in numerous wastewater treatment plants (WWTPs).

The fundamental feature of the engineered EBPR ecosystem is the decoupled and cyclic availability of an external carbon source and terminal electron acceptor. This cycling is often referred to as “feast-famine” conditions and provides a strong selective pressure for traits such as polymer cycling. Accumulation of intracellular polyphosphate through cyclic anaerobic-aerobic conditions ultimately results in net phosphorus removal and accomplishes the EBPR process [24, 25]. One of the most well-studied polyphosphate accumulating organisms (PAOs) belongs to the uncultivated bacterial lineage ‘*Candidatus* Accumulibacter phosphatis’ (hereby referred to as Accumulibacter) [24, 26]. Numerous genome-resolved ‘omics methods have been used to investigate the physiology and regulation of this model PAO enriched in engineered lab-scale enrichment bioreactor systems [27–34]. However, novel and putative PAOs have been discovered that remove phosphorus without exhibiting the hallmark traits of Accumulibacter [35–39]. Additionally, although these lab-scale systems are designed to specifically enrich for Accumulibacter, a diverse "flanking community" persists in these environments [27], and their ecological roles have largely remained unexplored. As a result, the general adaptations of microbial lineages inhabiting the EBPR community are not well understood. Using genome-resolved metagenomics and metatranscriptomics, we assembled 66 species-representative genomes spanning several significant EBPR lineages and identified the distribution of expression-based attributes. Using our novel trait-based comparative 'omics approach, we show that while some expression attributes are distributed in few genomes, many are redundant and shared across many lineages. Furthermore, we find that a majority of core traits (as defined by the presence of marker genes) have multiple attributes, suggesting that identifying niche-differentiating expression attributes may be used to reveal a large hidden metabolic versatility when investigating genomic data alone.

MATERIALS AND METHODS

Metagenomic Assembly, Annotation, and Metatranscriptomic Mapping

Three metagenomes sampled from an EBPR bioreactor with linked time-series metatranscriptomics data [40] were collected for metagenomic sequencing and assembled into 66 species-representative bins as described in the Supplemental Methods. All bins are greater than 75% complete and contain less than 10% contamination, with a large majority (44/66) >95% complete and <5% redundant as calculated by CheckM [41] (Table 1). Each bin was functionally annotated using the

KEGG database through an HMM-based approach under KEGG release 93.0 using the command-line KofamKOALA pipeline [42, 43], selecting annotations that were significant hits above the specific HMM threshold. This resulted in 117,657 total annotations with 5,228 unique annotations. We used a metatranscriptomic dataset of six timepoints collected over a single EBPR cycle from Oyserman et al. 2016 [40], with three timepoints from the anaerobic phase and three from the aerobic phase. Raw reads were quality filtered using BBtools suite v38.07 [44] and ribosomal rRNA was removed from each sample using SortMeRNA [45]. Reads from each sample were mapped against the concatenated set of open reading frames from all 66 bins using kallisto v0.44.0 and parsed using the R package tximport [46, 47].

TbasCO Method Implementation

The TbasCO package identifies expression-based attributes of predefined traits using time-series (meta)transcriptomics data (Figure 1). In general, traits are defined by the presence of a pathway or other collection of genes from an externally provided database. A weighted distance metric between expression patterns for all genes that define a trait is calculated, and statistically significant similarity is determined based on the background distribution of a trait of equal size. Thereby, two or more organisms with a statistically similar expression pattern for a trait share an *attribute.*

Input and Preprocessing

The input that is accepted by TbasCO is a table of RNAseq counts in csv format. Each row is treated as gene that has columns for the gene/locus name, counts per sample, the genome the gene belongs to, and the KEGG Orthology (KO) identifier. The RNAseq counts table may be provided pre-normalized or can be normalized by the program. The default normalization method is designed to minimize compositional bias in the differential abundance and activity of constituent populations in metatranscriptomics studies. Raw RNA expression counts are therefore normalized by genomic bin and sample [40]. These normalization factors are then applied to each sample for each bin individually. Alternatively, custom normalization methods may be implemented. After normalization, a pruning step is introduced to filter genes that have zero counts or a mean absolute deviation of less than one. To make inter-organismal comparisons of the relative contribution of a gene to total measured organismal RNA, an additional statistic is calculated ranking the expression counts from each sample from highest to lowest. The ranks for each sample are then normalized by dividing them by the maximum rank value in that sample. This normalization is applied to make ranks comparable between organisms with different genome sizes.

To assess the statistical significance of the calculated distances between the expression patterns of all genes within a trait, random background distributions are created for 1) individual genes and 2) traits of N genes. For individual genes, three different distributions were calculated, based on the distances between randomly sampled open reading frames, randomly sampled genes with an annotation (but not necessarily the same annotation), and randomly sampled genes with the same annotation. The background distribution for a trait of N genes is based on the distances between randomly composed sets of genes. For each gene pair, two distances metrics are calculated, the Pearson Correlation (PC) and the Normalized Rank Euclidean Distance (NRED). In practice, it is often found that a certain annotation is assigned to multiple genes in the same genome. If this occurs, there is an option to use either a random selection, or the highest scoring pair. In the latter case, a correction for multiple testing is implemented. This process is repeated N-times, where N equals the number of genes in any given trait. The background distribution for traits is determined by first randomly sampling two genomes, identifying the overlap in annotations, and finally artificially defining a trait containing N annotations. For each annotation in the trait, the distances are calculated between genome A and genome B, as described in the previous section. As modules vary in size, this process is repeated for traits of different sizes.

Identifying Attributes

TbasCO provides both a cluster-based and pair-wise approach to identify attributes. In both methods, the distance between expression patterns of a trait between two genomes is first calculated based on a composite Z score of the PC and NRED for each gene composing the trait. In the cluster-based analysis, the distances are subsequently clustered using the Louvain clustering algorithm to identify trait attributes. To determine if an attribute is significantly similar or not, a one-sided T-test between the attribute and the random background distribution of traits is conducted. This is done for both cluster-based and model-based comparisons. Many traits are complex and represented in databases such as KEGG by numerous alternative routes. To deal with this complexity, each pathway is expanded into the Disjunctive Normative Form (DNF). Due to the extremely high number of DNFs for some traits, attributes are pruned based on a strict requirement of 100% completion.

Distance Calculations

To determine the similarity in expression patterns between genes, two distance metrics are calculated: the PC between RNAseq counts across samples, and the NRED, where ranks are a measure of relative abundance of RNA in each sample, normalized the abundance of RNA in the corresponding genome. These distance scores are converted to Z scores using a background distribution of distances between randomly sampled genes as previously described. To determine statistically significant similarities between the expression patterns of a trait between two genomes, a composite distance score is calculated based on the distance between genes in two different genomes. For each of these genes the PC and NRED are calculated and transformed to Z scores and combined as (-1*PC + NRED). The distance of the trait between two genomes is defined as the average of these composite distance scores, and then normalized by the Jaccard distance between these genomes.

$$(-PC\ + NRED) * (1 - dJ)$$

Statistical Assessment of Trait Attributes

In both model-based and pair-wise approaches, the distance is first calculated between expression patterns of a trait between two genomes based on the composite Z score of the PC and NRED for each gene composing the trait. In the clustering-based analysis, the distances are subsequently clustered using the Louvain clustering algorithm to identify trait-attributes. To determine if attributes are significantly similar, a one-sided T-test is conducted between the attribute and a background distribution of randomly sampled traits with the same number of genes. To derive the random background distributions, multiple distributions are calculated ranging in gene numbers from the smallest trait to the largest trait in the dataset as described previously. For each background distribution, N (default: 10,000) traits are randomly composed. The distances between these artificial traits are calculated in the same way as for the actual traits. In addition to a statistical pruning step, the attributes are pruned based on a strict requirement of 100% completion of each DNF module. A benchmarking analysis to examine the effects of different parameters was conducted to determine their influence on the number of attributes identified and may be found in the supplementary materials (Supplementary Table 1, Supplementary Figures 2-4).

Data and Code Availability

All supplementary files and figures including functional annotations and transcriptome count files are available at https://figshare.com/projects/EBPR_Trait-Based_Comparative_Omics/90437. All 64 flanking genomes have been deposited in NCBI at Bioproject PRJNA714686. The remaining two reassembled Accumulibacter genomes have not been deposited in NCBI to not confuse between the original CAPIA and CAPIIA assemblies [27, 28]. These contemporary assemblies are available at the Figshare repository. The three metagenomes and six metatranscriptomes used in this study are available on the JGI/IMG at accession codes 3300026302, 3300026286, 3300009517, and 3300002341-46, respectively. All code for performing metagenomic assembly, binning, and annotation can be found at https://github.com/elizabethmcd/EBPR-MAGs. The TbasCO method has been implemented as a reproducible R package and can be accessed at https://github.com/Jorisvansteenbrugge/TbasCO.

RESULTS AND DISCUSSION

Reconstructing a Diverse EBPR SBR Community

To explore trait-based transcriptional dynamics of a semi-complex microbial community, we applied genome-resolved metagenomics and metatranscriptomics to an EBPR sequencing-batch reactor (SBR) ecosystem (Figure 2). We previously performed a metatranscriptomics time-series experiment over the course of a normally operating EBPR cycle to investigate the regulatory controls of Accumulibacter gene expression [40]. In this experiment, six samples were collected for RNA sequencing: three from the anaerobic phase and three from the aerobic phase (Figure 2A). Additionally, three metagenomes were collected from the same month of the metatranscriptomic experiment, including a sample from the same date of the experiment. We reassembled contemporary Accumulibacter clade IIA and IA genomes that were previously assembled from the same bioreactor system [27, 28]. The genomes of Accumulibacter clades IA and IIA are similar by approximately 85% average-nucleotide identity [28], and although this is well below the common species-resolved cutoff of 95% [48], we refer to the clade nomenclature defined based on polyphosphate kinase (*ppk1*) sequence identity [49, 50]. During the experiment, the bioreactor was highly enriched in Accumulibacter clade IIA, accounting for approximately 50% of the mapped metagenomic reads and the highest transcriptional counts (Figures 2B and 2C) [40]. Whereas Accumulibacter clade IA exhibited low abundance patterns but was within the top 10 genomes with the highest total transcriptional counts (Figure 2C).

Although this bioreactor system was highly enriched in Accumulibacter, a diverse flanking community persisted and was active in this ecosystem (Figure 2B, C). We reconstructed representative population genomes of the microbial community of the SBR system, resulting in 64 metagenome-assembled genomes (MAGs) of the flanking community. Interestingly, we recovered genomes of experimentally verified and putative PAOs, including two *Tetrasphaera spp.* (TET1 and TET2) '*Candidatus Obscuribacter phosphatis'* (OBS1)*,* and *Gemmatimonadetes* (GEMMA1). Pure cultures of *Tetrasphaera* have been experimentally shown to cycle polyphosphate without incorporating PHA [36], deviating from the hallmark Accumulibacter PAO model. The first cultured representative of the *Gemmatimonadetes* phylum *Gemmatimonas aurantiaca* was isolated from an SBR simulating EBPR and was shown to accumulate polyphosphate through Neisser and DAPI staining [51]. Additionally, *Ca. Obscuribacter phosphatis* has been hypothesized to cycle phosphorus based on the presence of genes for phosphorus transport, polyphosphate incorporation, and potential for both anaerobic and aerobic respiration [37], and has also been enriched in photobioreactor EBPR systems [52]. Both *Tetrasphaera spp.* TET1 and TET2, OBS1, and GEMMA1 groups exhibit higher relative abundance patterns than CAPIA but have similar relative transcriptional levels (Figure 2B and 2C, Table 1).

Numerous SBR MAGs among the *Actinobacteria* and *Proteobacteria* contain the metabolic potential for phosphorus cycling based on the presence of the high-affinity phosphorus transporter *pstABCS* system, polyphosphate kinase *ppk1*, and the low-affinity *pit* phosphorus transporter (Supplementary Figure 5). Additionally, select MAGs within the *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria* contain all required subunits for polyhydroxyalkanoate synthesis (Supplementary Figure 5). Other abundant and transcriptionally active groups in the SBR ecosystem that are not predicted to be PAOs are members of the *Bacteroidetes* such as CHIT1 within the *Chitinophagaceae,* and *Cytophagales* members *Runella* sp*.* RUN1 and *Leadbetterella* sp. LEAD1 (Figure 2B and 2C, Table 1). Interestingly, an uncharacterized group within the *Bacteroidetes* BAC1 contributed the third most to the pool of transcripts (Figure 2C), and did not show phylogenetic similarity to MAGs assembled from Danish full-scale wastewater treatment systems [39] (Supplementary Figure 1). Other groups from which we assembled MAGs for that do not exhibit clear roles in EBPR systems were *Chloroflexi* ANAER1 and HERP1 MAGs, *Armatimonadetes* FIMBRI1, *Firmicutes* FUSI1, and *Patescibacteria* SACCH1. Members of the *Chloroflexi* are filamentous bacteria that have been associated with bulking and foaming events in full-scale WWTPS [53–55], but also aid in forming the scaffolding around floc aggregates and degrade complex polymers [55–57]. The *Patescibacteria* (formerly TM7) are widespread but low abundant members of natural and engineered ecosystems, contain reduced genome sizes, and may contribute to filamentous bulking in activated sludge [21, 58]. To summarize, lab-scale SBRs designed to enrich for Accumulibacter contain diverse flanking community members [27, 32], but their ecological functions and putative interactions remain to be fully understood in the context of the EBPR ecosystem.

Identifying Expression-Based Trait Attributes Among the EBPR SBR Community with TbasCO

Current metatranscriptomics approaches often employ either a gene-centric [31, 59–61] or genome-centric approaches [40, 62–64]. In both approaches, highly, differentially, or co-expressed genes are identified and tested for enrichment of specific functions. Enrichment- or annotation-based approaches are employed in numerous metatranscriptomics tools such as MG-RAST, MetaTrans, SAMSA2, COMAN, IMP, and Anvi'o [65–70]. Here, we expand on the use of molecular markers as traits by defining expression attributes by leveraging *a priori* knowledge from predefined trait libraries, such as the KEGG database [71], to statistically assess inter-species expression patterns of genes that together form a trait (Figure 1). First, our results showed that there is statistically significant transcriptional conservation of genes at the community level; genes that share an annotation were significantly more similar than expected using two different distance metrics (NRED: p-value < 2.2e-16, PC: p-value < 2.2e-16). Extending this statistical analysis to the trait level, we identified 1674 attributes distributed across the 66 genomes. On average, we identified 9.12 genomes per attribute (SD - 5.22), with a minimum of 3 genomes and a maximum of 35 (Figure 3A). Based on these statistics, we defined redundant attributes as those two standard deviations above the mean (19 genomes). With this cutoff applied, we identified 79 redundant trait attributes mostly belonging to pathways among carbohydrate metabolism, purine metabolism, and fatty acid metabolism categories (Table 2). Of 290 traits, we identified 97 traits with two or more attributes identified (33%). Of these, traits in 10 or more genomes were twice as likely to have two or more attributes (65%), suggesting that divergent expression patterns for a trait are common, and may represent a niche-differentiating feature (Figure 3A). Henceforth, when multiple attributes are identified for a trait, we refer to these as niche-differentiating attributes.

From the ecosystem perspective, a clear phylogenetic signal is observed in the distribution of attributes, as genomes cluster together by shared trait attributes by phylum with some exceptions, such as genomes belonging to the *Bacteroidetes, Actinobacteria,* and *Proteobacteria* clustering together, respectively (Figure 3B). For simplicity, we filtered the network to only include nodes with more than 5 connections. Highly redundant trait attributes belonged to modules in the lipid metabolism, energy metabolism, and nucleotide metabolism KEGG functional categories. In contrast, more specialized trait attributes on the periphery of the network or amongst group-specific clusters such as within the *Actinobacteria* or subsets of the *Proteobacteria* belonged to amino acid metabolism, biosynthesis of terpenoids and polyketides, metabolism of cofactors and vitamins, and carbohydrate metabolism KEGG modules. Pathways of note that showed a high level of redundancy include the TCA cycle, isoleucine biosynthesis, acyl-CoA synthesis, threonine biosynthesis, and cytochrome c oxidase activity (Table 2). Large pathways with hundreds of possible routes such as glycolysis, the TCA cycle, gluconeogenesis, and the pentose phosphate pathway are not included in the main network and are displayed as individual networks (Supplementary Figure 6).

We next explored the distribution of non-redundant attributes (e.g. 3-18 genomes) (Figure 3A). A total of 796 trait attributes with low redundancy were identified belonging to pathways involved in carbohydrate cofactor and vitamin metabolism including glycolysis, gluconeogenesis, parts of the TCA cycle, tetrahydrofolate biosynthesis, tryptophan biosynthesis, and the pentose phosphate pathway (Table 3). Different sets of low redundancy trait attributes were identified within respective phyla (Supplementary Figure 7). Between genomes belonging to the *Actinobacteria*, *Alphaproteobacteria, Bacteroidetes, Betaproteobacteria,* and *Gammaproteobacteria,* low redundancy attributes (belonging to less than half of the total genomes within the phylum) include carbohydrate metabolism, amino acid metabolism and metabolism of cofactors and vitamins (Supplementary Figure 7). Redundant trait attributes within individual phyla belong to core energy metabolism pathways, fatty acid biosynthesis, and carbohydrate metabolism. However, even within individual phyla, non-redundant attributes include different amino acids and cofactors (Extended Table 1 - available on Figshare https://figshare.com/articles/dataset/Lineage-Specific_Core_and_Niche_Differentiating_Traits/15001200).

As noted previously, one of the most striking findings is that a majority, 65% of traits present in 10 or more genomes have multiple expression attributes. Thus, it seems that while the presence of marker genes suggests many organisms share a particular trait, the presence of niche-differentiating expression profiles suggest an alternative story, that there is a level of hidden metabolic diversity. For example, central carbon metabolism and energy pathways such as the TCA cycle, glycolysis, gluconeogenesis, and the pentose phosphate pathway are oftentimes considered core traits when only analyzing the presence and/or absence of individual markers belonging to these pathways. Among over 1000 high-quality MAGs assembled from full-scale Danish WWTPs, the TCA cycle and pentose phosphate pathway are highly represented among the abundant microorganisms, with glycolysis less so [39]. Whereas the TCA cycle and pentose phosphate pathway are present among a high number of genomes in the EBPR SBR community, different routes or parts of these pathways have niche-differentiating distributions (Supplementary Figure 4, Tables 2 and 3). These finer-scale differences in expression of “core” traits may explain the persistence of a diverse community when solely fed acetate, as different lineages could employ similar carbon utilization pathways differently or in more versatile ways. Another salient aspect of this analysis is the astonishingly high number of possible routes within individual pathways here represented by their Disjunctive Normal Forms. For example, accounting for all alternative routes and enzymes, the glycolysis pathway has 100s of possible routes. Layering upon this many expression attributes reveals a large hidden metabolic versatility.

Dimensionality of the High-Affinity Phosphorus Transporter System *PstABCS*

The EBPR ecosystem is characterized by its highly dynamic phosphorus cycles. To explore how different lineages respond to fluctuating phosphorus concentrations, we explored the expression-based attributes for the KEGG module of the high-affinity phosphorus transporter *pstABCS* (Figure 4). The *pstABCS* system is an ABC-type transporter that strongly binds phosphate under phosphorus-limiting conditions; therefore, it would be expected that the highest expression levels would be at the end of the aerobic cycle [72]. In contrast, we found that expression of the *pstABCS* was characterized by two different trait attributes. In the first attribute shared by 14 community members, all components of *pstABCS* displayed the highest activity towards the end of the aerobic cycle, when phosphorus concentrations were depleted (Figure 4, Attribute 1). Conversely, 11 community members displayed an alternate attribute where the highest activity of *pstABCS* was at the transition from anaerobic to aerobic phases when phosphorus concentrations are highest (Figure 4, Attribute 2).

These results are in agreement with previous results showing that Accumulibacter clade IIC has a canonical *pstABCS* expression pattern (as in Figure 4, Attribute 1) , whereas the Accumulibacter clade IA has a non-canonical expression (as in Figure 4, Attribute 2) [31]. By assigning trait attributes, we are able to extend these findings beyond Accumulibacter to other flanking community members in the SBR ecosystem suggesting that there are conserved ecological pressures driving niche differentiating expression patterns in *pstABCS* within the EBPR community.

Distribution and Expression of Truncated Denitrification Steps Among EPBR Community Members

Understanding the induction of denitrification is an important ecosystem property linked to the redox status of an environment. In EBPR communities, there are many diverse and incomplete denitrification pathways, distributed across many lineages denitrification steps expected in denitrifying systems (Figure 5). Among all 66 MAGs, we did not identify any single MAG with a complete denitrification pathway consisting of the genetic repertoire necessary to fully reduce nitrate to nitrogen gas (Supplementary Figure 5). Instead, we identified multiple groups of organisms with truncated denitrification pathways, with steps distributed among cohorts of community members (Figure 5).

For the first steps of reducing nitrate to nitrite, we explored expression attributes of the *napAB* and *narGH* pathways (Figure 5B, C). For the *narGH* pathway, two attributes were identified (Figure 5B). The first *narGH* attribute was characterized by high expression in the anaerobic phase, with decreasing activity by the second time point of the anaerobic phase. Genomes containing this attribute included the experimentally verified and putative PAOs *Tetrasphaera* (TET1 and TET2) and *Ca.* Obscuribacter (OBS1), respectively. The second attribute was exhibited among members of the *Actinobacteria* (PROP2, PHYC2, PROP3, and NANO1), *Proteobacteria* (BEIJ4), and *Bacteroidetes* (BAC1). The attribute identified for *napAB* was also more highly expressed anaerobically and included CAPIA, CAPIIA, ALIC1, REYR2, RUBRI1, and BEIJ3. Interestingly, this *napAB* attribute had expression patterns that quickly decreased in the first aerobic time point, suggesting a tighter regulation than Attribute 1 for *narGH*. Together, this suggests that the regulation of denitrification within the EBPR ecosystem is a niche-differentiating feature whereby the induction of denitrification pathways occurs either anaerobically or only after anaerobic carbon contact.

A smaller cohort contained the genetic repertoire to reduce nitrite to nitrogen gas and exhibited hallmark anaerobic-aerobic expression patterns (Figure 5E) These members within the *Proteobacteria* (OTTO2, BEIJ3, VITREO1, and ZOO1) contained the *nirS* nitrite reductase, the *norBC* nitric oxide reductase, and *nosZ,* and showed highest expression of these subunits towards the beginning of the anaerobic cycle, slowly decreasing over the aerobic period to their lowest in the end of the aerobic cycle. Although BEIJ2 was lacking the *norBC* system, it contained the *nirS* nitrite reductase and *nosZ* subunit, and exhibited similar expression patterns to others in this cohort. Other *Proteobacteria* lineages only contained the *norBC* subunits but were expressed in similar fashions (RHODO2, FLAVO1, RHIZO1, and LEAD1) (Figure 5D). Accumulibacter clades IA and IIA as well as ALIC1 were the only lineages with near-complete denitrification pathways. These lineages contained the *napAB* nitrate reductase system as mentioned above, the *nirS* nitrite reductase, *norB* (missing a confident hit for the *norC* subunit), and *nosZ.* These three lineages also exhibited hallmark upregulation of all steps in the anaerobic phase, with decreased activity after aerobic contact (Figure 5F).

Interestingly, Accumulibacter clade IA exhibited a higher magnitude of expression of denitrification steps when activity levels were normalized relative to clade IIA, supporting the hypothesis that denitrification is a niche-differentiating feature among clades [28, 31, 73], and possibly a strain-specific trait since denitrification traits cannot be predicted based on *ppk1* clade designations [32]. For example, independent observations in differences among denitrification activities among strains within Accumulibacter clade IC are inconsistent [34, 74]. Within the same bioreactor environment, coexisting Accumulibacter clades differ between denitrification abilities and expression profiles [31, 33, 75]. Truncated denitrification pathways have also been previously shown to be distributed among community members, with the complete denitrification genetic repertoire only present in few members [33, 75], which could be due to extensive horizontal gene transfer of genes comprising denitrification steps [75, 76]. Although this experiment was not conducted under denitrifying conditions, our approach could be applied to denitrifying EBPR systems to further understand the distribution of denitrification traits among community members and how to selectively enrich for diverse DPAOs.

Biosynthetic Potential and Expression Dynamics of Amino Acid and Vitamin Synthesis Pathways

Although SBRs are designed to enrich for Accumulibacter by providing acetate as the sole carbon source, a diverse flanking community persists in these setups [27, 75]. One hypothesis for the persistence of flanking community members may be cooperative interactions due to underlying auxotrophies of amino acid and vitamin biosynthetic pathways in Accumulibacter. Amino acids and vitamin cofactors are metabolically expensive to synthesize, and widespread auxotrophies have been widely documented among microbial communities [77, 78]. Specifically, auxotrophies of vitamin cofactors have been shown to fuel bacterial and cross-kingdom interactions with *de novo* bacterial and cross-kingdom interactions with *de novo* synthesizers [79, 80]. To explore this hypothesis in the EPBR SBR community, we analyzed the presence of amino acid and vitamin biosynthetic pathways and their expression patterns among the top 15 genomes based on transcript abundance (Figure 6).

Within Accumulibacter, there are a few key vitamin cofactor and amino acid auxotrophies that could fuel potential interactions with flanking community members. Both Accumulibacter clade genomes are missing the riboflavin pathway for FAD cofactor synthesis, as well as the pathways for serine and aspartic acid (Figure 6A). The biosynthetic pathway for aspartic acid is distributed among members of the *Bacteroidetes* and *Proteobacteria*, whereas only TET2 contains the pathway for serine synthesis (Figure 5A). The lack of serine biosynthesis pathways in Accumulibacter and other flanking genomes seems striking given that serine is one of the least metabolically costly amino acids to synthesize [81]. Interestingly, Accumulibacter clade IIA does not contain the biosynthetic machinery for thiamine and pantothenate synthesis, whereas clade IA does (Figure 6A). Only the CAULO1, HYPHO1, and PSEUDO1 genomes within the *Proteobacteria* can synthesize thiamine, whereas several other members can synthesize pantothenate (Figure 6A). The absence of the pantothenate biosynthetic pathway in Accumulibacter CAP IIA is particularly interesting given that coenzyme A is essential for polyhydroxyalkanoate biosynthesis, which fuels the polymer cycling PAO phenotype of Accumulibacter [24].

In addition to flanking community members potentially supporting the growth of Accumulibacter due to underlying auxotrophies, the reciprocal logic may be possible as well. Both Accumulibacter clades contain the pathways for synthesizing tyrosine and phenylalanine, which are missing in a majority of the top 15 active flanking genomes

(Figure 6A). Only two other members within the *Proteobacteria* can synthesize tyrosine and phenylalanine, where RAM1 can synthesize both and PSEUDO1 only phenylalanine. Interestingly, phenylalanine and tyrosine are the second and third most metabolically expensive amino acids to synthesize, respectively, with tryptophan the most costly [81]. Additionally, a few highly active flanking community members lack the biosynthetic machinery for several vitamin cofactors and amino acids, such as FLAVO1 and BAC3 within the *Bacteroidetes* and the putative PAO *Ca.* Obscuribacter phosphatis OBS1 (Figure 6A). Particularly, RAM1 within the *Proteobacteria* is missing the biosynthetic machinery for all vitamin cofactors but can synthesize most amino acids including the most metabolically expensive as mentioned above.

We next analyzed the distribution of trait-attributes of vitamin and amino acid pathways among these genomes to understand how these biosynthetic pathways are expressed similarly or differently in the EBPR SBR ecosystem (Figure 6B and C). Members of the *Proteobacteria* containing thiamine and cobalamin biosynthetic pathways all express these traits similarly (Figure 6B). However, the pantothenate synthesis pathway contains two trait-attributes and is expressed differently among two cohorts. In the first attribute, RUN1, TET1, CAULO1, CAPIA, and PSEUDO1 express the pantothenate pathway similarly. However, OBS1 and TET2 express the pantothenate pathway differently (Figure 6B). Because tetrahydrofolate can be synthesized through different metabolic routes, we analyzed the differences in trait attribute expression for all routes in genomes that contained sufficient coverage of this trait. Members of the *Bacteroidetes* and *Proteobacteria* mostly cluster together among tetrahydrofolate attributes, whereas the TET1 and TET2 genomes are differentiated (Figure 6B).

Expression of various groups of amino acids show more differentiated patterns of expression for genomes with these pathways. Several amino acids also contain different metabolic routes for biosynthesis, and we analyzed all trait attributes for each amino acid for all routes grouped by type (Figure 6C). For the charged amino acids arginine, histidine, and lysine, members of the *Proteobacteria* and *Bacteroidetes* cluster within their phylogenetic groups, respectively, with lysine and histidine expressed differently among these groups (Figure 6C). In contrast, arginine is expressed similarly among all *Proteobacteria* genomes. Among the polar charged amino acids, TET2 is the only genome among the top 15 genomes that contains the metabolic pathway to synthesize serine (Figure 6A). Several groups contain the pathway for threonine synthesis, and expression of different threonine routes are differentiated among the *Proteobacteria, Bacteroidetes,* and *Tetrasphaera spp.,* but mostly clusters phylogenetically (Figure 6C). Notably, the expression patterns for the cysteine and proline biosynthetic pathways do not cluster phylogenetically, such as both *Tetrasphaera* genomes expressing the proline pathway more similarly to other *Proteobacteria* and *Bacteroidetes* (Figure 6C). The few lineages that can synthesize tyrosine and phenylalanine (CAPIA, CAPIIA, RAM1, PSEUDO1) show different patterns of expression. These results show that beyond the presence or absence of key vitamin cofactor and amino acid biosynthetic pathways, EBPR SBR organisms also display coherent and differentiated patterns of expression for these traits, of which the functional consequences remain to be further understood.

CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, we applied a novel trait-based ‘omics pipeline to a semi-complex, engineered bioreactor microbial community to explore ecosystem-level and niche-differentiating traits. Through assembling high-quality MAGs of the EBPR SBR community and using a time-series metatranscriptomics experiment, we were able to extend functional predictions and ecosystem inferences beyond hypotheses made from gene presence/absence data. Using our novel trait-based comparative ‘omics pipeline, we identified how similarities and differences in the expression of significant EBPR traits are conferred among community members such as phosphorus cycling, denitrification, and amino acid metabolism. Specifically, we demonstrate that traits with similar expression profiles may be clustered into attributes providing a new layer to trait-based approaches.

We believe that identifying expression-based attributes will be a powerful tool to explore microbial traits in natural, engineered, and host-associated microbiomes. Outside of activated sludge systems, trait-based approaches could illuminate how similar secondary metabolite clusters are expressed among different species in a community [82, 83], how auxotrophies for amino acid and vitamin cofactors govern interactions [84], how rhizosphere microorganisms respond to day-night cycles, and identify putative traits that universally exhibit ecosystem-level or niche-differentiating patterns across ecosystems [19, 23]. Importantly, our trait-based approach can be used to screen for expected expression patterns of a key trait compared to a model organism, and then prioritize specific microbial lineages for downstream experimental verification with techniques such as Raman-FISH [85, 86]. Overall, our trait-based comparative ‘omics pipeline is a novel and high-throughput approach to understand how microbial traits connect to ecosystem-level processes in diverse microbiomes.

ACKNOWLEDGEMENTS

We thank Caitlin Singleton for providing early access to high-quality genomes from a full-scale WWTP to compare our MAGs against. Metagenomic and metatranscriptomic sequencing was provided through a Joint Genome Institute Community Science Proposal (Proposal ID 873). This work was supported by funding from the National Science Foundation (MCB-1518130) to K.D.M and D.R.N. Funding was provided to E.A.M. by a fellowship through the Department of Bacteriology at the University of Wisconsin – Madison. Funding for B.O.O was in part provided by the Technology Foundation of the Dutch National Science Foundation (NWO-TTW). This research was performed in part using the Wisconsin Energy Institute computing cluster, which is supported by the Great Lakes Bioenergy Research Center as a part of the U.S. Department of Energy Office of Science (DE-SC0018409).

REFERENCES CITED

1. Violle C, Navas M-L, Vile D, Kazakou E, Fortunel C, Hummel I, et al. Let the concept of trait be functional! *Oikos* 2007; **116**: 882–892.

2. Lavorel S, Garnier E. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Funct Ecol* 2002; **16**: 545–556.

3. Hooper DU, Chapin FS, Ewel JJ, Hector A, Inchausti P, Lavorel S, et al. EFFECTS OF BIODIVERSITY ON ECOSYSTEM FUNCTIONING: A CONSENSUS OF CURRENT KNOWLEDGE. *Ecol Monogr* 2005; **75**: 3–35.

4. Pianka ER. On r-and K-Selection. *Am Nat* 1970.

5. Wright IJ, Reich PB, Westoby M, Ackerly DD, Baruch Z, Bongers F, et al. The worldwide leaf economics spectrum. *Nature* 2004; **428**: 821–827.

6. Krause S, Le Roux X, Niklaus PA, Van Bodegom PM, Lennon JT, Bertilsson S, et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front Microbiol* 2014; **5**: 251.

7. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD, et al. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *bioRxiv* 2019.

8. Guittar J, Shade A, Litchman E. Trait-based community assembly and succession of the infant gut microbiome. *Nat Commun* 2019; **10**: 512.

9. Wolfe BE, Button JE, Santarelli M, Dutton RJ. Cheese Rind Communities Provide Tractable Systems for In Situ and In Vitro Studies of Microbial Diversity. *Cell* 2014; **158**: 422–433.

10. Enke TN, Datta MS, Schwartzman J, Barrere J, Pascual-García A, Cordero Correspondence OX. Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. *Curr Biol* 2019; **29**.

11. Herrera Paredes S, Gao T, Law TF, Finkel OM, Mucyn T, Teixeira PJPL, et al. Design of synthetic bacterial communities for predictable plant phenotypes. *PLOS Biol* 2018; **16**: e2003962.

12. Lindemann SR, Bernstein HC, Song H-S, Fredrickson JK, Fields MW, Shou W, et al. Engineering microbial consortia for controllable outputs. *ISME J* 2016; **10**: 2077–2084.

13. Oyserman BO, Medema MH, Raaijmakers JM. Road MAPs to engineer host microbiomes. *Curr Opin Microbiol* 2018; **43**: 46–54.

14. Lawson CE, Harcombe WR, Hatzenpichler R, Lindemann SR, Löffler FE, O’Malley MA, et al. Common principles and best practices for engineering microbiomes. *Nat Rev Microbiol* . 2019. Nature Publishing Group. , **17**: 725–741

15. Gutierrez CF, Sanabria J, Raaijmakers JM, Oyserman BO. Restoring degraded microbiome function with self-assembled communities. *FEMS Microbiol Ecol* 2020; **96**.

16. Allison SD. A trait-based approach for modelling microbial litter decomposition. *Ecol Lett* 2012; **15**: 1058–1070.

17. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. *Science* 2005; **308**: 554–7.

18. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial

genomes from the environment. *Nature* 2004; **428**: 37–43.

19. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 2016; **7**: 13219.

20. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, et al. Genome-centric view of carbon processing in thawing permafrost. *Nature* 2018; **560**: 49–54.

21. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013; **31**: 533–538.

22. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 2016; **7**: 13219.

23. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 2012; **337**: 1661–5.

24. Hesselmann RPX, Werlen C, Hahn D, van der Meer JR, Zehnder AJB. Enrichment, Phylogenetic Analysis and Detection of a Bacterium That Performs Enhanced Biological Phosphate Removal in Activated Sludge. *Syst Appl Microbiol* 1999; **22**: 454–465.

25. Seviour RJ, Mino T, Onuki M. The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* 2003; **27**: 99–127.

26. Crocetti GR, Hugenholtz P, Bond PL, Schuler A, Ju¨ J, Keller J, et al. Identification

of Polyphosphate-Accumulating Organisms and Design of 16S rRNA-Directed Probes for Their Detection and Quantitation. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY* . 2000.

27. Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* 2006; **24**: 1263–1269.

28. Flowers JJ, He S, Malfatti S, del Rio TG, Tringe SG, Hugenholtz P, et al. Comparative genomics of two ‘Candidatus Accumulibacter’ clades performing biological phosphorus removal. *ISME J* 2013; **7**: 2301–2314.

29. Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, et al. Ancestral genome reconstruction identifies the evolutionary basis for trait acquisition in polyphosphate accumulating bacteria. *ISME J* 2016; **10**: 2931–2945.

30. Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B, et al. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* 2008; **2**: 853–864.

31. McDaniel E, Moya-Flores F, Keene Beach N, Oyserman B, Kizaric M, Hoe Khor E, et al. Metabolic differentiation of co-occurring Accumulibacter clades revealed through genome-resolved metatranscriptomics. *bioRxiv* 2020; 2020.11.23.394700.

32. Gao H, Mao Y, Zhao X, Liu WT, Zhang T, Wells G. Genome-centric metagenomics resolves microbial diversity and prevalent truncated denitrification

pathways in a denitrifying PAO-enriched bioprocess. *Water Res* 2019; **155**: 275–287.

33. Wang Y, Gao H, Wells G. Integrated Omics Analyses Reveal Differential Gene Expression and Potential for Cooperation Between Denitrifying Polyphosphate and Glycogen Accumulating Organisms. *bioRxiv* . 2020. bioRxiv. , 2020.01.10.901413

34. Camejo PY, Oyserman BO, McMahon KD, Noguera DR. Integrated Omic Analyses Provide Evidence that a 'Candidatus Accumulibacter phosphatis' Strain Performs Denitrification under Microaerobic Conditions. *mSystems* 2019; **4**: e00193-18.

35. Kong Y, Nielsen JL, Nielsen PH. Identity and ecophysiology of uncultured actinobacterial polyphosphate-accumulating organisms in full-scale enhanced biological phosphorus removal plants. *Appl Environ Microbiol* 2005; **71**: 4076–85.

36. Kristiansen R, Nguyen HTT, Saunders AM, Nielsen JL, Wimmer R, Le VQ, et al. A metabolic model for members of the genus Tetrasphaera involved in enhanced biological phosphorus removal. *ISME J* 2013; **7**: 543–54.

37. Soo R, Skennerton CT, Sekiguchi Y, Imelfort M, Paech S, Dennis P, et al. An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4040986/. Accessed 11 Jul 2020.

38. Petriglieri F, Singleton C, Peces M, Petersen JF, Nierychlo M, Nielsen H. 'Candidatus Dechloromonas phosphatis' and 'Candidatus Dechloromonas phosphovora', two novel polyphosphate accumulating organisms abundant in

wastewater treatment systems. *bioRxiv* 2020.

39. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun* 2021; **12**: 2009.

40. Oyserman BO, Noguera DR, del Rio TG, Tringe SG, McMahon KD. Metatranscriptomic insights on gene expression and regulatory controls in Candidatus Accumulibacter phosphatis. *ISME J* 2016; **10**: 810–822.

41. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043–55.

42. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 2016; **428**: 726–731.

43. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *bioRxiv* 2019; 602110.

44. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* 2017; **12**: e0185056.

45. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012; **28**: 3211–3217.

46. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016; **34**: 525–527.

47. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 2015; **4**: 1521.

48. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018; **9**: 5114.

49. He S, Gall DL, McMahon KD. 'Candidatus accumulibacter' population structure in enhanced biological phosphorus removal sludges as revealed by polyphosphate kinase genes. *Appl Environ Microbiol* 2007; **73**: 5865–5874.

50. Camejo PY, Owen BR, Martirano J, Ma J, Kapoor V, Santo Domingo J, et al. Candidatus Accumulibacter phosphatis clades enriched under cyclic anaerobic and microaerobic conditions simultaneously use different electron acceptors. *Water Res* 2016; **102**: 125–137.

51. Zhang H, Sekiguchi Y, Hanada S, Hugenholtz P, Kim H, Kamagata Y, et al. Gemmatimonas aurantiaca gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. *Int J Syst Evol Microbiol* 2003; **53**: 1155–1163.

52. McDaniel EA, Wever R, Oyserman BO, Noguera DR, McMahon KD. Genome-Resolved Metagenomics of a Photosynthetic Bioreactor Performing Biological Nutrient Removal. *Microbiol Resour Announc* 2021; **10**.

53. Speirs LBM, Rice DTF, Petrovski S, Seviour RJ. The Phylogeny, Biodiversity, and Ecology of the Chloroflexi in Activated Sludge. *Front Microbiol* . 2019. Frontiers Media S.A. , **10**: 2015

54. Andersen MH, McIlroy SJ, Nierychlo M, Nielsen PH, Albertsen M. Genomic insights into Candidatus Amarolinea aalborgensis gen. nov., sp. nov., associated with settleability problems in wastewater treatment plants. *Syst Appl Microbiol* 2019; **42**: 77–84.

55. Nierychlo M, Miłobędzka A, Petriglieri F, McIlroy B, Nielsen PH, McIlroy SJ. The morphology and metabolic potential of the Chloroflexi in full-scale activated sludge wastewater treatment plants. *FEMS Microbiol Ecol* 2019; **95**.

56. McIlroy SJ, Karst SM, Nierychlo M, Dueholm MS, Albertsen M, Kirkegaard RH, et al. Genomic and in situ investigations of the novel uncultured Chloroflexi associated with 0092 morphotype filamentous bulking in activated sludge. *ISME J* 2016; **10**: 2223–2234.

57. Kragelund C, Levantesi C, Borger A, Thelen K, Eikelboom D, Tandoi V, et al. Identity, abundance and ecophysiology of filamentous Chloroflexi species present in activated sludge treatment plants. *FEMS Microbiol Ecol* 2007; **59**: 671–682.

58. Kindaichi T, Yamaoka S, Uehara R, Ozaki N, Ohashi A, Albertsen M, et al. Phylogenetic diversity and ecophysiology of Candidate phylum Saccharibacteria in activated sludge. *FEMS Microbiol Ecol* 2016; **92**: 1–11.

59. Mann E, Wetzels SU, Wagner M, Zebeli Q, Schmitz-Esser S. Metatranscriptome Sequencing Reveals Insights into the Gene Expression and Functional Potential of Rumen Wall Bacteria. *Front Microbiol* 2018; **9**: 43.

60. Jiang Y, Xiong X, Danska J, Parkinson J. Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. *Microbiome* 2016; **4**: 2.

61. Linz AM, Aylward FO, Bertilsson S, McMahon KD. Time-series metatranscriptomes reveal conserved patterns between phototrophic and heterotrophic microbes in diverse freshwater systems. *Limnol Oceanogr* 2019.

62. Lawson CE, Wu S, Bhattacharjee AS, Hamilton JJ, McMahon KD, Goel R, et al. Metabolic network analysis reveals microbial community interactions in anammox granules. *Nat Commun* 2017; **8**: 15416.

63. Aylward FO, Eppley JM, Smith JM, Chavez FP, Scholin CA, DeLong EF. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc Natl Acad Sci* 2015; **112**: 5443–5448.

64. Hao L, Michaelsen TY, Singleton CM, Dottorini G, Kirkegaard RH, Albertsen M, et al. Novel syntrophic bacteria in full-scale anaerobic digesters revealed by genome-centric metatranscriptomics. *ISME J* 2020; **14**: 906–918.

65. Glass EM, Meyer F. The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. *Handb Mol Microb Ecol I Metagenomics Complement Approaches* 2011; **8**: 325–331.

66. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, et al. MetaTrans: an open-source pipeline for metatranscriptomics. *Sci Rep* 2016; **6**: 26447.

67. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics* 2018; **19**: 175.

68. Ni Y, Li J, Panagiotou G. COMAN: a web server for comprehensive metatranscriptomics analysis. *BMC Genomics* 2016; **17**: 622.

69. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated

metagenomic and metatranscriptomic analyses. *Genome Biol* 2016; **17**: 260.

70. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015; **3**: e1319.

71. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**: D457–D462.

72. Wanner BL. Gene regulation by phosphate in enteric bacteria. *J Cell Biochem* 1993; **51**: 47–54.

73. Flowers JJ, He S, Yilmaz S, Noguera DR, McMahon KD. Denitrification capabilities of two biological phosphorus removal sludges dominated by different 'Candidatus Accumulibacter' clades. *Environ Microbiol Rep* 2009; **1**: 583–588.

74. Rubio-Rincón FJ, Weissbrodt DG, Lopez-Vazquez CM, Welles L, Abbas B, Albertsen M, et al. 'Candidatus Accumulibacter delftensis': A clade IC novel polyphosphate-accumulating organism without denitrifying activity on nitrate. *Water Res* 2019; **161**: 136–151.

75. Gao H, Mao Y, Zhao X, Liu WT, Zhang T, Wells G. Genome-centric metagenomics resolves microbial diversity and prevalent truncated denitrification pathways in a denitrifying PAO-enriched bioprocess. *Water Res* 2019; **155**: 275–287.

76. Parsons C, Stüeken EE, Rosen CJ, Mateos K, Anderson RE. Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history. *Geobiology* 2020.

77. Gómez-Consarnau L, Sachdeva R, Gifford SM, Cutter LS, Fuhrman JA, Sañudo-Wilhelmy SA, et al. Mosaic patterns of B-vitamin synthesis and utilization in a natural marine microbial community. *Environ Microbiol* 2018; **20**: 2809–2823.

78. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, et al. Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. *mSystems* 2017; **2**: e00091-17.

79. McClure RS, Overall CC, Hill EA, Song H-S, Charania M, Bernstein HC, et al. Species-specific transcriptomic network inference of interspecies interactions. *ISME J* 2018; 1.

80. Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* 2005; **438**: 90–93.

81. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* 2002; **99**: 3695–3700.

82. Lozano GL, Bravo JI, Diago MFG, Park HB, Hurley A, Peterson SB, et al. Introducing THOR, a Model Microbiome for Genetic Dissection of Community Behavior. *MBio* 2019; **10**: e02846-18.

83. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* 2018; **558**: 440–444.

84. Zengler K, Zaramela LS. The social network of microorganisms — how

auxotrophies shape complex communities. *Nat Rev Microbiol* 2018; **16**: 383–390.

85. Fernando EY, McIlroy SJ, Nierychlo M, Herbst FA, Petriglieri F, Schmid MC, et al. Resolving the individual contribution of key microbial populations to enhanced biological phosphorus removal with Raman–FISH. *ISME J* 2019; **13**: 1933–1946.

86. Petriglieri F, Petersen JF, Peces M, Nierychlo M, Hansen K, Baastrand CE, et al. Quantification of biologically and chemically bound phosphorus in activated sludge from full-scale plants with biological P-removal. *bioRxiv* 2021.

87. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2019.

88. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**.

FIGURE AND TABLE LEGENDS

Figure 1. Overview of Trait-based Comparative Transcriptomics Approach

In genome-resolved metagenomics approaches, representative MAGs are assembled from a microbial community of interest, and the presence and/or absence of key metabolic pathways are used to make inferences of metabolic potential and ecosystem processes. However, metagenomic data alone can only assess the metabolic potential of a given pathway, and do not provide other biologically relevant information such as the timing or induction of these traits. Using time-series metatranscriptomics, we developed a trait-based comparative ‘omics (TbasCO) pipeline that statistically assesses the inter-organismal differences in gene expression pattern of a given trait to cluster into trait attributes.

Figure 2. Genome-Resolved Metatranscriptomics Approach of an EBPR System

Application of a genome-resolved metatranscriptomics approach to a lab-scale sequencing batch reactor (SBR) designed to enrich for Accumulibacter. **1A)** Schematic of the main cycle parameters and analyte dynamics of an SBR simulating EBPR. Six samples were taken for RNA sequencing within the cycle at time-points denoted by arrows. **1B)** Phylogenetic identity and abundance patterns of 66 assembled MAGs from the EBPR system. The phylogenetic tree was constructed from concatenated markers contained in the GTDB-tk with muscle, calculated with RAxML, and visualized in iTOL. A phylogenetic tree of all 66 MAGs with reference genomes and high-quality genomes from Singleton et al. constructed with concatenated markers from GTDB-tk are provided in Supplementary Figure 1. Sizes of circles represent abundance patterns of metagenomic reads mapping back to genomes from the same day as the metatranscriptomic experiment and are not to scale. **1C)** Transcriptional patterns of each MAG in the anaerobic and aerobic phases of the EBPR cycle. RNA-seq reads from each time-point were competitively mapped to all 66 assembled MAGs and counts normalized by transcripts per million (TPM). Total counts in the anaerobic and aerobic phases for each genome were averaged separately and plotted on a log scale. Order of MAGs from left to right mirrors the order of MAGs in the phylogenetic tree in 1B from the top of the circle going clockwise.

Figure 3. Clustering and Distribution of Trait Attributes Across EBPR SBR Community Members. Using the TbasCO method, we identified expression-based trait attributes from predefined trait modules in the KEGG library and explored the distribution of these trait attributes across community members. **A)** Distribution of trait-attributes among sets of genomes. Bars represent the number of trait-attributes present in a set number of genomes and colored by KEGG module category. Among a total of 35 genomes, trait attributes present between 3-18 genomes are designated as niche differentiating, whereas trait attributes present in 19 or greater genomes are designated as core trait attributes. Inset figure demonstrates the maximum number of attributes for the maximum number of genomes. **B)** Cytoscape network showing the connectedness of genomes to trait attributes. The network was filtered to only include nodes with more than 5 connections, therefore filtering out both genomes with few trait attributes and trait attributes connected to less than 5 genomes. Genomes are represented as squares colored by phylum, and trait attributes are represented as circles colored by KEGG

category. The size of both the squares and circles represents the number of connections to that genome or trait attribute, respectively.

Figure 4. Trait Attributes of the High-Affinity Phosphorus Transporter System *pstABCS*

Using the TbasCO method, two trait attributes of the high-affinity phosphorus transporter system *pstABCS* were identified. The *pstABCS* system consists of a phosphate-binding protein and ABC-type transporter, and the corresponding KEGG orthologs for each subunit are shown. Timepoints 1-3 refer to the three anaerobic phase timepoints, and timepoints 4-6 refer to the three anaerobic phase timepoints (Figure 1). Expression values are log-transformed based on setting the lowest expression value within each genome across the time-series to 0 for each subunit. Specific subunits for some genomes in both attributes are missing to the high cutoff thresholds for annotations. However we kept genomes with 2/4 subunits to show similarities in expression profiles. The first *pstABCS* trait-attribute includes microbial lineages that exhibited the highest expression of all subunits towards the end of the aerobic cycle, when phosphate concentrations are expected to be lowest. This includes microbial lineages within the *Actinobacteria, Proteobacteria, Gemmatimonadetes,* and *Chloroflexi.* The second *pstABCS* trait-attribute includes lineages that exhibited highest expression of all subunits upon the switch from anaerobic to aerobic phases, or when phosphate concentrations are expected to be the highest. This includes lineages within the *Actinobacteria* and *Proteobacteria.*

Figure 5. Expression Dynamics of Distributed Denitrification Routes

Expression of denitrification traits distributed among community members in the EBPR SBR ecosystem. Timepoints 1-3 correspond to the anaerobic phase and timepoints 4-6 correspond to the aerobic phase as referenced in Figure 1. **A)** Complete denitrification pathway and associated genetic repertoire with each sequential step. **B)** Trait attributes of expression dynamics for community members with the *narGH* nitrate reductase system. This trait was the only denitrification trait identified with more than one attribute. **C)** Expression dynamics of the *napAB* nitrate reductase system. **D)** Expression dynamics of the *norBC* nitrous oxide reductase system. **E)** Expression of all steps of denitrification starting at nitrite reduction. **F)** Expression of the most complete denitrification route among three community members, with the *norC* subunit for nitrous oxide reduction missing. Note that OTTO1 only contains *nirS* but is included in this trait attribute because the expression dynamics are similar to that of the other three genomes for this subunit.

Figure 6. Biosynthetic Potential Compared to Expression of Amino Acid and Vitamin Synthesis Pathways for Top 15 Expressed MAGs

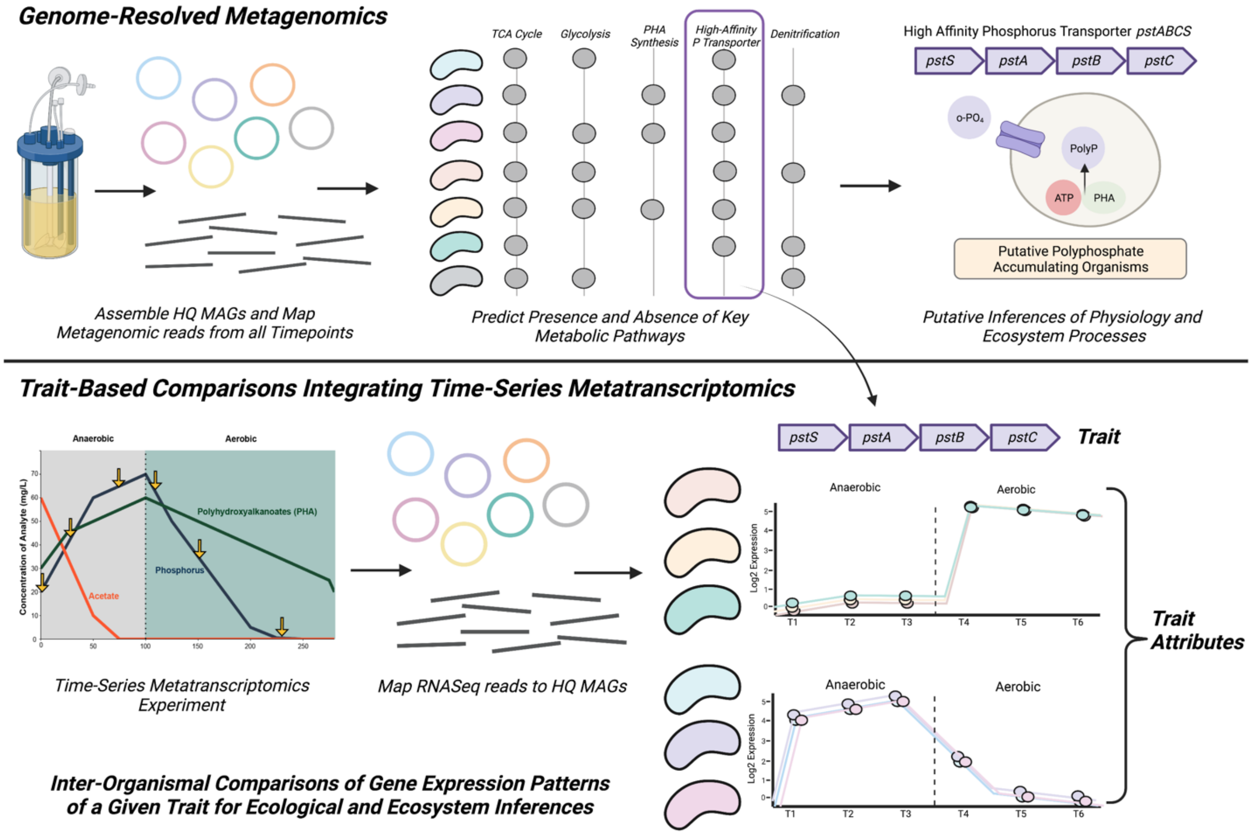
Biosynthetic potential and expression patterns of amino acid and vitamin pathways were analyzed for the top 15 genomes with the highest transcriptional counts (Table 1). **A)** For a pathway to be considered present for downstream analysis in the TbasCO pipeline, 80% of the pathway had to be present in a genome. Thus, we used this cutoff criterion to discern whether a specific pathway was present or absent in a genome (with the expectation of methionine, as all genomes did not contain at least 80% of the subunits in the KEGG methionine synthase pathway, we inferred the presence of the methionine synthase as presence of this pathway). Orange colored boxes for cofactor biosynthesis pathways represents the presence of that pathway, whereas grey infers absence. For amino acid biosynthetic pathways, amino acids are listed by their side chain groups – charged, polar, hydrophobic, and other. **B)** Mini-networks of vitamin co-factors. Squares are genomes with the colors matching the color bar in A. Nodes are attributes, where the colored nodes for the tetrahydrofolate attributes represent the different routes. **C)** Mini-networks of amino acid biosynthesis pathways split by type. Colors of nodes for each amino acid represent the different routes for that pathway. Squares represent genomes with colors matching the color bar in A.

Table 1. Genome quality statistics and relative abundance calculations for all 66 EBPR SBR MAGs. Genome code names match names used in all figures and within the text. Classifications were assigned using the GTDB-tk [87] and confirmed by comparing against select publicly available references and a subset of HQ MAGs from Singleton et al. 2021 [39]. Completeness and redundancy estimates and GC content were calculated by CheckM [41]. tRNA and rRNA predictions were performed with Barrnap as part of the Prokka software [88]. Relative abundance estimates reflect the proportion of reads mapped to the genome in that sample divided by the total number of reads mapped to all genomes as performed with SingleM. Table available at https://figshare.com/articles/dataset/EBPR_SBR_MAGs_Metadata/13063874.

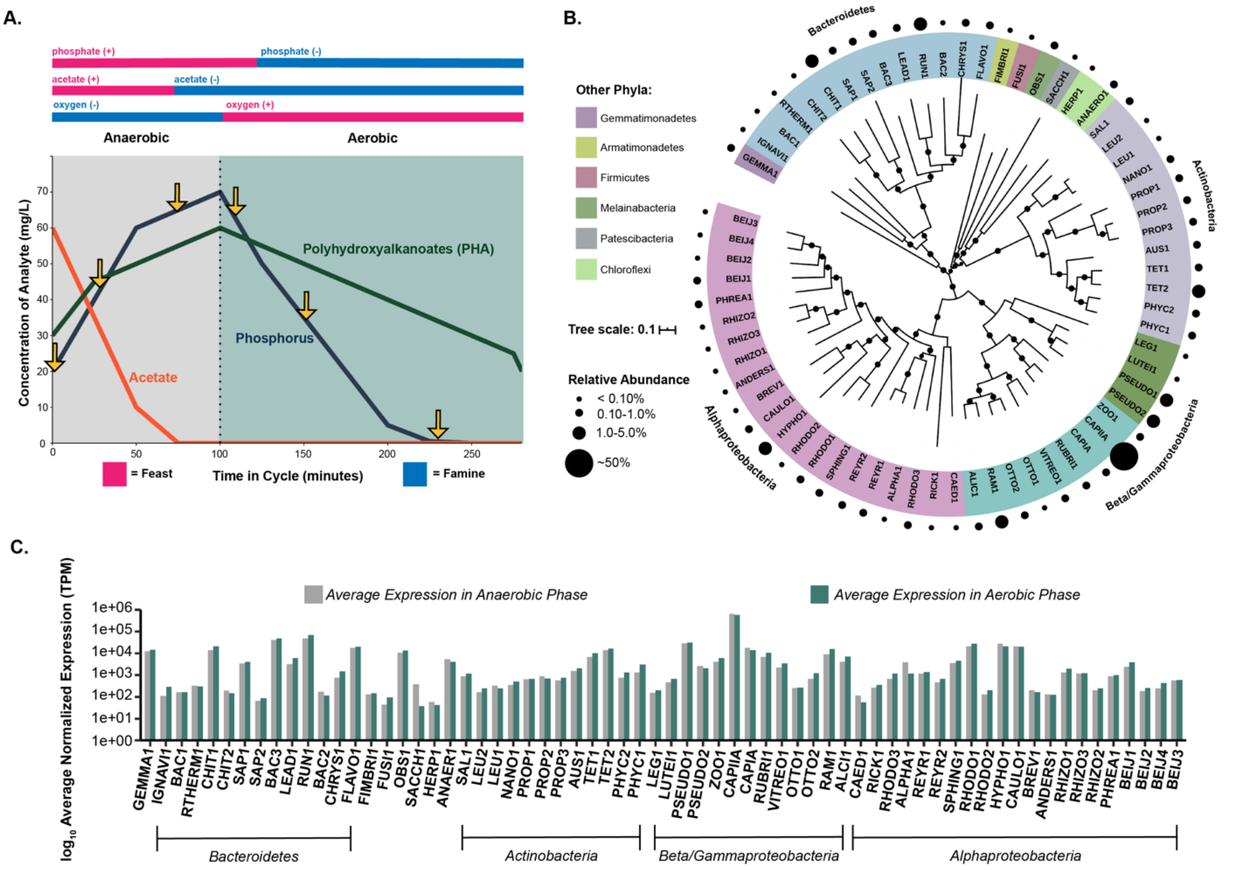
Table 2. KEGG Pathways for core trait-attributes present in greater than 19 genomes.

Table 3. KEGG Pathways for differentiating trait-attributes present between 3 and 18 genomes.

Figure 1.


Genome-Resolved Metagenomics
Assemble HQ MAGs and Map Metagenomic reads from all Timepoints
TCA Cycle
Glycolysis
PHA Synthesis
High-Affinity P Transporter
Denitrification
Predict Presence and Absence of Key Metabolic Pathways
High Affinity Phosphorus Transporter pstABCS
pstS
pstA
pstB
pstC
o-PO₄
PolyP
ATP
PHA
Putative Polyphosphate Accumulating Organisms
Putative Inferences of Physiology and Ecosystem Processes
Trait-Based Comparisons Integrating Time-Series Metatranscriptomics
Anaerobic
Aerobic
Concentration of Analyte (mg/L)
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Polyhydroxyalkanoates (PHA)
Phosphorus
Acetate
Time-Series Metatranscriptomics Experiment
Map RNASeq reads to HQ MAGs
pstS
pstA
pstB
pstC
Trait
Anaerobic
Aerobic
Log2 Expression
5
4
3
2
1
0
T1
T2
T3
T4
T5
T6
Trait Attributes
Anaerobic
Aerobic
Log2 Expression
5
4
3
2
1
0
T1
T2
T3
T4
T5
T6
Inter-Organismal Comparisons of Gene Expression Patterns of a Given Trait for Ecological and Ecosystem Inferences

983 **Figure 2.**



984

985

986

987

988

989

990

991

992

993

Figure 3.

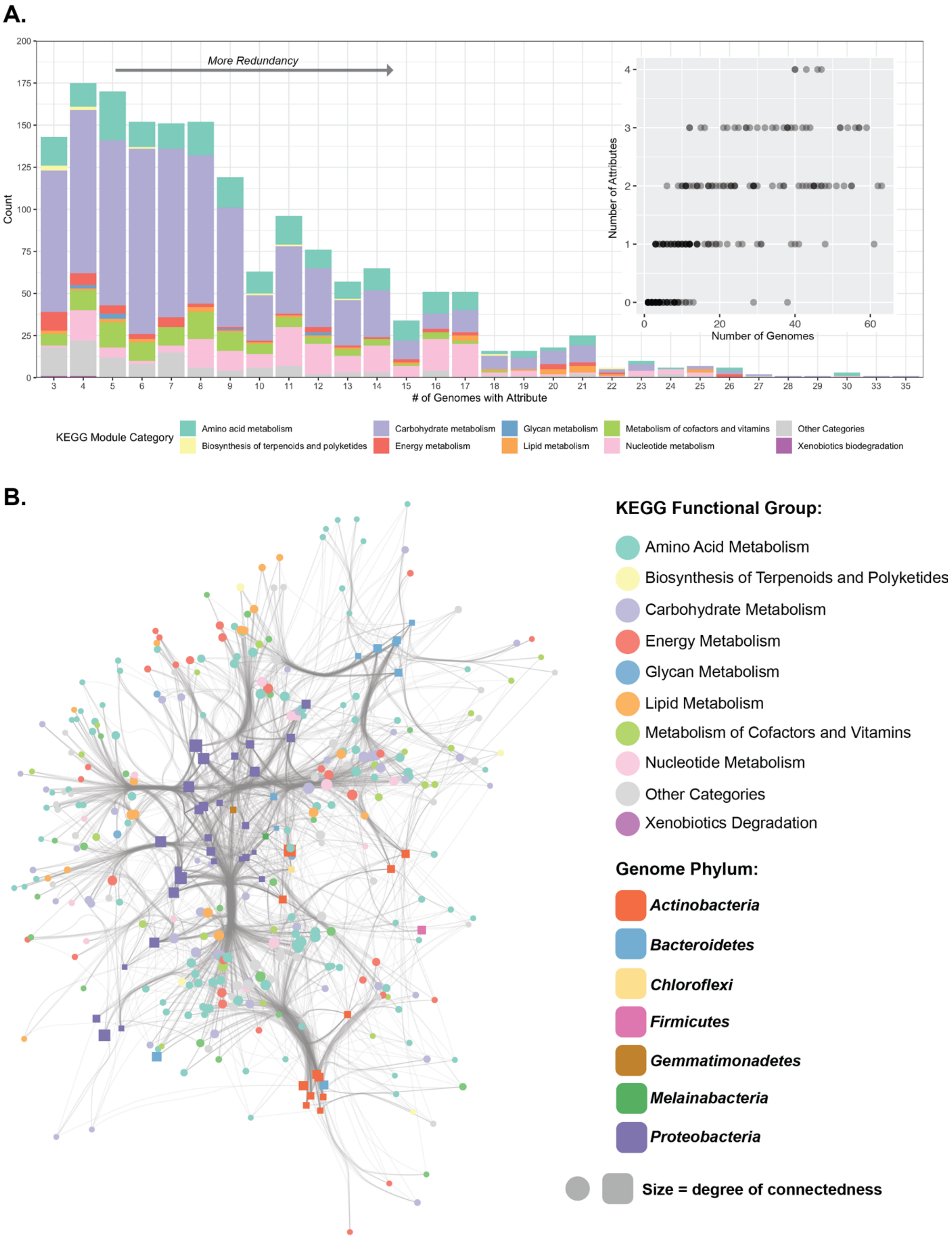

A.
More Redundancy
Count
200
175
150
125
100
75
50
25
0
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
33
35
# of Genomes with Attribute
Number of Attributes
4
3
2
1
0
0
20
40
60
Number of Genomes
KEGG Module Category
Amino acid metabolism
Biosynthesis of terpenoids and polyketides
Carbohydrate metabolism
Energy metabolism
Glycan metabolism
Lipid metabolism
Metabolism of cofactors and vitamins
Nucleotide metabolism
Other Categories
Xenobiotics biodegradation
B.
KEGG Functional Group:
Amino Acid Metabolism
Biosynthesis of Terpenoids and Polyketides
Carbohydrate Metabolism
Energy Metabolism
Glycan Metabolism
Lipid Metabolism
Metabolism of Cofactors and Vitamins
Nucleotide Metabolism
Other Categories
Xenobiotics Degradation
Genome Phylum:
Actinobacteria
Bacteroidetes
Chloroflexi
Firmicutes
Gemmatimonadetes
Melainabacteria
Proteobacteria
Size = degree of connectedness

Figure 4.

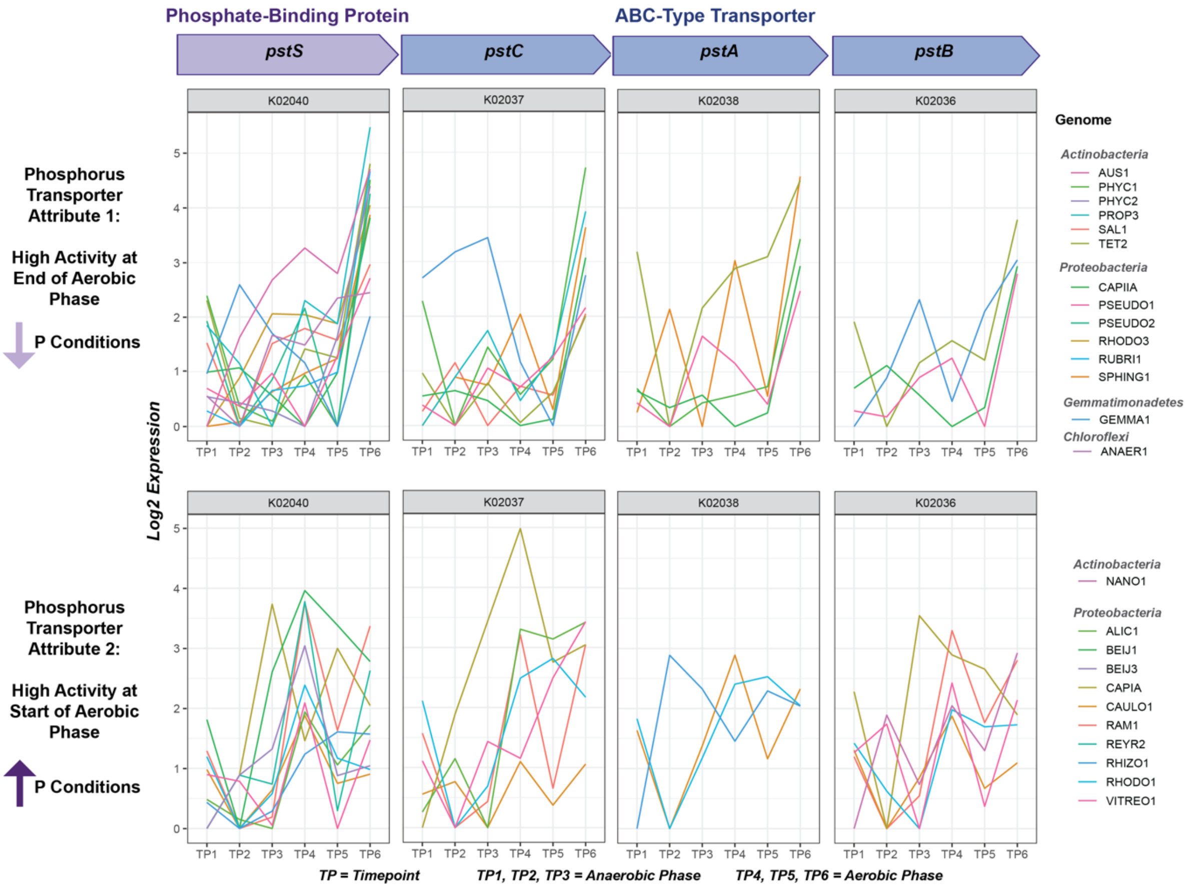

Phosphate-Binding Protein
ABC-Type Transporter
pstS
pstC
pstA
pstB
K02040
K02037
K02038
K02036
Phosphorus Transporter Attribute 1:
High Activity at End of Aerobic Phase
P Conditions
Phosphorus Transporter Attribute 2:
High Activity at Start of Aerobic Phase
P Conditions
Log2 Expression
0
1
2
3
4
5
TP1
TP2
TP3
TP4
TP5
TP6
Genome
Actinobacteria
AUS1
PHYC1
PHYC2
PROP3
SAL1
TET2
Proteobacteria
CAPIIA
PSEUDO1
PSEUDO2
RHODO3
RUBRI1
SPHING1
Gemmatimonadetes
GEMMA1
Chloroflexi
ANAER1
Actinobacteria
NANO1
Proteobacteria
ALIC1
BEIJ1
BEIJ3
CAPIA
CAULO1
RAM1
REYR2
RHIZO1
RHODO1
VITREO1
TP = Timepoint
TP1, TP2, TP3 = Anaerobic Phase
TP4, TP5, TP6 = Aerobic Phase

Figure 5.

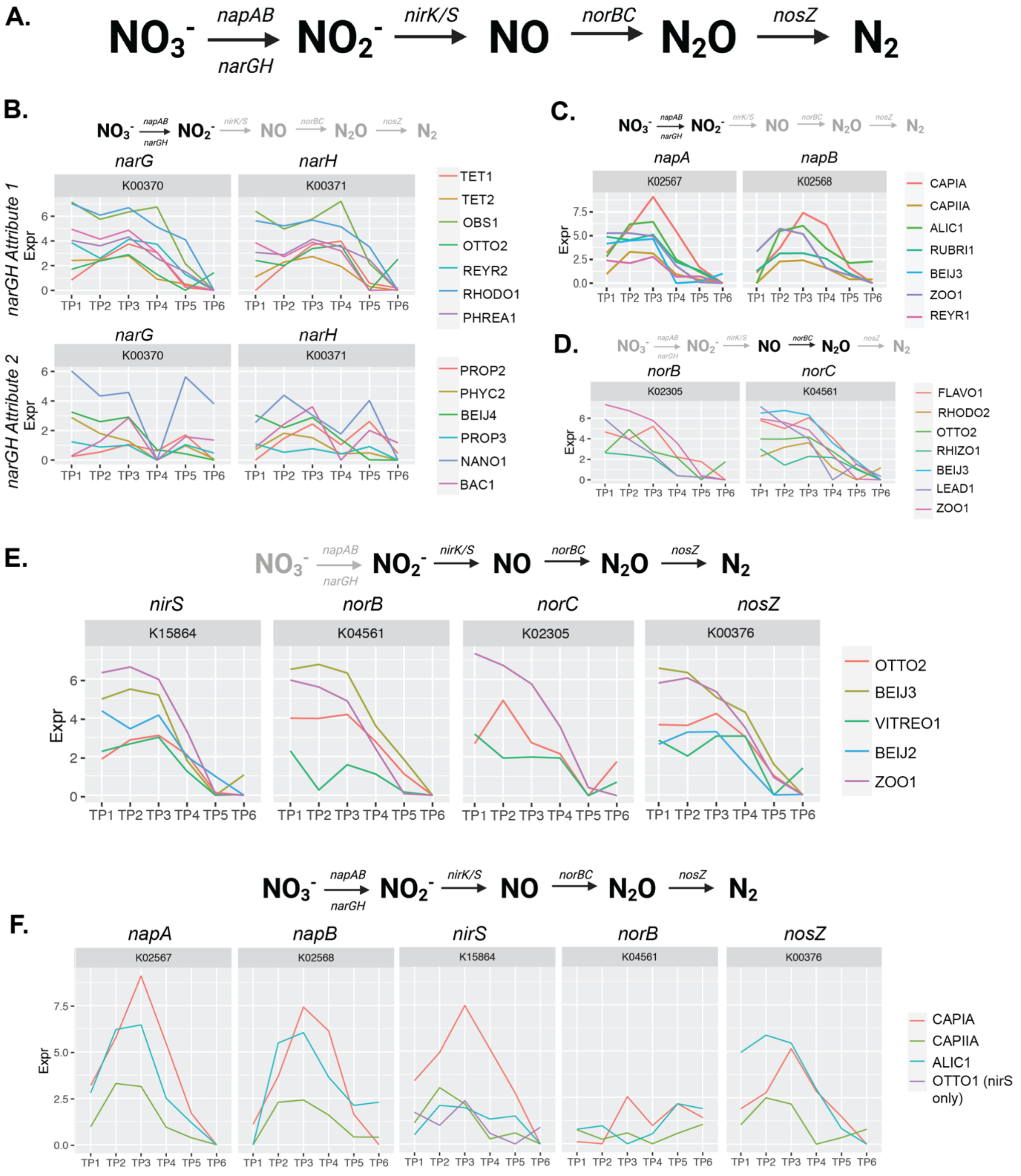

A.
NO_3^- napAB narGH NO_2^- nirK/S NO norBC N_2O nosZ N_2
B.
narG
narH
K00370
K00371
narGH Attribute 1
narGH Attribute 2
Expr
0
2
4
6
TP1 TP2 TP3 TP4 TP5 TP6
TET1
TET2
OBS1
OTTO2
REYR2
RHODO1
PHREA1
PROP2
PHYC2
BEIJ4
PROP3
NANO1
BAC1
C.
napA
napB
K02567
K02568
0.0
2.5
5.0
7.5
CAPIA
CAPIIA
ALIC1
RUBRI1
BEIJ3
ZOO1
REYR1
D.
norB
norC
K02305
K04561
FLAVO1
RHODO2
OTTO2
RHIZO1
BEIJ3
LEAD1
ZOO1
E.
nirS
norB
norC
nosZ
K15864
K04561
K02305
K00376
OTTO2
BEIJ3
VITREO1
BEIJ2
ZOO1
F.
napA
napB
nirS
norB
nosZ
K02567
K02568
K15864
K04561
K00376
CAPIA
CAPIIA
ALIC1
OTTO1 (nirS only)

Figure 6.

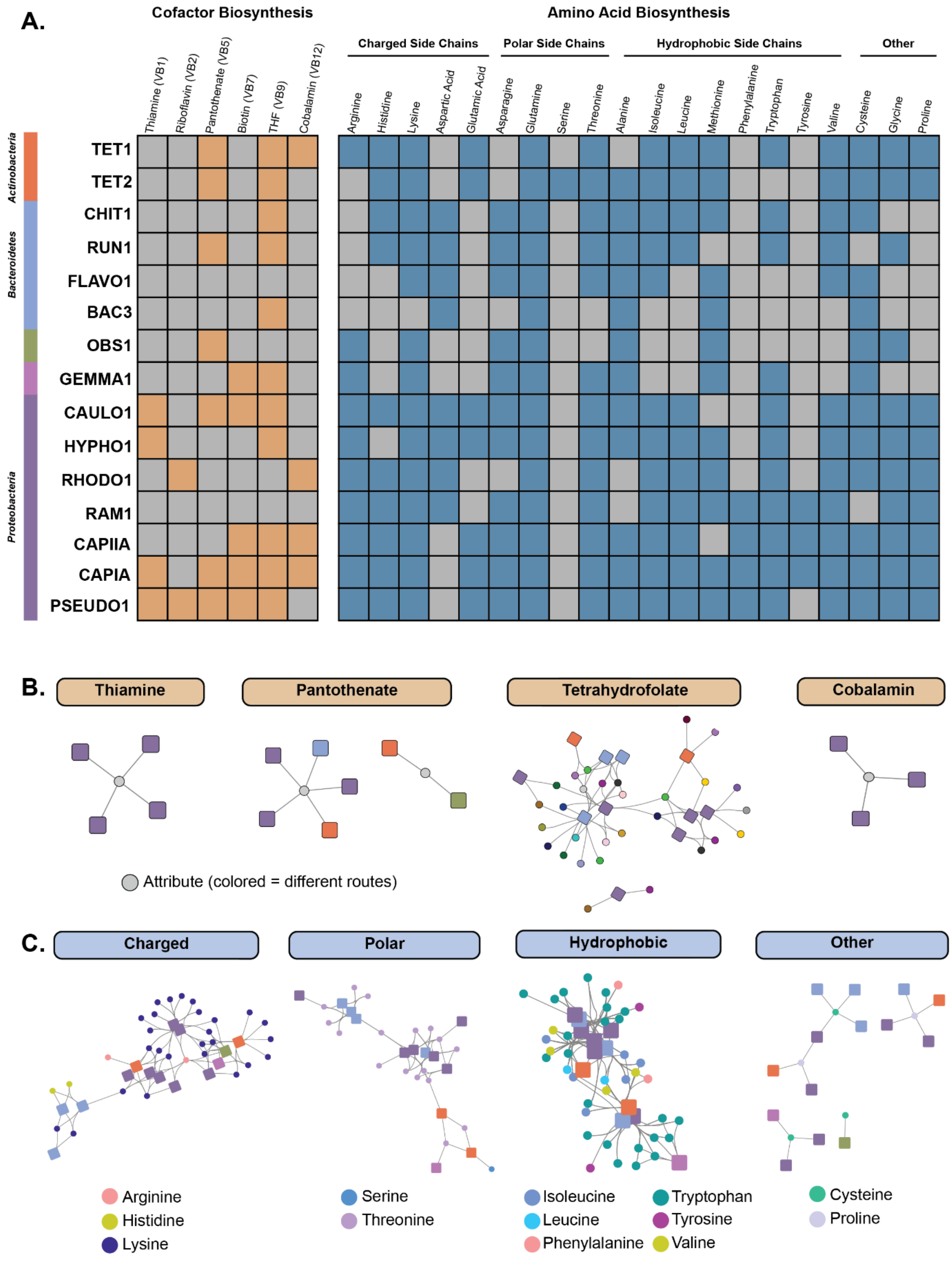
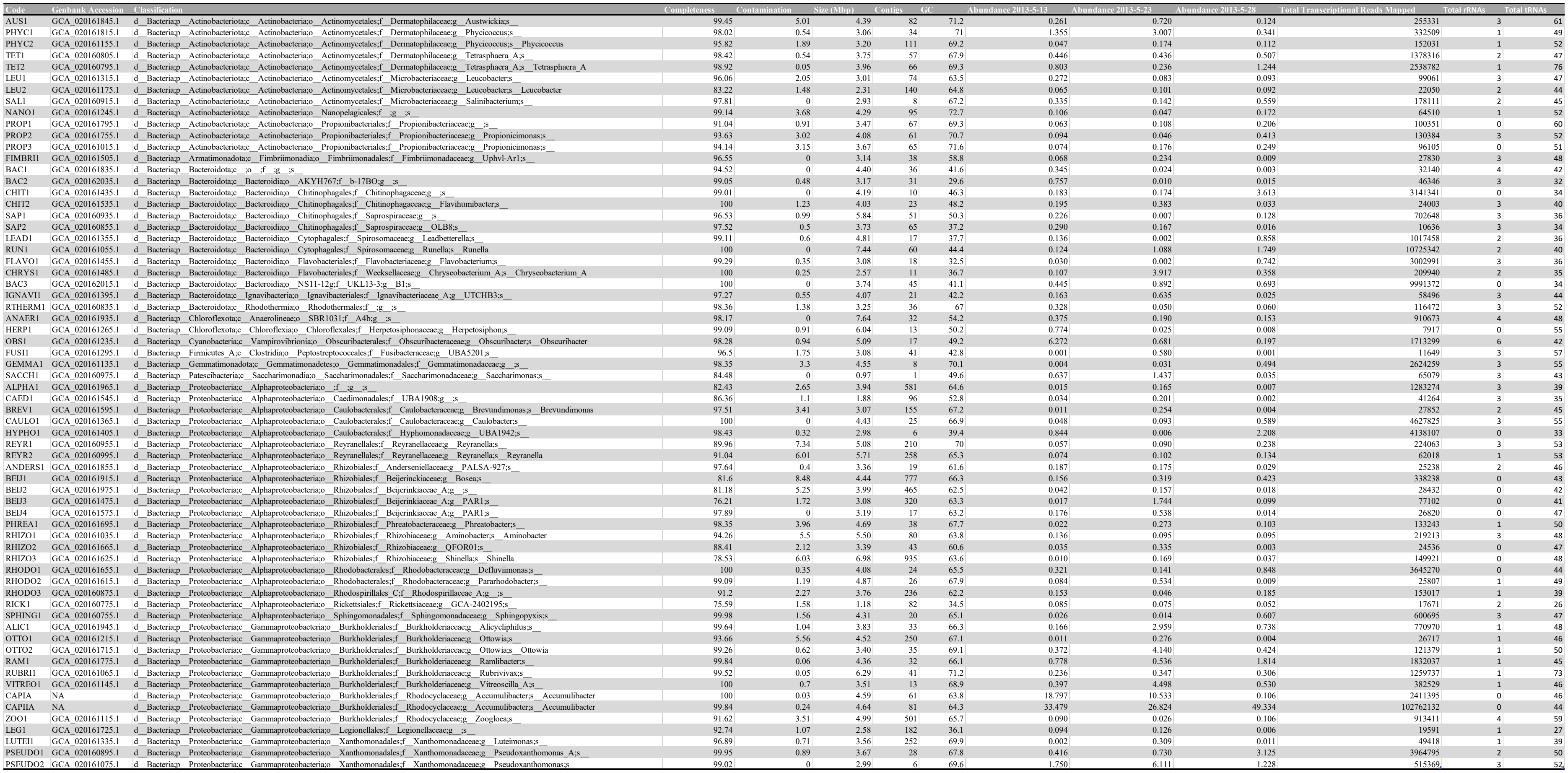

A.
Cofactor Biosynthesis
Amino Acid Biosynthesis
Charged Side Chains
Polar Side Chains
Hydrophobic Side Chains
Other
Thiamine (VB1)
Riboflavin (VB2)
Pantothenate (VB5)
Biotin (VB7)
THF (VB9)
Cobalamin (VB12)
Arginine
Histidine
Lysine
Aspartic Acid
Glutamic Acid
Asparagine
Glutamine
Serine
Threonine
Alanine
Isoleucine
Leucine
Methionine
Phenylalanine
Tryptophan
Tyrosine
Valine
Cysteine
Glycine
Proline
Actinobacteria
Bacteroidetes
Proteobacteria
TET1
TET2
CHIT1
RUN1
FLAVO1
BAC3
OBS1
GEMMA1
CAULO1
HYPHO1
RHODO1
RAM1
CAPIIA
CAPIA
PSEUDO1
B.
Thiamine
Pantothenate
Tetrahydrofolate
Cobalamin
Attribute (colored = different routes)
C.
Charged
Polar
Hydrophobic
Other
Arginine
Histidine
Lysine
Serine
Threonine
Isoleucine
Leucine
Phenylalanine
Tryptophan
Tyrosine
Valine
Cysteine
Proline

Table 1.



Code	Genbank Accession	Classification	Completeness	Contamination	Size (Mbp)	Contigs	GC	Abundance 2013-5-13	Abundance 2013-5-23	Abundance 2013-5-28	Total Transcriptional Reads Mapped	Total rRNAs	Total tRNAs
AUS1	GCA_020161845.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Dermatophilaceae;g__Austwickia;s__	99.45	5.01	4.39	82	71.2	0.261	0.720	0.124	255331	3	61
PHYC1	GCA_020161815.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Dermatophilaceae;g__Phycicoccus;s__	98.02	0.54	3.06	34	71	1.355	3.007	0.341	332509	1	49
PHYC2	GCA_020161155.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Dermatophilaceae;g__Phycicoccus;s__Phycicoccus	95.82	1.89	3.20	111	69.2	0.047	0.174	0.112	152031	1	52
TET1	GCA_020160805.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Dermatophilaceae;g__Tetrasphaera_A;s__	98.42	0.54	3.75	57	67.9	0.446	0.436	0.507	1378316	2	47
TET2	GCA_020160795.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Dermatophilaceae;g__Tetrasphaera_A;s__Tetrasphaera_A	98.92	0.05	3.96	66	69.3	0.803	0.236	1.244	2538782	1	76
LEU1	GCA_020161315.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae;g__Leucobacter;s__	96.06	2.05	3.01	74	63.5	0.272	0.083	0.093	99061	3	47
LEU2	GCA_020161175.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae;g__Leucobacter;s__Leucobacter	83.22	1.48	2.31	140	64.8	0.065	0.101	0.092	22050	2	44
SAL1	GCA_020160915.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae;g__Salinibacterium;s__	97.81	0	2.93	8	67.2	0.335	0.142	0.559	178111	2	45
NANO1	GCA_020161245.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Nanopelagicales;f__;g__;s__	99.14	3.68	4.29	95	72.7	0.106	0.047	0.172	64510	1	52
PROP1	GCA_020161795.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Propionibacteriales;f__Propionibacteriaceae;g__;s__	91.04	0.91	3.47	67	69.3	0.063	0.108	0.206	100351	0	60
PROP2	GCA_020161755.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Propionibacteriales;f__Propionibacteriaceae;g__Propionicimonas;s__	93.63	3.02	4.08	61	70.7	0.094	0.046	0.413	130384	3	52
PROP3	GCA_020161015.1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Propionibacteriales;f__Propionibacteriaceae;g__Propionicimonas;s__	94.14	3.15	3.67	65	71.6	0.074	0.176	0.249	96105	0	51
FIMBRI1	GCA_020161505.1	d__Bacteria;p__Armatimonadota;c__Fimbriimonadia;o__Fimbriimonadales;f__Fimbriimonadaceae;g__Uphvl-Ar1;s__	96.55	0	3.14	38	58.8	0.068	0.234	0.009	27830	3	48
BAC1	GCA_020161835.1	d__Bacteria;p__Bacteroidota;c__;o__;f__;g__;s__	94.52	0	4.40	36	41.6	0.345	0.024	0.003	32140	4	42
BAC2	GCA_020162035.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__AKYH767;f__b-17BO;g__;s__	99.05	0.48	3.17	31	29.6	0.757	0.010	0.015	46346	3	32
CHIT1	GCA_020161435.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__;s__	99.01	0	4.19	10	46.3	0.183	0.174	3.613	3141341	0	34
CHIT2	GCA_020161535.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Flavihumibacter;s__	100	1.23	4.03	23	48.2	0.195	0.383	0.033	24003	3	40
SAP1	GCA_020160935.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Saprospiraceae;g__;s__	96.53	0.99	5.84	51	50.3	0.226	0.007	0.128	702648	3	36
SAP2	GCA_020160855.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Saprospiraceae;g__OLB8;s__	97.52	0.5	3.73	65	37.2	0.290	0.167	0.016	10636	3	34
LEAD1	GCA_020161355.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Leadbetterella;s__	99.11	0.6	4.81	17	37.7	0.136	0.002	0.858	1017458	2	36
RUN1	GCA_020161055.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Runella;s__Runella	100	0	7.44	60	44.4	0.124	1.088	1.749	10725342	2	40
FLAVO1	GCA_020161455.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium;s__	99.29	0.35	3.08	18	32.5	0.030	0.002	0.742	3002991	3	36
CHRYS1	GCA_020161485.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Weeksellaceae;g__Chryseobacterium_A;s__Chryseobacterium_A	100	0.25	2.57	11	36.7	0.107	3.917	0.358	209940	2	35
BAC3	GCA_020162015.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__NS11-12g;f__UKL13-3;g__B1;s__	100	0	3.74	45	41.1	0.445	0.892	0.693	9991372	0	34
IGNAVI1	GCA_020161395.1	d__Bacteria;p__Bacteroidota;c__Ignavibacteria;o__Ignavibacteriales;f__Ignavibacteriaceae_A;g__UTCHB3;s__	97.27	0.55	4.07	21	42.2	0.163	0.635	0.025	58496	3	44
RTHERM1	GCA_020160835.1	d__Bacteria;p__Bacteroidota;c__Rhodothermia;o__Rhodothermales;f__;g__;s__	98.36	1.38	3.25	36	67	0.328	0.050	0.060	116472	3	52
ANAER1	GCA_020161935.1	d__Bacteria;p__Chloroflexota;c__Anaerolineae;o__SBR1031;f__A4b;g__;s__	98.17	0	7.64	32	54.2	0.375	0.190	0.153	910673	4	48
HERP1	GCA_020161265.1	d__Bacteria;p__Chloroflexota;c__Chloroflexia;o__Chloroflexales;f__Herpetosiphonaceae;g__Herpetosiphon;s__	99.09	0.91	6.04	13	50.2	0.774	0.025	0.008	7917	0	55
OBS1	GCA_020161235.1	d__Bacteria;p__Cyanobacteria;c__Vampirovibrionia;o__Obscuribacterales;f__Obscuribacteraceae;g__Obscuribacter;s__Obscuribacter	98.28	0.94	5.09	17	49.2	6.272	0.681	0.197	1713299	6	42
FUSI1	GCA_020161295.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__Fusibacteraceae;g__UBA5201;s__	96.5	1.75	3.08	41	42.8	0.001	0.580	0.001	11649	3	57
GEMMA1	GCA_020161135.1	d__Bacteria;p__Gemmatimonadota;c__Gemmatimonadetes;o__Gemmatimonadales;f__Gemmatimonadaceae;g__;s__	98.35	3.3	4.55	8	70.1	0.004	0.031	0.494	2624259	3	55
SACCH1	GCA_020160975.1	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__Saccharimonadaceae;g__Saccharimonas;s__	84.48	0	0.97	1	49.6	0.637	1.437	0.035	65079	3	43
ALPHA1	GCA_020161965.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__;f__;g__;s__	82.43	2.65	3.94	581	64.6	0.015	0.165	0.007	1283274	3	39
CAED1	GCA_020161545.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caedimonadales;f__UBA1908;g__;s__	86.36	1.1	1.88	96	52.8	0.034	0.201	0.002	41264	3	35
BREV1	GCA_020161595.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__Brevundimonas;s__Brevundimonas	97.51	3.41	3.07	155	67.2	0.011	0.254	0.004	27852	2	45
CAULO1	GCA_020161365.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__Caulobacter;s__	100	0	4.43	25	66.9	0.048	0.093	0.589	4627825	3	55
HYPHO1	GCA_020161405.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Hyphomonadaceae;g__UBA1942;s__	98.43	0.32	2.98	6	39.4	0.844	0.006	2.208	4138107	0	33
REYR1	GCA_020160955.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Reyranellales;f__Reyranellaceae;g__Reyranella;s__	89.96	7.34	5.08	210	70	0.057	0.090	0.238	224063	3	53
REYR2	GCA_020160995.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Reyranellales;f__Reyranellaceae;g__Reyranella;s__Reyranella	91.04	6.01	5.71	258	65.3	0.074	0.102	0.134	62018	1	53
ANDERS1	GCA_020161855.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Anderseniellaceae;g__PALSA-927;s__	97.64	0.4	3.36	19	61.6	0.187	0.175	0.029	25238	2	46
BEIJ1	GCA_020161915.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Bosea;s__	81.6	8.48	4.44	777	66.3	0.156	0.319	0.423	338238	0	43
BEIJ2	GCA_020161975.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinkiaceae_A;g__;s__	81.18	5.25	3.99	465	62.5	0.042	0.157	0.018	28432	0	42
BEIJ3	GCA_020161475.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinkiaceae_A;g__PAR1;s__	76.21	1.72	3.08	320	63.3	0.017	1.744	0.099	77102	0	41
BEIJ4	GCA_020161575.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinkiaceae_A;g__PAR1;s__	97.89	0	3.19	17	63.2	0.176	0.538	0.014	26820	0	47
PHREA1	GCA_020161695.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Phreatobacteraceae;g__Phreatobacter;s__	98.35	3.96	4.69	38	67.7	0.022	0.273	0.103	133243	1	50
RHIZO1	GCA_020161035.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhizobiaceae;g__Aminobacter;s__Aminobacter	94.26	5.5	5.50	80	63.8	0.136	0.095	0.095	219213	3	48
RHIZO2	GCA_020161665.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhizobiaceae;g__QFOR01;s__	88.41	2.12	3.39	43	60.6	0.035	0.335	0.003	24536	0	47
RHIZO3	GCA_020161625.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhizobiaceae;g__Shinella;s__Shinella	78.53	6.03	6.98	935	63.6	0.010	0.169	0.037	149921	0	48
RHODO1	GCA_020161655.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__Defluviimonas;s__	100	0.35	4.08	24	65.5	0.321	0.141	0.848	3645270	0	44
RHODO2	GCA_020161615.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__Pararhodobacter;s__	99.09	1.19	4.87	26	67.9	0.084	0.534	0.009	25807	1	49
RHODO3	GCA_020160875.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales_C;f__Rhodospirillaceae_A;g__;s__	91.2	2.27	3.76	236	62.2	0.153	0.046	0.185	153017	1	39
RICK1	GCA_020160775.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__Rickettsiaceae;g__GCA-2402195;s__	75.59	1.58	1.18	82	34.5	0.085	0.075	0.052	17671	2	26
SPHING1	GCA_020160755.1	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingopyxis;s__	99.98	1.56	4.31	20	65.1	0.026	0.014	0.607	600695	3	47
ALIC1	GCA_020161945.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Alicycliphilus;s__	99.64	1.04	3.83	33	66.3	0.166	2.959	0.738	770970	1	48
OTTO1	GCA_020161215.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ottowia;s__	93.66	5.56	4.52	250	67.1	0.011	0.276	0.004	26717	1	46
OTTO2	GCA_020161715.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ottowia;s__Ottowia	99.26	0.62	3.40	35	69.1	0.372	4.140	0.424	121379	1	50
RAM1	GCA_020161775.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ramlibacter;s__	99.84	0.06	4.36	32	66.1	0.778	0.536	1.814	1832037	1	45
RUBRI1	GCA_020161065.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Rubrivivax;s__	99.52	0.05	6.29	41	71.2	0.236	0.347	0.306	1259737	1	73
VITREO1	GCA_020161145.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Vitreoscilla_A;s__	100	0.7	3.51	13	68.9	0.397	4.498	0.530	382529	1	46
CAPIA	NA	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Rhodocyclaceae;g__Accumulibacter;s__Accumulibacter	100	0.03	4.59	61	63.8	18.797	10.533	0.106	2411395	0	46
CAPIIA	NA	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Rhodocyclaceae;g__Accumulibacter;s__Accumulibacter	99.84	0.24	4.64	81	64.3	33.479	26.824	49.334	102762132	0	44
ZOO1	GCA_020161115.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Rhodocyclaceae;g__Zoogloea;s__	91.62	3.51	4.99	501	65.7	0.090	0.026	0.106	913411	4	59
LEG1	GCA_020161725.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Legionellales;f__Legionellaceae;g__;s__	92.74	1.07	2.58	182	36.1	0.094	0.126	0.006	19591	1	27
LUTEI1	GCA_020161335.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Luteimonas;s__	96.89	0.71	3.56	252	69.9	0.002	0.309	0.011	49418	1	39
PSEUDO1	GCA_020160895.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Pseudoxanthomonas_A;s__	99.95	0.89	3.67	28	67.8	0.416	0.730	3.125	3964795	2	50
PSEUDO2	GCA_020161075.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Pseudoxanthomonas;s	99.02	0	2.99	6	69.6	1.750	6.111	1.228	515369	3	52

File at https://figshare.com/account/projects/90437/articles/13063874

Table 2.

Module Description	Number of Attributes
Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate [PATH:map00020 map01200 map01100]	13
Citrate cycle (TCA cycle, Krebs cycle) [PATH:map00020 map01200 map01100]	10
Shikimate pathway, phosphoenolpyruvate + erythrose-4P => chorismate [PATH:map00400 map01230 map01100 map01110]	8
Fatty acid biosynthesis, initiation [PATH:map00061 map01212 map01100]	7
Glycolysis, core module involving three-carbon compounds [PATH:map00010 map01200 map01230 map01100]	7
Adenine ribonucleotide biosynthesis, IMP => ADP,ATP [PATH:map00230 map01100]	4
Guanine ribonucleotide biosynthesis IMP => GDP,GTP [PATH:map00230 map01100]	4
Inosine monophosphate biosynthesis, PRPP + glutamine => IMP [PATH:map00230 map01100]	4
Isoleucine biosynthesis, threonine => 2-oxobutanoate => isoleucine [PATH:map00290 map01230 map01100]	3
NADH:quinone oxidoreductase, prokaryotes [PATH:map00190]	3
beta-Oxidation, acyl-CoA synthesis [PATH:map00061 map00071 map01212 map01100]	2
F-type ATPase, prokaryotes and chloroplasts [PATH:map00190 map00195]	2
Valine/isoleucine biosynthesis, pyruvate => valine / 2-oxobutanoate => isoleucine [PATH:map00290 map00770 map01210 map01230 map01100 map01110]	2
CAM (Crassulacean acid metabolism), dark [PATH:map00620 map00710 map01200 map01100 map01120]	1
Cytochrome c oxidase, cbb3-type [PATH:map00190]	1
Cytochrome c oxidase, prokaryotes [PATH:map00190]	1
dTDP-L-rhamnose biosynthesis [PATH:map00521 map00523 map01100 map01130]	1
Leucine biosynthesis, 2-oxoisovalerate => 2-oxoisocaproate [PATH:map00290 map01210 map01230 map01100 map01110]	1
Phosphatidylethanolamine (PE) biosynthesis, PA => PS => PE [PATH:map00564 map01100]	1
PRPP biosynthesis, ribose 5P => PRPP [PATH:map00030 map00230 map01200 map01230 map01100]	1
Pyruvate oxidation, pyruvate => acetyl-CoA [PATH:map00010 map00020 map00620 map01200 map01100]	1
Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P [PATH:map00030 map01200 map01100 map01120]	1
Threonine biosynthesis, aspartate => homoserine => threonine [PATH:map00260 map01230 map01100 map01110]	1

Table 3.

Module_description	Number of Attributes
Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate [PATH:map00010 map01200 map01100]	279
Citrate cycle (TCA cycle, Krebs cycle) [PATH:map00020 map01200 map01100]	208
Gluconeogenesis, oxaloacetate => fructose-6P [PATH:map00010 map00020 map01100]	76
Inosine monophosphate biosynthesis, PRPP + glutamine => IMP [PATH:map00230 map01100]	45
Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate [PATH:map00020 map01200 map01100]	31
Heme biosynthesis, plants and bacteria, glutamate => heme [PATH:map00860 map01100 map01110]	27
Tetrahydrofolate biosynthesis, GTP => THF [PATH:map00790 map00670 map01100]	25
Tryptophan biosynthesis, chorismate => tryptophan [PATH:map00400 map01230 map01100 map01110]	25
Ornithine biosynthesis, glutamate => ornithine [PATH:map00220 map01210 map01230 map01100]	24
Histidine biosynthesis, PRPP => histidine [PATH:map00340 map01230 map01100 map01110]	17
Pentose phosphate pathway (Pentose phosphate cycle) [PATH:map00030 map01200 map01100 map01120]	16
Lysine biosynthesis, succinyl-DAP pathway, aspartate => lysine [PATH:map00300 map01230 map01100]	12
Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP [PATH:map00240 map01100]	11