Metabolic pathways for the whole community

Niels W. Hanson[1], Kishori M. Konwar[2], Alyse K. Hawley[2], Tomer Altman[3], Peter D. Karp[4], and Steven J. Hallam[1,2,*]

[1]Graduate Program in Bioinformatics, University of British Columbia, Canada

[2]Department of Microbiology & Immunology, University of British Columbia, Canada

[3]Biomedical Informatics, Stanford University, USA

[4]Bioinformatics Research Group, SRI International, USA

*To whom correspondence should be addressed: University of British Columbia, Department of Microbiology & Immunology. 2552-2350 Health Sciences Mall, Vancouver, British Columbia, V6T1Z3, Canada. Office: (604) 827-3420 FAX: (604) 822-6041 e-mail: shallam@mail.ubc.ca

ABSTRACT

Background: A convergence of high-throughput sequencing and computational power is transforming biology into information science. Despite these technological advances, converting bits and bytes of sequence information into meaningful insights remains a challenging enterprise. Biological systems operate on multiple hierarchical levels from genomes to biomes. Holistic understanding of biological systems requires agile software tools that permit comparative analyses across multiple information levels (DNA, RNA, protein, and metabolites) to identify emergent properties, diagnose system states, or predict responses to environmental change. **Results:** Here we adopt the MetaPathways annotation and analysis pipeline and Pathway Tools to construct environmental pathway/genome databases (ePGDBs) that describe microbial community metabolism using MetaCyc, a highly curated database of metabolic pathways and components covering all domains of life. We evaluate Pathway Tools' performance on three datasets with different complexity and coding potential, including simulated metagenomes, a symbiotic system, and the Hawaii Ocean Time-series. We define accuracy and sensitivity relationships between read length, coverage and pathway recovery and evaluate the impact of taxonomic pruning on ePGDB construction. Resulting ePGDBs provide interactive metabolic maps, predict emergent metabolic pathways associated with biosynthesis and energy production and differentiate between genomic potential and phenotypic expression across defined environmental gradients.

Conclusions: This multi-tiered analysis provides the user community with specific operating guidelines, performance metrics and prediction hazards for more reliable ePGDB construction and interpretation. Moreover, it demonstrates the power of Pathway Tools in predicting metabolic interactions in natural and human engineered ecosystems.

BACKGROUND

Community interactions between uncultivated microorganisms give rise to dynamic metabolic networks integral to ecosystem function and global scale biogeochemical cycles [1]. Metagenomics bridges the "cultivation gap" through plurality or single-cell sequencing by providing direct and quantitative insight into microbial community structure and function [2,3]. Although, new technologies are rapidly expanding our capacity to chart microbial sequence space, persistent computational and analytical bottlenecks impede comparative analyses across multiple information levels (DNA, RNA, protein and metabolites) [4,5]. This in turn limits our ability to convert the genetic potential and phenotypic expression of microbial communities into predictive insights and technological or therapeutic innovations.

Functional genes operate within the structure of metabolic pathways and reactions that define metabolic networks. Despite this fact, few metagenomic studies use pathway-centric approaches to predict microbial community interaction networks based on known biochemical rules. Recently, algorithms for pathway prediction and metabolic flux have been developed for environmental sequence information including the Human Microbiome Project Unified Metabolic Analysis Network (HUMAnN) and Predicted Relative Metabolic Turnover (PRMT). The HUMAnN uses an integer optimization algorithm that conservatively computes a parsimonious minimum set of reactions along KEGG pathways based on pathway presence, absence or completion [6,7]. PRMT infers metabolic flux based on normalized enzyme activity counts mapped to KEGG pathways across multiple metagenomes [8]. Because KEGG pathways are coarse and do not discriminate between pathway variants, both modes of analysis have limited metabolic resolution [9]. Moreover, neither HUMAnN nor PRMT provides a coherent structure for exploring and interpreting predicted KEGG pathways.

One alternative to HUMAnN and PRMT is Pathway Tools, a production-quality software environment supporting metabolic inference and flux balance analysis based on the MetaCyc database of metabolic pathways and enzymes representing all domains of life [10-13]. Unlike KEGG or SEED subsystems, MetaCyc emphasizes smaller, evolutionarily conserved or co-regulated units of metabolism and contains the largest collection (over 2000) of experimentally validated metabolic pathways. Extensively commented pathway descriptions, literature citations, and enzyme properties combined within a pathway/genome database (PGDB) provide a coherent structure for exploring and interpreting predicted pathways. Although initially conceived for cellular organisms, recent development of the MetaPathways pipeline extends the PGDB concept to environmental sequence information enabling pathway-centric insights into microbial community structure and function [14].

Here we provide essential guidelines for generating and interpreting ePGDBs inspired by the multi-tiered structure of BioCyc [15] **(Figure 1)**. We begin with genome and metagenome simulations to assess performance on datasets manifesting different read length, coverage and taxonomic diversity. We then demonstrate the power of Pathway Tools to predict emergent metabolism in simulated metagenomes and a previously characterized symbiotic system [16]. Finally, we generate ePGDBs using coupled metagenomic and metatranscriptomic datasets from the Hawaii Ocean Time-series (HOT) to compare and contrast genetic potential and phenotypic expression along defined environmental gradients in the ocean [17-19].

RESULTS AND DISCUSSION

Performance Considerations

Environmental pathway/genome database (ePGDB) construction commences with the MetaPathways automated annotation pipeline using environmental sequence information as input **(Materials and Methods)**. Resulting annotations are used by the PathoLogic algorithm implemented in Pathway Tools to predict metabolic pathways based on multiple criteria including proportion of pathways found, pathway specific enzymatic reactions, and purported taxon-specific pathway distributions. PathoLogic is known to perform well when compared to machine learning methods using the genomes of cellular organisms as input [20]. We previously reported PathoLogic's performance on combined and incomplete genomes using two simulated metagenomes (Sim1 and Sim2) derived from 10 BioCyc tier-2 PGDBs manifesting different coverage and taxonomic diversity using MetaSim [14, 21]. Simulations on increasing proportions of the total component genome length (G_m) showed that the performance of pathway recovery based on multiple metrics (F-measure, Matthews Correlation Coefficient, etc.) increased with sequence coverage and sample diversity nearing an asymptote at higher coverage **(Figure 2a)**. This suggests that pathway prediction follows a collector's curve in which common core pathways accumulate in the early part of the curve followed by less common accessory pathways near the asymptote.

To better constrain pathway recovery and performance in relation to ePGDB construction we compared results of MetaSim experiments using the *Escherichia coli K12 substr. MG1655* genome (basis of the EcoCyc database), Sim1 and Sim2, and a subsampled 25 m metagenome from HOT [18] **(Materials and Methods, Tables S1-S4 and Figure S1 in Additional File 1)**. Simulations were performed at progressively larger G_m coverage. Consistent with previous observations for Sim1 and Sim2, all experiments showed that pathway recovery percentage and performance sensitivity increased with sequence coverage and sample diversity nearing an

asymptote at higher coverage **(Fig. 2a-b)**. The absolute values of these patterns were sensitive to read length and likely reflected limits imposed by open reading frame prediction and BLAST/LAST-based annotation. In contrast, performance specificity was high (>85%) regardless of read length, coverage, or taxonomic diversity **(Fig. 2b)**. The rate of pathway recovery increased proportionally with increasing sample diversity at lower coverage values, as seen in the reduction of pathway recovery percentage between Sim1, Sim2 and *E. coli* for long read (~700 bp) and between HOT, Sim1/2 and *E. coli* for short read (~160 bp) datasets. Additional performance metrics can be found in **Additional File 1 (Tables S5–S8)**. Because PathoLogic performance improves with increasing read length, coverage and sample diversity, sequencing platform selection and use of assembled versus unassembled sequence information should be considered when generating ePGDBs.

In addition to performance, we evaluated the impact of enabling or disabling taxon-specific pathway distributions on pathway recovery (**Table S9 in Additional File 1)**. PathoLogic uses a process called taxonomic pruning to constrain the pathway prediction process such that pathways whose expected taxonomic range does not include the organism whose pathways are being predicted must have all enzymes present to be predicted, e.g., decreasing the probability that a plant pathway will be predicted for a bacterium. This rule is probably not appropriate for metagenomic pathway analysis. While this increases algorithm specificity it has the potential to decrease sensitivity, i.e., pathway recovery below the selected taxonomic threshold. To address this issue, we ran PathoLogic on Sim1/2 and 25 m HOT datasets with the 'Unclassified sequences' pruning threshold and without pruning. With taxonomic pruning enabled, long read and short read Sim1 ePGDBs exhibited a reduction of 56% (206 compared to 604) and 61% (194 compared to 499) predicted pathways, respectively. Interestingly, the subsampled 25 m HOT

dataset exhibited a 28% reduction (425 compared to 593) in pathway recovery with and without pruning suggesting that increased sample complexity can partially offset taxon specific sensitivity losses. Based on these results, taxonomic pruning should be disabled when creating ePGDBs, opening PathoLogic to microbial community metabolism across all three domains of life.

Distributed Metabolic Pathways

Public good dynamics play an integral role in shaping microbial interactions through distributed networks of metabolite exchange [22]. Such networks promote increased fitness and resilience and may explain the underlying difficulty in cultivating most environmental microorganisms [23-25]. Because ePGDBs are constructed from environmental sequence information, predicted pathways are represented by multiple donor genotypes providing different levels of sequence coverage for each reaction. By comparing pathway recovery for individual reference genomes to pathway recovery for combinations of reference genomes, it becomes formally possible to use Pathway Tools to identify distributed metabolic pathways that emerge between multiple interacting partners. To test this hypothesis, we selected four Tier-2 reference genomes used in simulation experiments and constructed ePGDBs using all possible pair-wise genome combinations **(Table S10 in Additional File 1)**. Thirty distributed pathways were identified in pair-wise genome combinations that were not predicted in PGDBs for individual cellular organisms using set-difference analysis **(Table S11 in Additional File 1)**. Common and unique reactions associated with distributed pathways could be identified as composite glyphs in the Pathway Tools genome browser **(Figure S2 in Additional File 1)**.

To provide a real world example of distributed metabolic pathway prediction we selected a symbiotic system with known nutritional provisioning requirements. The reduced genomes of *Candidatus Moranella endobia* and *Candidatus Tremblaya princeps* (GenBank NC-015735 and NC-015736), bacterial endosymbionts of the mealybug *Planococcus citri* have been previously described by McCutcheon and colleagues to distribute biosynthetic pathways for essential amino acids in a process known as "inter-pathway complementarity." Environmental PGDB construction using the combined *Moranella* and *Tremblaya* genomes recovered 43 out of 44 reactions and all 9 distributed amino acid biosynthesis pathways previously reported **(Figure 3 and Figure S3 in Additional File 1)**. Given these results, combinatorial ePGDB construction has enormous potential to predict distributed metabolic pathways within defined microbial assemblages e.g., co-cultures or more complex microbial communities.

Comparative Community Metabolism

To evaluate Pathway Tools' performance on complex microbial communities at different information levels we compared and contrasted coupled metagenome (DNA) and metatranscriptome (RNA) datasets from 25, 75, 110 m (sunlit or euphotic) and 500 m (dark) ocean depth intervals from HOT [18]. A total of 1026 unique pathways from approximately 1.2 billion base pairs of environmental sequence information were recovered spanning defined environmental gradients including luminosity, salinity, pressure, and oxygen concentration **(Table S12 in Additional File 1)**. Of these pathways, 840 met minimal quality control (QC) standards **(Materials and Methods)** and were used for subsequent set-difference analysis **(Figure 4a)**.

More than 600 pathways were shared in common between the sunlit and dark ocean based on combined DNA and RNA datasets consistent with a conserved metabolic core **(Figure 4b)**. A total of 14 unique pathways were predicted exclusively in sunlit samples with 20 pathways predicted at the intersection of 25, 75 and 110 m depth intervals **(Figure 4b)**. More than 100 unique pathways were predicted for the 500 m compliment consistent with increased metabolic potential and niche-specialization with increasing depth **(Figure 4b)**. Interestingly, the normalized proportion of genetic potential (DNA) versus expressed metabolic pathways (DNA/RNA) increased linearly between 25, 75 and 110 m depth intervals (0.4, 0.7 and 1.2, respectively) before plateauing at 500 m (1.2) **(Figure 4c)**. It remains to be determined if this trend reflects an asymptote or an inflection point in pathway expression co-varying as a function of metabolic status, environmental conditions or sample coverage and QC.

A total of 30 pathways were identified exclusively in RNA datasets including 11 pathway variants **(Figure 4c and Figure S4 in Additional File 1)**. Expressed cholesterol degradation and tetrahydrobiopterin biosynthesis I were common to all depth intervals. Unique expressed photorespiration and glycolate degradation III pathways were recovered at 25 and 75 m, while ammonia oxidation III, methane oxidation to methanol II, and arginine biosynthesis III were unique to 500 m **(Figure S4 in Additional File 1)**. More than 590 pathways were identified exclusively in DNA datasets, while 495 were shared in common between DNA and RNA datasets **(Figure 4d)**. With respect to functional classes, unique Degradation, Biosynthesis and Energy-Metabolism pathways increased as a function of depth in DNA datasets (**Figure S5a in Additional File 1)**. Within unique degradation classes a progression from amino acids to aromatic-compounds and secondary metabolites was observed between 25, 75, 110 and 500 m depth intervals. A similar progression was observed for a subset of Biosynthetic classes

including polyamines, lipids, and cofactors and for Energy-Metabolism including C1-compounds and fermentation (**Figure S5b in Additional File 1)**.

An evaluation of the 72 most abundant pathways recovered from the combined datasets indicated that 53 were both present and expressed at 25, 75, 110, and 500 m depth intervals. Moreover, several of the most abundant pathways including ammonium transport, Rubisco shunt, NADH to cytochrome electron transfer, pyruvate fermentation, denitrification, Calvin-Benson-Bassham cycle, cysteine biosynthesis I and arginine biosynthesis III exhibited depth-dependent trends in gene expression **(Figure S6 in Additional File 1)**. A number of abundant pathways common to 25, 75, 110, and 500 m depth intervals in the DNA datasets were exclusively expressed in sunlit or dark ocean waters **(Figure 5)**. In sunlit waters these included photosynthesis light reactions, hydrogen production VIII, flavonoid biosynthesis, cofactors including heme, vitamin B-complex (thiamin, adenosylcobalamin), and glutathione for oxidative stress **(Figure 5)**. Below the euphotic zones, the 500 m depth interval exclusively expressed pathways for ribotol, rhamnose, guanosine nucleotide, 2-methylcitrate, and threonine degradation as well as pathways for cofactor biosynthesis including phosphopantothenate, menaquinol-8 (vitamin K), and coenzyme M and several carbohydrate and amino acid biosynthetic pathways including CMP-N-acetylneuraminate I, ADP-L-glycero-beta-D-manno-heptose and glycine biosynthesis IV **(Figure 5)**.

Consistent with previous reports, sunlit waters expressed many photosynthesis-related pathways including aerobic electron transfer, hydrogen production, and cofactors including ubiquinol, heme, vitamin B-complex (nicotinate, thiamine, cobalamin, tetrahydrofolate), chlorophyll a, and retinol biosynthesis [18,19] **(Figures S7 and S8 in Additional File 1)**. In addition to photosynthesis, 25 and 75 m depth intervals (upper euphotic) sets included pathways

associated with degradation of plant metabolites including phytate, glucuronate, mannitol, chitin, xylose, arabinose, gallate, and quinolate. Other pathways of interest identified in sunlit waters included organophosphate, urea, and aminobutyrate degradation, as well as pathways for conversion of the plant hormone indole-3 acetic acid and mercury detoxification. Below the euphotic zone, the 500 m depth interval expressed unique pathways for intra-aerobic nitrite reduction, dissimilatory nitrate reduction, the reductive monocarboxylic acid cycle, ammonia oxidation, and methane oxidation to methanol I **(Figure S9 in Additional File 1)**. Thus, comparative ePGDB analysis using the combined DNA and RNA datasets differentiated between genomic potential and phenotypic expression across defined environmental gradients in the ocean and revealed known and novel patterns of functional specialization with potential implications for nutrient and energy flow within sunlit and dark ocean waters.

Pathway Prediction Hazards

While the construction of ePGDBs promotes pathway-centric analysis of environmental sequence information, prediction hazards need to be considered for optimal interpretive power. One common hazard relates to the 'multiple mapping problem,' arising when an enzyme catalyzes conserved or promiscuous reaction steps across multiple pathways. This problem typically manifests in pathway variants sharing a number of common or reversible reaction steps and has been described previously by Caspi and colleagues in the context of PGDB construction for cellular organisms [26]. For example, the tricarboxylic acid cycle (TCA) cycle has at least 8 pathway variants associated with different taxonomic groups and several incomplete or reversible forms that share multiple reactions steps. Pathologic has difficulty differentiating between TCA cycle variants when reversible pathway components are present even when a

diagnostic step such as ATP-citrate lyase for the reductive TCA cycle is missing from the input data. A related problem occurs when a regulatory protein is used to provide evidence that a pathway exists even when catalytic pathway components are missing from the input. Given that ePGDBs are constructed without taxonomic pruning and that Pathologic uses automated annotations from multiple taxonomic groups when predicting pathways from environmental sequence information, taxon specific pathways such as plant hormone biosynthesis or innate immunity can be predicted even when organisms known to encode such pathways are absent from the dataset. The extent to which these predictions reflect previously unrecognized pathway variants remains to be determined. Moreover, this hazard has the potential to confound distributed metabolic pathway identification when sequence coverage is low or microbial community composition is extremely uneven.

The identification of dissimilatory nitrate reduction (denitrification), intra-aerobic nitrite reduction and ammonia oxidation in the combined 500 m HOT DNA and RNA datasets provides a real world example of hazard navigation. Denitrification is a distributed form of energy metabolism resulting in the production of nitrogen gas in oxygen-deficient waters (< 20 μM O_2 per kg) [27,28]. The first step in denitrification is nitrate reduction to nitrite. In the combined HOT DNA and RNA datasets the predicted pathway variant nitrate reduction IV included a subset of ORFs/transcripts for 'nitrate reductase gamma subunit' (24 in DNA, 79 in RNA) while the predicted pathway variant nitrate reduction I included ORFs/transcripts for multiple nitrate reductase subunits **(Figure 6)**. While open reading frames for nitrate reductase subunits originated from a number of different taxa including Alphaproteobacteria, Gammaproteobacteria, Nitrospira and Planctomycetes, 435 out of 523 (83%) predicted nitrate reductase transcripts originated from Nitrospira and Planctomycetes consistent with a role in

nitrite oxidation [29-32] **(Figure 6)**. The second step in denitrification is nitrite reduction to nitric oxide. Within the DNA dataset both bacterial and archaeal ORFs encoding nitrite reductase were recovered while transcripts originating from ammonia oxidizing archaea dominated the RNA dataset **(Figure 6)**. Open reading frames/transcripts for downstream pathway components including nitric oxide reductase and nitrous oxide reductase were not detected, although CbbQ/NirQ/NorQ family regulators necessary for inorganic carbon fixation in the Calvin-Benson-Bassham cycle, nitrite and nitric oxide reduction were identified in DNA and RNA datasets [33] **(Figure 6)**. Given that the mean oxygen concentration at 500 m is ~120 μM O_2 per kg [17,19], these results are consistent with active water column nitrite and ammonia oxidation processes. Recent studies in the Eastern Tropical South Pacific OMZ observed changes in the frequency distribution of denitrification genes between free-living (0.2-1.6 μm) and particle-associated (>1.6 μm) size fractions, with nitric oxide reductase and nitrous oxide reductase encoding genes enriched on particles [34]. The extent to which denitrification or anammox processes partition between free-living and particle-associated microoganisms in the HOT water column remains to be determined.

CONCLUSIONS

While advances in high throughput sequencing technologies are rapidly giving rise to tens of thousands of environmental datasets, the computational and analytic powers needed to organize, interpret and mobilize these datasets have lagged behind. Conventional BLAST-based annotation methods combined with gene-centric analyses tend to overlook the network properties of microbial communities driving ecological and biogeochemical interactions. We argue that pathway-centric analyses via the MetaPathways pipeline and Pathway Tools provides the

scientific user community with an end-to-end solution for comparing ePGDBs constructed from environmental sequence information revealing known and novel network properties. As with any automated analysis, this method is no replacement for manual curation. Indeed, we have highlighted specific instances where idiosyncratic annotation, multifunctional enzymes, regulatory functions, and reversible enzymatic forms predicted by Pathway Tools result in interpretive hazards that require expert knowledge to resolve.

Continued development efforts are needed to improve on existing features and add new functionality to both the MetaPathways pipeline and Pathway Tools. Specifically, improved import features amenable to categorical variables or metadata e.g., taxonomic origin, location, depth, etc need to be integrated with the 'groups' features, a feature that enables users to integrate external data to group pathways and objects within Pathway Tools. The 'groups' feature in turn needs to be better integrated into the 'omics' viewer allowing for improved pathway navigation and page summaries within the Pathway Tools browser. Tooltip enhancements that include categorical data mentioned above could further enhance the browsing experience. Current ePGDBs are constructed using concatenated ORF sequences and improved viewing features are needed that map coverage and noncoding sequence information onto complete contigs. Finally, new pruning rules are needed that incorporate taxonomic affiliation for individual pathways or pathway components.

Despite current limitations, ePGDBs provide an interactive and holistic data structure in which to investigate distributed metabolism and differentiate between microbial community genetic potential and phenotypic expression. Thus, ePGBDs provide a functional blueprint of microbial community metabolism that can be harnessed to construct microbial consortia with defined emergent properties. These properties can in turn be transferred to industrial strains or

modeled using MetaFlux to improve process performance [13]. Although the set-difference and visual inspection methods used to identify distributed metabolic pathways described here do not scale for big datasets, future algorithmic improvements will enable comparisons of reference genomes and metagenomes in large numbers. Indeed, splitting the proverbial "reaction arrows" for each step in a given metabolic pathway into taxonomic bins provides a basis for integer optimization methods that compute "distribution" scores and a baseline for monitoring changes in the reaction network associated with environmental change or even human health status. Looking forward, we envision an open source collection of ePGDBs analogous to BioCyc[15] that can be queried and compared online revealing the network properties of microbial communities in natural and human engineered ecosystems on a truly global scale.

MATERIALS AND METHODS

Metabolic Pathway Analysis

Environmental PGDBs were constructed from public datsets using MetaPathways [14] with default parameter settings: open reading frame (ORF) detection by Prodigal (minimum length 60 amino acids), functional annotation by BLAST (e-value 1e-5, blast-score ratio 0.4) against protein databases KEGG [35], COG [36], MetaCyc [11] (version 16.0), and RefSeq [37] (Downloaded August 2012), and pathway prediction via the PathoLogic algorithm with taxonomic pruning disabled. Predicted pathways and associated ORFs were extracted from created ePGDBs using the utility script extract_pathway_table_from_pgdb.pl included with MetaPathways.

Pathway Prediction on Simulated Data

Simulated sequencing experiments were performed using MetaSim [21] with the parameter settings: Long read: clone size 36000bp, Gaussian error, mean read length 700bp, standard deviation 100bp; Short read: Gaussian error, mean 160bp, standard deviation 40bp) against the *E. coli K12 MG1655* complete nucleotide genome (GenBank: NC_000913) at a series of fractional levels (1/32, 1/16, 1/8, 1/4, 1/2, 1/1) of the total combined length of starting component genomes (G_m). Pathways were predicted using the MetaPathways pipeline, as described above, against each of the resulting sequence sets. A classification performance analysis was performed; True positives (TP) were pathways found in both the simulated sample pathways (test set) and the complete gold standard *E. coli* genome. True negatives (TN) were pathways not predicted in the test set or gold standard. False positives (FP) were pathways found in the test set but not in the gold standard. Finally, false negatives (FN) were pathways found in the gold standard but not in the test set. Multiple summary statistics for the resulting confusion tables (Sensitivity (Recall), Specificity, Precision, Accuracy, F-measure, and Matthew's Correlation Coefficient (MCC)) were calculated. A summary of these performance statistics is provided in the supplement **(Note S1: 'A Note on Confusion Table Statistics' in Additional File 1)**.

Simulated Metagenomes: Sim1, Sim2

Simulated sequencing experiments of metagenomes Sim1 and Sim2 were generated and analyzed as described above for *E. coli*. To minimize name mapping problems, we used prokaryotic genomes from the tier-2 BioCyc database collection [20]. The Sim1 metagenome was composed of ten tier-2 BioCyc genomes **(Table S2 in Additional File 1)** in equal copy number, while Sim2 was composed of the *Caulobacter cresentus NA1000* genome in 20-fold

excess relative to other genomes **(Figure S1 in Additional File 1)**. A classification performance analysis was performed as described above with the set of 646 pathways predicted from the complete tier-2 genomes used to derive Sim1 and Sim2 representing the gold standard.

Simulated Metagenomes: HOT (25m)

A 25m metagenome from the Hawaii ocean time series was sub-sampled with replacement to different factional levels (1/20, 1/10, 3/20, 1/5, 2/5, 3/5, 4/5, and 1/1) and pathways were predicted as described above. Similarly, a classification performance analysis was performed with the set of 864 pathways predicted from the complete 454 run representing the gold standard.

Distributed metabolic pathway prediction

Four genomes of similar size and complexity from the tier-2 dataset were combined in a pairwise manner: *Aurantimonas manganoxydans SI85-9A* (GenBank: NZ_AAPJ00000000.1), *Bacillus subtilis subtilis 168* (GenBank: AL009126.3), *Caulobacter crescentus NA1000* (GenBank: CP001340.1), and *Helicobacter pylori 26695* (GenBank: AE000511.1), abbreviated by the first character of their proper names, A, B, C, and H, respectively. The six pair-wise and four original genomes were analyzed as described above for *E. coli*. Pathways predicted in the combined PGDBs were considered candidates for distributed metabolism if they were absent from PGDBs for individual genomes (i.e., found in A and B combined, but not in either A or B individually). Candidate pathways were manually inspected and deemed 'plausible' if there was sufficient coverage, i.e., 75% of reactions in a pathway had associated ORFs from both taxa **(Figure S2 in Additional File 1)**.

Similarly, the *Candidatus Moranella endobia* and *Candidatus Tremblaya princeps* genomes (GenBank: NC-015735 and NC-015736) were downloaded from NCBI and analyzed as described above for *E. coli*. Resulting PGDBs for individual and combined genomes were manually inspected for amino acid biosynthetic pathways described in McCutcheon and Dohlen [16].

Hawaii Ocean Time-series

Unassembled metagenomic and transcriptomic pyrosequences from the Hawaii Ocean Time-series (10m, 75m, 110m, and 500m) were obtained from the NCBI Sequence Read Archive (SRA_Accession: SRX007372, SRX007369, SRX007370, SRX007371, SRX016893, SRX016897, SRX156384, SRX156385) and run through the MetaPathways pipeline using default settings **(Additional File 3)**. Only pathways with more than ten mapped ORFs in an individual sample were used in downstream analysis **(Figure 4a, Additional File 4)**. Pathway ORF counts for each sample were normalized to the total number of unannotated ORFs in each dataset. Count data was then converted to percentages providing relative ORF abundance for each pathway **(Additional File 5)**. Relative ORF abundance of the top-40 pathways from DNA and RNA datasets were compared **(Figure S6 in Additional File 1)**. In addition, pathways predicted in the DNA and RNA datasets were compared at each depth interval to give simple-wise factions for each depth e.g., DNA-only, DNA-RNA, and RNA-only **(Figure 4c)**. Given the small number of pathways in the RNA-only sets no set-difference analysis was needed **(Figure S4 in Additional File 1)**. The DNA-only sets were declined and tabulated at various levels of the MetaCyc pathway hierarchy **(Figure S5 in Additional File 1)**. A final four-way set analysis was performed on the DNA-only and DNA-RNA pathways at each depth **(Figure 4d, Additional**

Files 6 and 7). DNA-RNA set-difference subsets with more than 5 predicted pathways were compared in detail (**Figures S7-S12 in Additional File 1)**. All data transformations, set operations, and comparisons were performed in the R statistical environment (http://www.r-project.org), and visualized using the ggplot graphical package (http://ggplot2.org) and d3.js graphical library (http://d3js.org/).

COMPETING INTERESTS

The authors are unaware of any competing interests.

AUTHORS' CONTRIBUTIONS

NWH conducted simulated metagenome and distributed metabolism experiments, comparative ePGDB analysis and co-wrote the paper. KMK provided pipeline and high performance computing support and co-wrote the paper. AKH participated in interpreting nitrogen cycle hazards in the HOT datasets and provided essential feedback on data products. TA and PDK provided computational and interpretive support related to MetaPathways, PathoLogic, and the MetaCyc database. SJH supervised the group, participated in data interpretation, provided essential feedback on data products, integration and formatting, and co-wrote the paper.

ACKNOWLEDGEMETNS

This work was carried out under the auspices of Genome Canada, Genome British Columbia, Genome Alberta, the Natural Science and Engineering Research Council (NSERC) of Canada, the Canadian Foundation for Innovation (CFI) and the Canadian Institute for Advanced Research (CIFAR) through grants awarded to S.J.H. The Western Canadian Research Grid (WestGrid)

provided access to high-performance computing resources. KMK was supported by the Tula Foundation funded Centre for Microbial Diversity and Evolution (CMDE). NWH was supported by a four year doctoral fellowship (4YF) administered through the UBC Graduate Program in Bioinformatics. We would like to thank Suzanne Paley, Ron Caspi, and Quang Ong of SRI International for their patience, technical support, and lucid discussions on the function of Pathway Tools and the PathoLogic algorithm, Antoine Pagé for his participation in preliminary performance evaluations and all members of the Hallam Lab for helpful comments along the way.

References

1. Falkowski PG, Fenchel T, Delong EF: **The microbial engines that drive Earth's biogeochemical cycles**. *Science* 2008, **320**:1034–1039.
2. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol. Mol. Biol. Rev. 2005*, **69**:195–195.
3. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS: **Genomic sequencing of single microbial cells from environmental samples**. *Current Opinion in Microbiology 2008*, **11**:198–204.
4. Wooley JC, Ye Y: **Metagenomics: Facts and Artifacts, and Computational Challenges**. *J Comput Sci Technol 2009*, **25**:71–81.
5. Hey AJ, Tansley S, Tolle KM: *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research; 2009.
6. Ye Y, Doak TG: **A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes**. *PLoS Comput Biol 2009,* **5**:e1000465.
7. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, *et al.*: **Metabolic reconstruction for metagenomic data and its application to the human microbiome**. *PLoS Comput Biol 2012,* **8**:e1002358.
8. Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, *et al.*: **Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset**. *Microbial Informatics and Experimentation 2011,* **1**:4.
9. Altman T, Travers M, Kothari A, Caspi R, Karp, PD: **A systematic comparison of the MetaCyc and KEGG pathway databases**. *BMC Bioinformatics 2013,* **14**:112.
10. Karp PD, Paley S, Romero P: **The pathway tools software**. *Bioinformatics 2002*, **18**:S225–S232.
11. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, *et al.*: **MetaCyc: a multiorganism database of metabolic pathways and enzymes**. *Nucleic Acids Research 2006*, 34:D511–6.
12. Karp PD, Latendresse M, Caspi R: **The pathway tools pathway prediction algorithm**. *Stand Genomic Sci 2011*, **5:**424–429.
13. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion of flux balance models from pathway databases**. *Bioinformatics 2012*, **28**:388–396.
14. Konwar KM, Hanson NW, Pagé AP, Hallam SJ: **MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information**. *BMC Bioinformatics 2013* **14**:202.
15. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, *et al.*: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes**. *Nucleic Acids Research 2005*, **33**:6083–6089.
16. McCutcheon JP, von Dohlen CD: **An interdependent metabolic patchwork in the**

nested symbiosis of mealybugs. *Curr. Biol. 2011*, **21**:1366–1372.

17. Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, *et al.*: **Community genomics among stratified microbial assemblages in the ocean's interior**. *Science 2006*, **311**:496–503.

18. Stewart FJ, Sharma AK, Bryant JA, Eppley JM, Delong EF: **Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities**. *Genome Biol 2011*, **12**:R26.

19. Shi Y, Tyson GW, Eppley JM, Delong EF: **Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean**. *ISME J 2011*, **5**:999–1013.

20. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction**. *BMC Bioinformatics 2010*, **11**:15.

21. Richter, DC, Ott F, Auch AF, Schmid R, Huson DH: **MetaSim—A sequencing simulator for genomics and metagenomics**. *PLoS ONE 2008*, **3**:e3373.

22. Cordero OX, Ventouras L-A, Delong EF, Polz MF: **Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations**. *Proc. Natl. Acad. Sci. U.S.A 2012*, **109**:20059–20064.

23. Ellers J, Toby Kiers E, Currie CR, McDonald BR, Visser B: **Ecological interactions drive evolutionary loss of traits**. *Ecol Lett 2012*, **15**:1071–1082.

24. Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, *et al.*: **Species interactions alter evolutionary responses to a novel environment**. *PLoS Biol 2012*, **10**: e1001330.

25. Morris JJ, Lenski RE, Zinser ER: **The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss**. *MBio 2012*, **3**:e00036–12.

26. Caspi R, Dreher K, Karp PD: **The challenge of constructing, classifying, and representing metabolic pathways**. *FEMS Microbiol Lett 2013*, **345**:85–93.

27. Lam P, Kuypers MMM. **Microbial nitrogen cycling processes in oxygen minimum zones**. *Ann Rev Mar Sci 2011*, **3**:317–345.

28. Wright JJ, Konwar KM, Hallam SJ: **Microbial ecology of expanding oxygen minimum zones**. *Nat. Rev. Microbiol 2012*, **10**:381–394.

29. Ehrich S, Behrens D, Lebedeva E, Ludwig W, Bock E: **A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium, Nitrospira moscoviensis sp. nov. and its phylogenetic relationship**. *Arch. Microbiol. 1995*, **164**:16–23.

30. Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, *et al.*: **Deciphering the evolution and metabolism of an anammox bacterium from a community genome**. *Nature 2006*, **440**:790–794.

31. Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, *et al.*: **A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria**. *Proc. Natl. Acad. Sci. U.S.A. 2010*, **107**:13479–13484.

32. Kartal B, Maalcke WJ, de Almeida NM, Cirpus I, Gloerich J, *et al.*: **Molecular**

mechanism of anaerobic ammonium oxidation. *Nature 2011*, **479**:127–130.

33. Zumft WG: **Cell biology and molecular basis of denitrification**. *Microbiol. Mol. Biol. Rev. 1997*, **61**:533–616.

34. Ganesh S, Parris DJ, Delong EF, Stewart FJ: **Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone**. *ISME J 2013*, doi:10.1038/ismej.2013.144

35. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research 2000*, **28**:27–30.

36. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, *et al.*: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. *Nucleic Acids Research 2001*, **29**:22–28.

37. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Research 2007*, **35**:D61–5.

Illustrations and Figures

Figure 1. A multi-tiered approach to ePGDB validation. (a) In the absence of highly curated and validated datasets, we took inspiration from the curation-tiered structure of available pathway/genome databases within the BioCyc family. **(b/c)** Through *in silico* simulated sequencing experiments on the *E. coli K12* genome and two simulated metagenomes, we evaluated the performance of the PathoLogic algorithm under changing sequence coverage and taxonomic distributions. **(d)** We reanalyzed the genomes of *Candidatus Moranella endobia* and *Candidatus Tremblaya princeps*, two symbiotic taxa with reduced genomes, sharing a number of essential amino acid pathways. **(e)** Finally, we predicted pathways from a previously analyzed paired metagenomic and metatranscriptomic dataset from the Hawaii Ocean Time-series to validate on previously identified pathways and metabolic functions.

Figure 2. Analysis on *in silico* simulated sequencing experiments across different levels of coverage, sequencing lengths, and taxonomic distributions. (a) Predicted pathway recovery as a percentage of the total pathways predicted from the full genomes. **(b)** Sensitivity (circles) and specificity (triangles) of predicted pathways of the *in silico* experiments using the pathways predicted on the full genomes as the gold standard.

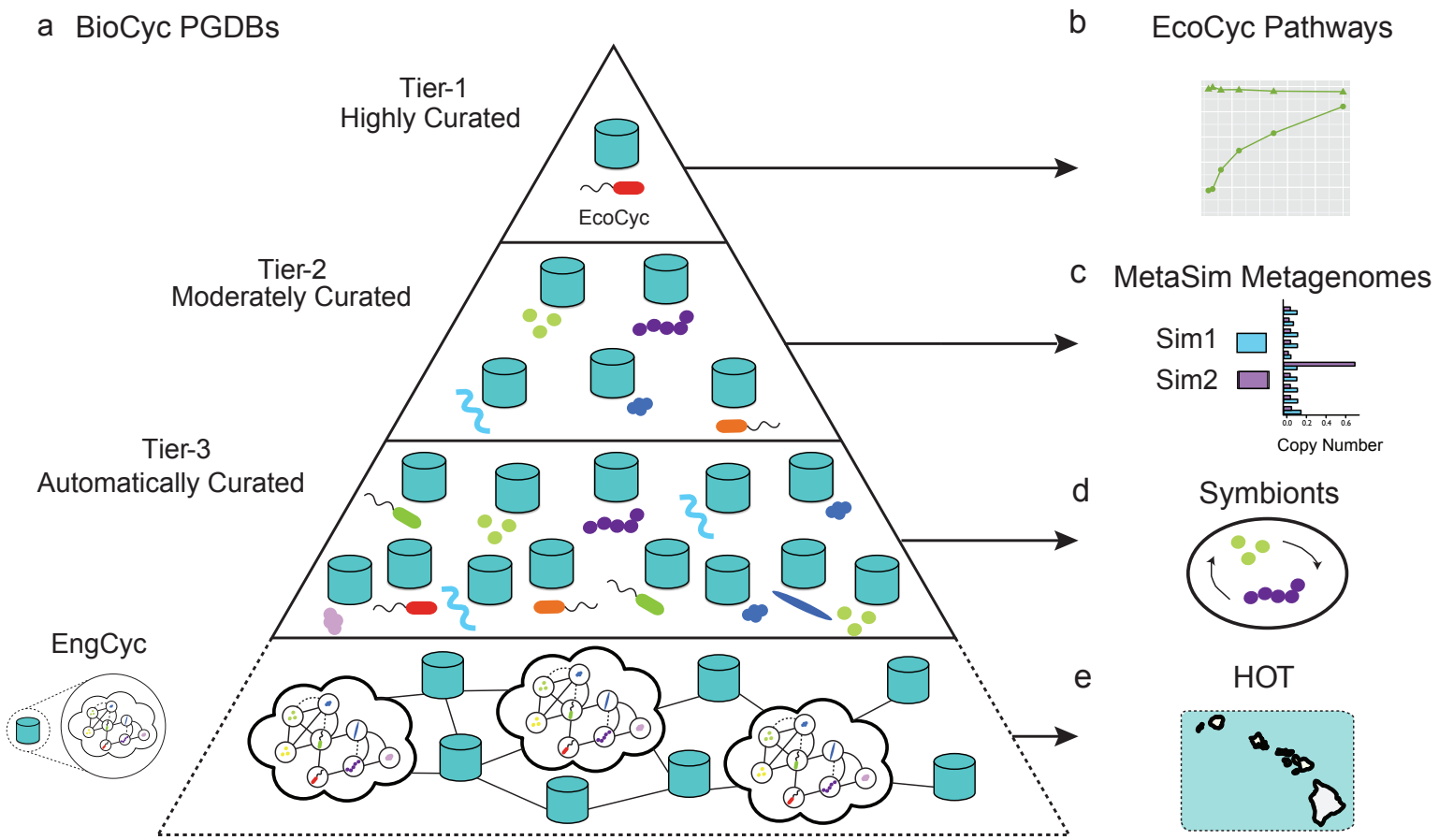
Figure 3. Examples of emergent amino acid metabolism shared between the *Moranella endobia* and *Tremblaya princeps* genomes. This figure illustrates examples of emergent metabolic pathways predicted between symbiotic prokaryotes *Candidatus Moranella endobia* and *Candidatus Tremblaya princeps*. Enyzmes found in *Moranella* (red), *Tremblaya* (blue), or both taxa (purple) are highlight in the pathway glyph diagrams, showing patterns of potentially emergent metabolism. A complete description of all amino acid pathways can be found in **Figure S3 in Additional File 1**.

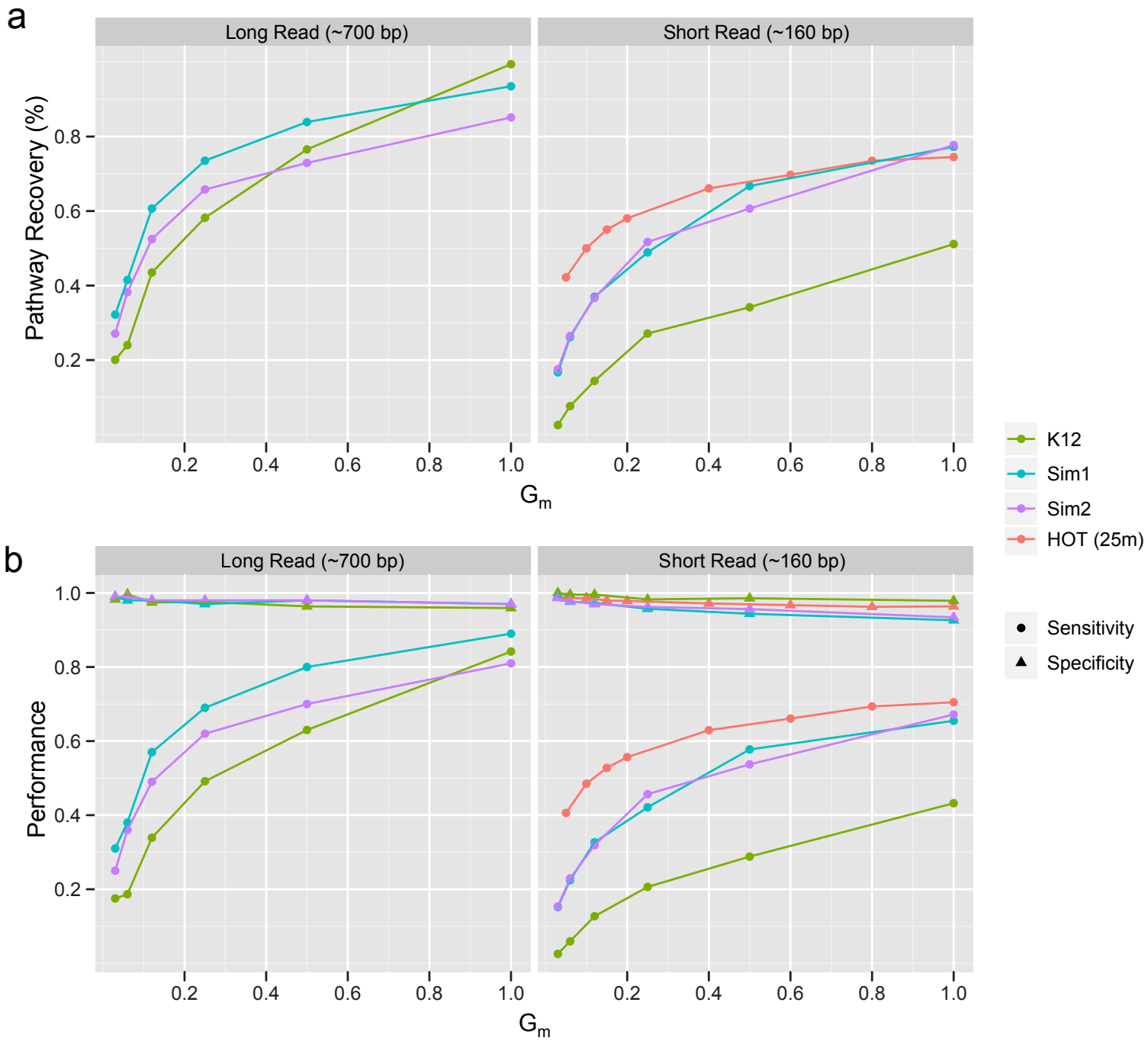
Figure 4. Analysis of predicted pathways from the Hawaii Ocean Time-series. (a) A total of 1033 unique pathways were predicted from the HOT samples **(Additional File 3)**, however only 840 unique pathways remained after all pathways in each sample with less than 10 ORFs were removed **(Additional File 4)**. **(b)** After normalizing by total predicted ORFs **(Additional File 5)**, a 4-way set analysis of these quality controlled (QC) pathways shows that there is a large core common pathways to the samples. **(c)** Separating unique pathways within the DNA and RNA of each sample revealed that very few pathways were unique to the RNA fraction of each sample. **(d)** Finally, at set analysis of the unique DNA fraction (light colors), and pathways common to

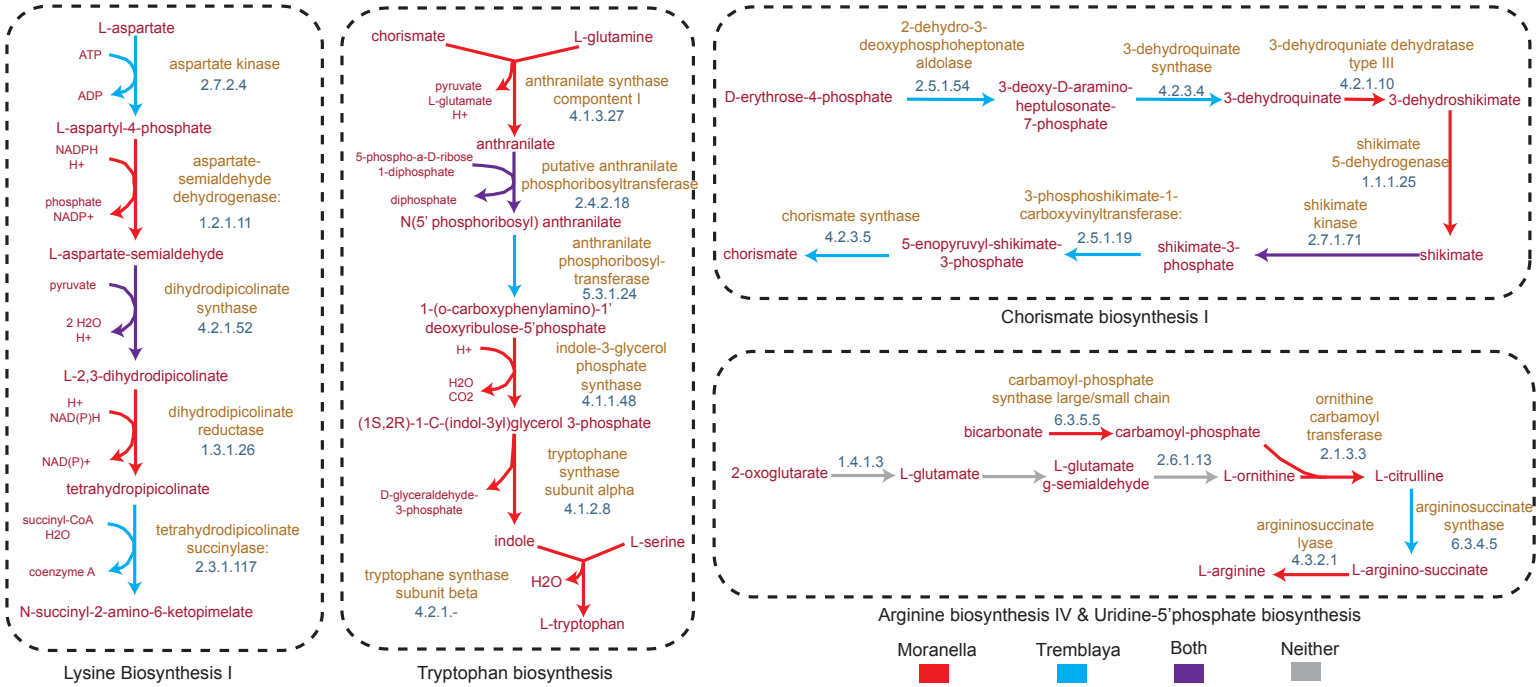
DNA and RNA from each sample (dark colors) found subsets of pathways unique to each fraction **(Additional Files 6 and 7)**.

Figure 5. Comparison of predicted genomic and transcriptomic pathways with unique expression in the 'sunlit' and 'dark' HOT samples. Sunlit metabolism was indicative photosynthesis and aerobic metabolism including photosynthesis light reactions and hydrogen production. Dark metabolism had significantly more degradation pathways and an anaerobic signal indicative of methanogenesis (coenzyme M biosynthesis).

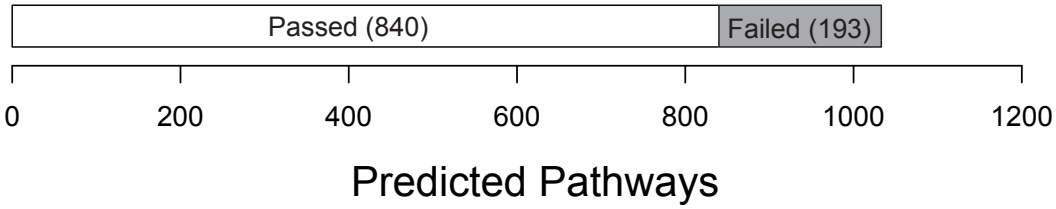
Figure 6. Taxonomic and functional breakdown of nitrogen cycling pathways. (**a**) Nitrogen cycling pathways and reactions assigned by PathoLogic. Arrow color indicates pathway, nitrate reduction I (denitrification) (brown), nitrate reduction IV (dissimilatory) (yellow), and intra-aerobic nitrite reduction (red). Grey numbers adjacent to arrows indicated number of reads assigned to the reaction in the DNA and RNA (RNA in parentheses). Overlapping circles indicate the distribution of reads across multiple pathways. **(b)** BLAST-based functional and taxonomic breakdown of reads assigned to reactions in given pathways as indicated by letters A-E. Function was determined by the top RefSeq BLAST hit, reported by the MetaPathways pipeline, and indicated by reaction arrows, with color corresponding to taxa or taxonomic group with known activity: taxa with nitrate and nitrite reducing activity (blue), nitrite oxidizing activity (green), and ammonia oxidizing activity (purple). Grey reactions indicate no reads for enzymatic activity were detected, only regulatory proteins which may be involved in gene expression regulation (*).


a BioCyc PGDBs
Tier-1
Highly Curated
EcoCyc
Tier-2
Moderately Curated
Tier-3
Automatically Curated
EngCyc
b
EcoCyc Pathways
c
MetaSim Metagenomes
Sim1
Sim2
0.0 0.2 0.4 0.6
Copy Number
d
Symbionts
e
HOT


a
Long Read (~700 bp)
Short Read (~160 bp)
Pathway Recovery (%)
0.2
0.4
0.6
0.8
1.0
G_m
b
Performance
K12
Sim1
Sim2
HOT (25m)
Sensitivity
Specificity

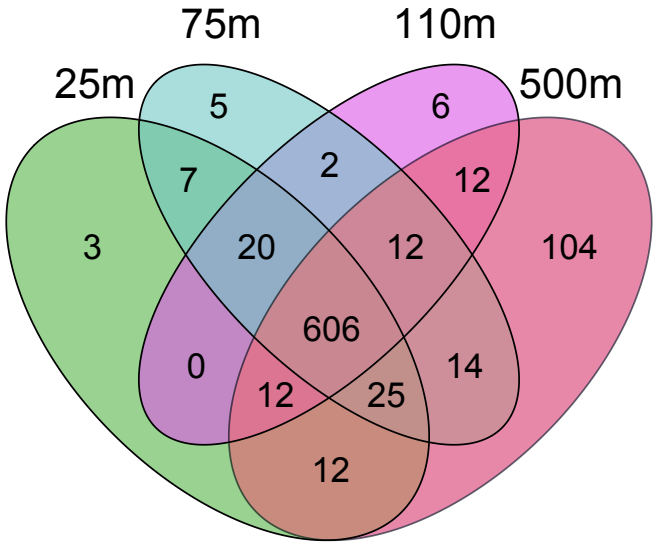

L-aspartate
ATP
ADP
aspartate kinase
2.7.2.4
L-aspartyl-4-phosphate
NADPH
H+
phosphate
NADP+
aspartate-semialdehyde dehydrogenase:
1.2.1.11
L-aspartate-semialdehyde
pyruvate
2 H2O
H+
dihydrodipicolinate synthase
4.2.1.52
L-2,3-dihydrodipicolinate
H+
NAD(P)H
NAD(P)+
dihydrodipicolinate reductase
1.3.1.26
tetrahydropipicolinate
succinyl-CoA
H2O
coenzyme A
tetrahydrodipicolinate succinylase:
2.3.1.117
N-succinyl-2-amino-6-ketopimelate
Lysine Biosynthesis I
chorismate
L-glutamine
pyruvate
L-glutamate
H+
anthranilate synthase compontent I
4.1.3.27
anthranilate
5-phospho-a-D-ribose 1-diphosphate
diphosphate
putative anthranilate phosphoribosyltransferase
2.4.2.18
N(5' phosphoribosyl) anthranilate
anthranilate phosphoribosyl-transferase
5.3.1.24
1-(o-carboxyphenylamino)-1'
deoxyribulose-5'phosphate
H+
H2O
CO2
indole-3-glycerol phosphate synthase
4.1.1.48
(1S,2R)-1-C-(indol-3yl)glycerol 3-phosphate
D-glyceraldehyde-3-phosphate
tryptophane synthase subunit alpha
4.1.2.8
indole
L-serine
tryptophane synthase subunit beta
4.2.1.-
H2O
L-tryptophan
Tryptophan biosynthesis
D-erythrose-4-phosphate
2-dehydro-3-deoxyphosphoheptonate aldolase
2.5.1.54
3-deoxy-D-aramino-heptulosonate-7-phosphate
3-dehydroquinate synthase
4.2.3.4
3-dehydroquinate
3-dehydroquniate dehydratase type III
4.2.1.10
3-dehydroshikimate
shikimate 5-dehydrogenase
1.1.1.25
shikimate
shikimate kinase
2.7.1.71
shikimate-3-phosphate
3-phosphoshikimate-1-carboxyvinyltransferase:
2.5.1.19
5-enopyruvyl-shikimate-3-phosphate
chorismate synthase
4.2.3.5
chorismate
Chorismate biosynthesis I
2-oxoglutarate
1.4.1.3
L-glutamate
L-glutamate g-semialdehyde
2.6.1.13
L-ornithine
carbamoyl-phosphate synthase large/small chain
6.3.5.5
bicarbonate
carbamoyl-phosphate
ornithine carbamoyl transferase
2.1.3.3
L-citrulline
argininosuccinate synthase
6.3.4.5
L-arginino-succinate
argininosuccinate lyase
4.3.2.1
L-arginine
Arginine biosynthesis IV & Uridine-5'phosphate biosynthesis
Moranella
Tremblaya
Both
Neither

a



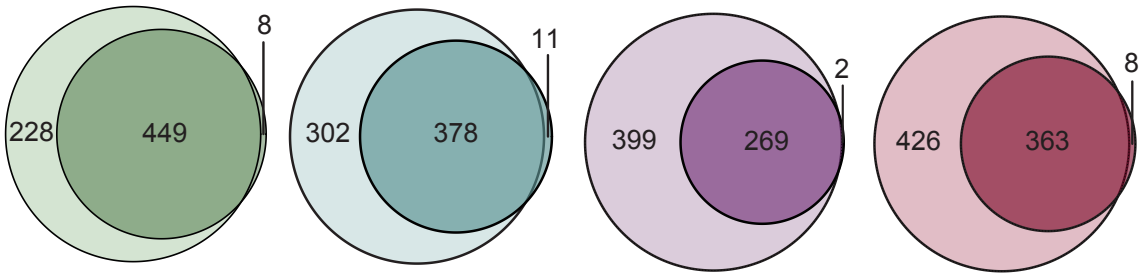
Predicted Pathways

b



QC Pathways (840)

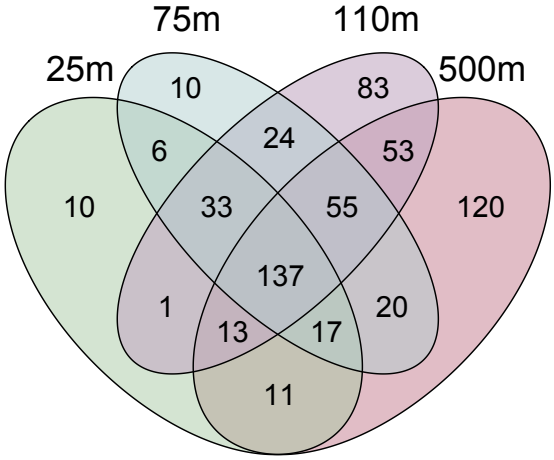
c



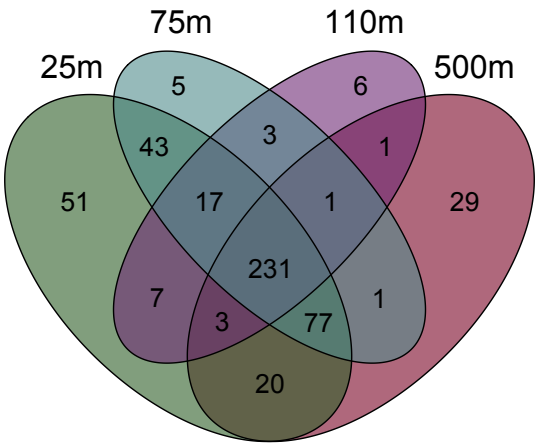
25m (685) 75m (691) 110m (670) 500m (797)

Sample-wise DNA/RNA

d

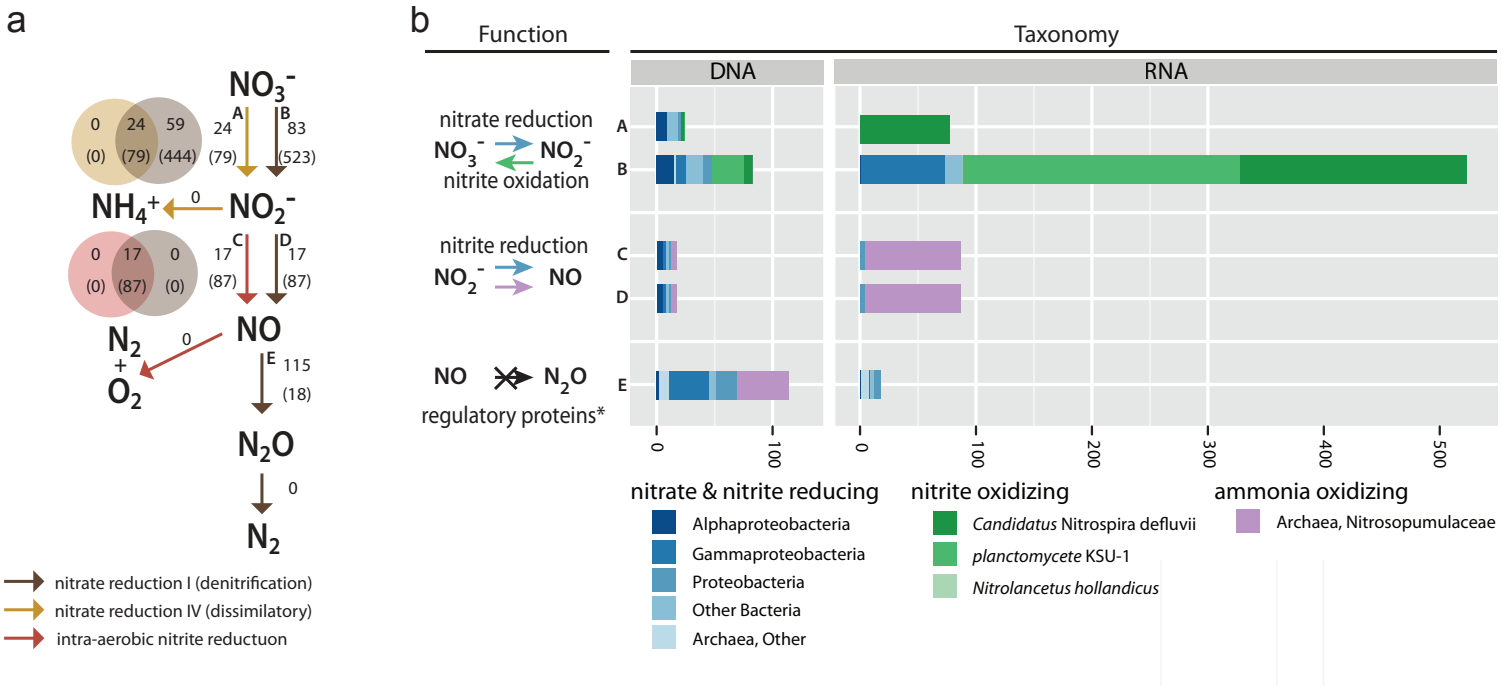


DNA Faction (593)



Shared DNA & RNA (495)


MetaCyc Pathways
Sunlit (25m, 75m, 110m) (RNA/DNA)
Dark (500m) (RNA/DNA)
Energy
Degradation
Biosynthesis
Photosynthesis
photosynthesis light reactions
Hydrogen production
hydrogen production VIII
Amino acids
methionine degradation II
Secondary metabolites
flavonoid biosynthesis
Cofactors
heme biosynthesis I (uroporphyrinogen-III)
thiamin diphosphate biosynthesis II
thiamin diphosphate biosynthesis I
adenosylcobalamin biosynthesis II
adenosylcobalamin biosynthesis I
thiamin diphosphate biosynthesis IV
lipoate biosynthesis and incorporation I
glutathione biosynthesis
biotin biosynthesis (7-keto-8-aminopelargonate)
trans-farnesyl diphosphate biosynthesis
UDP-N-acetyl-D-galactosamine biosynthesis II
Carbohydrates
glycogen biosynthesis I (ADP-D-Glucose)
Amino acids
homocysteine and cysteine interconversion
Energy
Degradation
Biosynthesis
Fermentation
(S)-acetoin biosynthesis
Secondary metabolites
ribitol degradation
sorbitol degradation I
Non-carbon nutrients
ammonia oxidation I (aerobic)
intra-aerobic nitrite reduction
nitrate reduction IV (dissimilatory)
Nucleotides
guanosine nucleotides degradation II
Carbohydrates
L-rhamnose degradation II
D-mannose degradation
Carboxylates
2-methylcitrate cycle II
acetate formation from acetyl-CoA II
citrate degradation
C1 compounds
reductive monocarboxylic acid cycle
methane oxidation to methanol I
Amino acids
threonine degradation II
threonine degradation III (to methylglyoxal)
Secondary metabolites
salidroside biosynthesis
diploterol and cycloartenol biosynthesis
Cofactors
phosphopantothenate biosynthesis III
menaquinol-8 biosynthesis
5,6-dimethylbenzimidazole biosynthesis
mycothiol biosynthesis
coenzyme M biosynthesis I
pyridoxal 5-phosphate biosynthesis II
coenzyme B/coenzyme M regeneration
Carbohydrates
CMP-N-acetylneuraminate biosynthesis I
ADP-L-glycero-beta-D-manno-heptose biosynthesis
Amino acids
selenocysteine biosynthesis II (archaea)
glycine biosynthesis IV
25m
75m
110m
500m
DNA
RNA
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
Relative ORF Abundance (%)


a
NO_3^-
A
B
0
24
59
(0)
(79)
(444)
24
(79)
83
(523)
NH_4^+
0
NO_2^-
C
D
0
17
0
(0)
(87)
(0)
17
(87)
17
(87)
N_2
+
O_2
0
NO
E
115
(18)
N_2O
0
N_2
nitrate reduction I (denitrification)
nitrate reduction IV (dissimilatory)
intra-aerobic nitrite reductuon
b
Function
Taxonomy
DNA
RNA
nitrate reduction
NO_3^- NO_2^-
nitrite oxidation
nitrite reduction
NO_2^- NO
NO N_2O
regulatory proteins*
A
B
C
D
E
0
100
0
100
200
300
400
500
nitrate & nitrite reducing
Alphaproteobacteria
Gammaproteobacteria
Proteobacteria
Other Bacteria
Archaea, Other
nitrite oxidizing
Candidatus Nitrospira defluvii
planctomycete KSU-1
Nitrolancetus hollandicus
ammonia oxidizing
Archaea, Nitrosopumulaceae

Additional File 1: Supplementary Material for Metabolic pathways for the whole community.

Table S1. Overview of the *E. coli K12* genome used for simulated sequencing experiments.

Taxa	GenBank	Size (bp)	Genes
Escherichia coli str. K-12 substr. MG1655	NC_000913	4,639,675	4,288

Table S2. Overview of the tier-2 BioCyc genomes used for simulated sequencing experiments.

Taxa	GenBank	Size (bp)	Genes
Agrobacterium tumefaciens C58	AE008687-AE008690	5,674,064	5,469
Aurantimonas manganoxydans SI85-9A1	NZ_AAPJ00000000.1	4,285,343	3,665
Bacillus subtilis subtilis 168	AL009126.3	4,215,606	4,428
Caulobacter crescentus CB15	AE005673	4,016,947	3,819
Caulobacter crescentus NA1000	CP001340.1	4,042,929	3,968
Helicobacter pylori 26695	AE000511.1	1,667,867	1,609
Mycobacterium tuberculosis CDC1551	AE000516	4,403,836	4,235
Mycobacterium tuberculosis H37Rv	AL123456	4,411,529	3,916
Synechococcus elongatus PCC 7942	NC_007604.1	2,695,903	2,664
Vibrio cholerae O1 biovar El Tor str. N16961	AE003852, AE003853	4,033,464	3,952
Total		39,447,488	37,725
Average		3,944,749	3,773

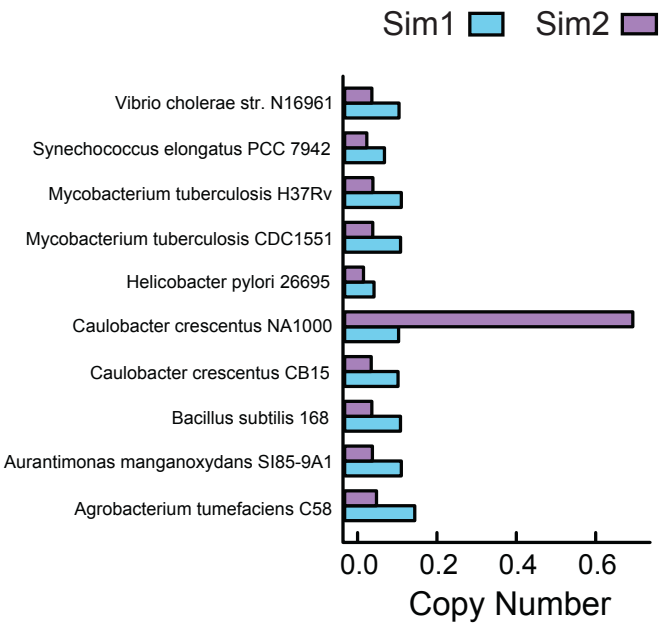


Figure S1. Copy number distributions for the simulated metagenomes Sim1 and Sim2. Sim1 (blue) has the ten BioCyc taxa in approximately equal proportion. Sim2 (purple) has the genome copy number of *Caulobacter crescentus NA1000* in approximately twenty times abundance. Taxa used were selected with approximately equal genome size and gene content.

Table S3. Overview of Long-read simulated sequencing experiments for *E. coli K12,* Sim1, and Sim2 taxonomic distributions.

Distribution	G_m	Size (bp)	Reads	ORFs	Annotated ORFs	Predicted Pathways	Recovered (%)
E. coli K12	0.03	143,339	180	243	150	71	0.20
E. coli K12	0.06	285,694	360	494	309	85	0.24
E. coli K12	0.12	575,821	720	973	604	154	0.44
E. coli K12	0.25	1,151,303	1,438	1,992	1,207	206	0.58
E. coli K12	0.50	2,305,899	2,876	4,030	2,401	271	0.77
E. coli K12	1.0	4,594,877	5,750	7,996	4,859	352	0.99
Sim1	0.03	1,249,104	1,564	2,978	771	208	0.32
Sim1	0.06	2,501,345	3,126	5,857	1,487	268	0.41
Sim1	0.12	5,003,783	6,250	11,769	3,120	392	0.61
Sim1	0.25	10,014,496	12,500	23,422	6,107	475	0.74
Sim1	0.5	19,991,551	25,000	47,304	12,139	542	0.84
Sim1	1.0	40,016,291	50,000	94,438	24,388	604	0.93
Sim2	0.03	1,245,781	1,562	2,946	760	175	0.27
Sim2	0.06	2,496,313	3,126	5,880	1538	247	0.38
Sim2	0.12	4,987,646	6,250	11,756	3154	339	0.52
Sim2	0.25	9,994,331	12,500	23,852	6330	425	0.66
Sim2	0.50	19,993,717	25,000	47,350	12,676	471	0.73
Sim2	1.0	40,006,531	50,000	94,366	25,139	550	0.85

Table S4. Overview of Short-read simulated sequencing experiments for *E. coli K12,* Sim1, Sim2, and HOT (25m) taxonomic distributions.

Distribution	G_m	Size (bp)	Reads	ORFs	Annotated ORFs	Predicted Pathways	Pathways Recovered (%)
E. coli K12	0.03	139,983	540	179	37	9	0.03
E. coli K12	0.06	292,642	1,125	356	68	27	0.08
E. coli K12	0.12	584,031	2,250	742	128	51	0.14
E. coli K12	0.25	1,168,390	4,500	1,445	269	96	0.27
E. coli K12	0.50	2,340,834	9,000	2,878	516	121	0.34
E. coli K12	1.0	4,676,245	18,000	5,884	1,013	181	0.51
Sim1	0.03	1,283,742	4,738	4,151	2,261	108	0.17
Sim1	0.06	2,570,031	9,476	8,266	4,576	169	0.26
Sim1	0.12	5,140,469	18,975	16,549	9,132	239	0.37
Sim1	0.25	10,271,637	37,904	33,164	18,270	316	0.49
Sim1	0.50	20,540,345	75,808	66,260	36,443	431	0.67
Sim1	1.0	41,097,945	151,616	132,577	72,937	499	0.77
Sim2	0.03	1,282,621	4,738	4,337	2,666	113	0.17
Sim2	0.06	2,567,379	9,476	8,657	5,193	171	0.26
Sim2	0.12	5,133,838	18,952	17,313	10,496	237	0.37
Sim2	0.25	10,264,228	37,904	34,624	21,301	334	0.52
Sim2	0.50	20,545,013	75,808	69,256	41,901	392	0.61
Sim2	1.0	41,074,096	151,616	138,593	83,929	502	0.78
HOT (25m)	0.05	8,012,746	31,178	6,668	5,978	336	0.42
HOT (25m)	0.10	16,025,492	62,356	13,478	12,087	398	0.50
HOT (25m)	0.15	24,038,238	93,534	20,054	17,953	438	0.55
HOT (25m)	0.20	32,050,984	124,712	26,836	23,972	462	0.58
HOT (25m)	0.40	64,101,968	249,424	53,300	47,617	526	0.66
HOT (25m)	0.60	96,152,695	374,135	80,080	71,599	555	0.70
HOT (25m)	0.80	128,203,679	498,847	106,985	95,766	585	0.73
HOT (25m)	1.0	160,254,663	623,559	133,836	119,867	593	0.74

Note S1: Confusion Table Statistics
In machine learning a *confusion table* (*contingency table*) is a method to assess the performance of a supervised classifier. Rows of the table represent class predictions, while columns represent the actual class. Given a predicted class and the known class, there are four possible outcomes for the prediction:

Correct Responses
True Positives (TP) - The classifier correctly identified the class as present.
True Negatives (TN) - The classifier correctly identified the class as absent.

Incorrect Responses
False Positives (FP) (Type 1 Error) - The classifier incorrectly predicted the class present when absent.
False Negative (FN) (Type 2 Error) - The classifier incorrectly predicted the class absent when present.

		Actual Class		
		Positive	Negative	
Prediction	Positive	True Positives (TP)	False Positives (FP)	Precision TP / (TP + FP)
	Negative	False Negatives (FN)	True Negatives (TN)	Negative Predictive Value TN / (FN + TN)
		Sensitivity TP / (TP + FN)	Specificity TN / (FP + TN)	

Summary Statistics
Since classifiers can have very different performance characteristics it is often important to consider different statistics of the confusion table. In most situations, there is often a trade off between the two types of errors that a classifier can make.

Sensitivity (Recall) – Represents the ability of the classifier to find positive results. Given that a class is actually in the sample, what is the probability that it is found? High values represent a low number of false negatives (Type-II errors).

Sensitivity = (# correctly predicted present) / (# actually present)

Sensitivity = TP / (TP + FN)

Specificity - The ability of the test to find negative results. What is the probability of correctly rejecting a class. High values represent a low number of false positives (Type-I errors).

Specificity = (# predicted absent) / (# actually absent)

Specificity = TN / (FP + TN)

Precision - Given a positive prediction, what is the probability that it is correct? High values represent a low number of false positives (Type-I errors).

Precision = (# correctly predicted present) / (# predicted present)

Precision = TP / (TP + FP)

Negative Predictive Value (NPV) - Given a negative prediction, what is the probability of actually being correct? High values represent low false negatives (Type-2 errors).

NPV = (# correctly predicted absent) / (# predicted absent)

NPV = TN / (FN + TN)

Ideally one will investigate the confusion table directly, however, because it is onerous to compare many tables, a number of statistics have been developed to summarize the performance described in a table.

Accuracy is the most intuitive, but can be misleading if the distribution of positive and negative results are not of similar magnitude. It asks, of all the decisions that the classifier made, how many were correct?

Accuracy = (Number of correct predictions) / (Total Predictions)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

E.g. Which classifier is the best: A, B, C, or D?

	TP	*TN*	*FP*	*FN*	*Accuracy (%)*
A	25	75	25	75	50
B	0	150	0	50	75
C	50	0	150	0	25
D	30	100	50	20	65

F-measure is the harmonic mean between precision and sensitivity. Therefore, it represents the number of correctly predicted values scaled between false-positive and false-negative errors. However, it does not take into account the number of true-negative responses, which can be important depending on the application.

$$\text{F-measure} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}$$

Matthew's Correlation Coefficient (MCC) is a comprehensive measure that controls for the population differences between total positive and negatives in a test or training sample. Essentially it is a correlation coefficient between observed and predicted responses where +1 is perfect prediction, -1 is total disagreement, and 0 is no better than randomly guessing (i.e. no correlation).

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Matthews Correlation is generally accepted to be the best overall summary statistic of a confusion table if you want to give equal weight for correct positive and negative responses, as well as taking into account Type-1 and Type-2 errors.

Table S5. Overview of pathway prediction performance for simulated Long-read metagenomes of *E. coli K12,* Sim1, and Sim2 at progressively larger genomic sequence coverage.

Distribution	G_m	Precision	Sensitivity	Specificity	Accuracy	F-measure	Matthews
E. coli K12	0.03	0.93	0.18	1.00	0.83	0.31	0.83
E. coli K12	0.06	0.70	0.18	0.98	0.81	0.28	0.56
E. coli K12	0.12	0.76	0.32	0.97	0.84	0.45	0.64
E. coli K12	0.25	0.85	0.50	0.98	0.88	0.63	0.77
E. coli K12	0.50	0.81	0.62	0.96	0.89	0.71	0.74
E. coli K12	1.0	0.84	0.85	0.96	0.93	0.84	0.80
Sim1	0.03	0.96	0.31	0.99	0.73	0.47	0.79
Sim1	0.06	0.91	0.38	0.98	0.75	0.53	0.73
Sim1	0.12	0.94	0.57	0.98	0.82	0.71	0.81
Sim1	0.25	0.94	0.69	0.97	0.86	0.80	0.83
Sim1	0.50	0.95	0.80	0.98	0.91	0.87	0.88
Sim1	1.0	0.95	0.89	0.97	0.94	0.92	0.91
Sim2	0.03	0.93	0.25	0.99	0.70	0.40	0.74
Sim2	0.06	0.95	0.36	0.99	0.75	0.53	0.78
Sim2	0.12	0.93	0.49	0.98	0.79	0.64	0.78
Sim2	0.25	0.95	0.62	0.98	0.84	0.75	0.83
Sim2	0.50	0.97	0.70	0.98	0.88	0.81	0.87
Sim2	1.0	0.95	0.81	0.97	0.91	0.87	0.87

Table S6. Confusion tables of pathway prediction using simulated Long-read sequencing upon the *E. coli K12* genome, Sim1, and Sim2 at progressively larger genomic sequence coverage.

Distribution	G_m	TP	TN	FP	FN	P	N
E. coli K12	0.03	71	1453	5	317	76	1770
E. coli K12	0.06	69	1428	30	319	99	1747
E. coli K12	0.12	124	1418	40	264	164	1682
E. coli K12	0.25	195	1423	35	193	230	1616
E. coli K12	0.50	242	1402	56	146	298	1548
E. coli K12	1.0	328	1397	61	60	389	1457
Sim1	0.03	200	1023	8	446	208	1469
Sim1	0.06	244	1007	24	402	268	1409
Sim1	0.12	368	1007	24	278	392	1285
Sim1	0.25	446	1002	29	200	475	1202
Sim1	0.50	517	1006	25	129	542	1135
Sim1	1.0	576	1003	28	70	604	1073
Sim2	0.03	163	1019	12	483	175	1502
Sim2	0.06	235	1019	12	411	247	1430
Sim2	0.12	316	1008	23	330	339	1338
Sim2	0.25	403	1009	22	243	425	1252
Sim2	0.50	455	1015	16	191	471	1206
Sim2	1.0	521	1002	29	125	550	1127

Table S7. Overview of pathway prediction performance for simulated Short-read sequencing experiments of the *E. coli K12*, Sim1, Sim2, and HOT 25m metagenome at progressively larger genomic sequence coverage.

Distribution	G_m	Precision	Sensitivity	Specificity	Accuracy	F-measure	Matthews
E. coli K12	0.03	1.00	0.03	1.00	0.79	0.05	0.89
E. coli K12	0.06	0.78	0.06	1.00	0.80	0.11	0.64
E. coli K12	0.12	0.88	0.13	1.00	0.81	0.22	0.77
E. coli K12	0.25	0.76	0.21	0.98	0.82	0.32	0.64
E. coli K12	0.50	0.84	0.29	0.99	0.84	0.43	0.74
E. coli K12	1.0	0.85	0.43	0.98	0.86	0.57	0.76
Sim1	0.03	0.91	0.15	0.99	0.67	0.26	0.69
Sim1	0.06	0.86	0.22	0.98	0.69	0.36	0.64
Sim1	0.12	0.88	0.33	0.97	0.72	0.48	0.69
Sim1	0.25	0.86	0.42	0.96	0.75	0.57	0.67
Sim1	0.5	0.87	0.58	0.94	0.80	0.69	0.71
Sim1	1.0	0.85	0.65	0.93	0.82	0.74	0.70
Sim2	0.03	0.88	0.15	0.99	0.67	0.26	0.65
Sim2	0.06	0.87	0.23	0.98	0.69	0.36	0.65
Sim2	0.12	0.87	0.32	0.97	0.72	0.47	0.67
Sim2	0.25	0.88	0.46	0.96	0.77	0.60	0.71
Sim2	0.5	0.89	0.54	0.96	0.79	0.67	0.73
Sim2	1.0	0.86	0.67	0.93	0.83	0.76	0.73
HOT (25m)	0.05	0.96	0.41	0.99	0.71	0.57	0.75
HOT (25m)	0.10	0.97	0.48	0.99	0.75	0.65	0.78
HOT (25m)	0.15	0.96	0.53	0.98	0.77	0.68	0.78
HOT (25m)	0.20	0.96	0.56	0.98	0.78	0.70	0.78
HOT (25m)	0.40	0.95	0.63	0.97	0.81	0.76	0.80
HOT (25m)	0.60	0.95	0.66	0.97	0.82	0.78	0.80
HOT (25m)	0.80	0.94	0.69	0.96	0.83	0.80	0.80
HOT (25m)	1.0	0.95	0.70	0.96	0.84	0.81	0.81

Table S8. Confusion tables of pathway prediction for simulated Short-read sequencing experiments of the *E. coli K12*, Sim1, Sim2, and the HOT 25 m metagenome at progressively larger genomic sequence coverage.

Distribution	G_m	TP	TN	FP	FN	P	N
E. coli K12	0.03	9	1323	0	345	9	1668
E. coli K12	0.06	21	1317	6	333	27	1650
E. coli K12	0.12	45	1317	6	309	51	1626
E. coli K12	0.25	73	1300	23	281	96	1581
E. coli K12	0.50	102	1304	19	252	121	1556
E. coli K12	1.0	153	1295	28	201	181	1496
Sim1	0.03	98	1021	10	548	108	1569
Sim1	0.06	145	1007	24	501	169	1508
Sim1	0.12	211	1003	28	435	239	1438
Sim1	0.25	272	987	44	374	316	1361
Sim1	0.5	373	973	58	273	431	1246
Sim1	1.0	423	955	76	223	499	1178
Sim2	0.03	99	1017	14	547	113	1564
Sim2	0.06	148	1008	23	498	171	1506
Sim2	0.12	206	1000	31	440	237	1440
Sim2	0.25	295	992	39	351	334	1343
Sim2	0.5	347	986	45	299	392	1285
Sim2	1.0	434	963	68	212	502	1175
HOT (25m)	0.05	323	868	13	473	336	1341
HOT (25m)	0.10	386	869	12	410	398	1279
HOT (25m)	0.15	420	863	18	376	438	1239
HOT (25m)	0.20	443	862	19	353	462	1215
HOT (25m)	0.40	501	856	25	295	526	1151
HOT (25m)	0.60	526	852	29	270	555	1122
HOT (25m)	0.80	552	848	33	244	585	1092
HOT (25m)	1.0	561	849	32	235	593	1084

Table S9. Taxonomic pruning pathway recovery results for simulated metagenomes Sim1 and Sim2 and the HOT 25 m metagenome using the 'Unclassified sequences' taxonomic level.

Distribution	Read Length	Pruning	No Pruning	Reduction (%)
Sim1	Long	260	604	56.95
Sim1	Short	194	499	61.12
Sim1	Long	222	550	59.64
Sim2	Short	184	502	63.35
HOT (25m)	N/A	425	593	28.33

Table S10. Total predicted pathways for pairwise combined tier-2 BioCyc genomes: *Aurantimonas manganoxydans SI85-9A* (A), *Bacillus subtilis subtilis 168* (B), *Caulobacter crescentus NA1000* (C), and *Helicobacter pylori 26695* (H).

	A	**B**	**C**	**H**
A	394	497	424	435
B		361	481	402
C			378	416
H				210

Table S11. Number of candidate pathways that are potentially distributed by set-difference calculation (full listing in **Additional File 2**).

	A	**B**	**C**	**H**
A	-	4	1	6
B		-	6	11
C			-	2
H				-

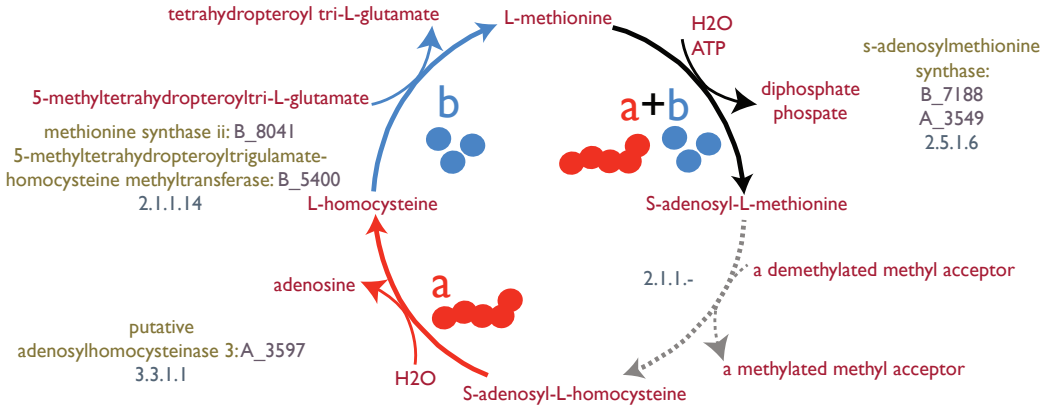


Figure S2. An example of a plausible emergent metabolism pattern. The completion of the pathway requires participation from multiple taxa, e.g., *Aurantimonas manganoxydans SI85-9A* (a) and *Bacillus subtilis subtilis 168* (b). Pathway glyphs produced by Pathway Tools can be supplemented with taxonomic information to enable the discovery of patterns of inter-pathway complementarity and potentially distributed metabolic pathways.

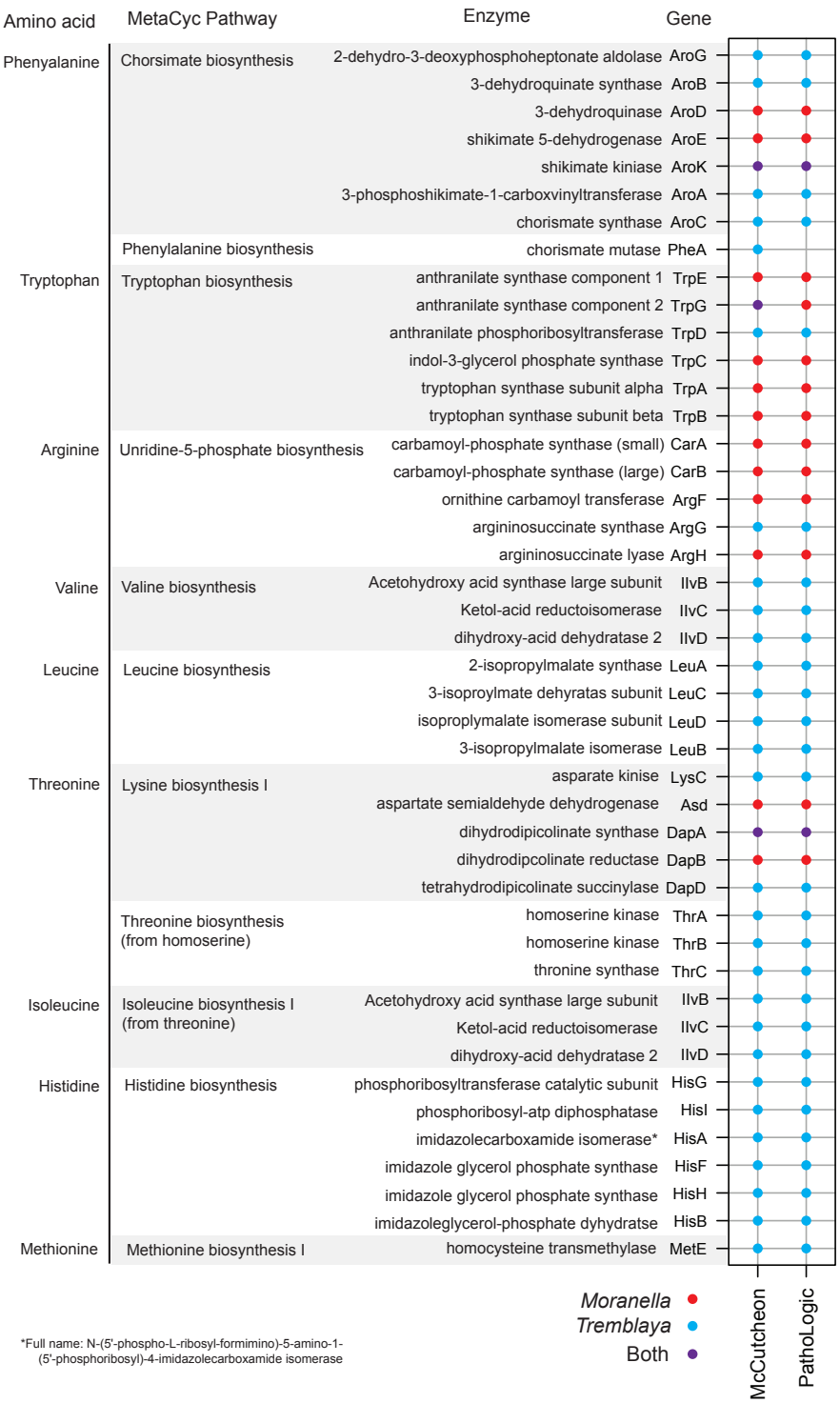


Figure S3. Comparison of predicted amino acid pathways in the *Candidatus Moranella endobia* and *Candidatus Tremblaya princeps* genomes. Dots represent detected presence of the pathway enzymes in *Moranella* (red), *Tremblaya* (blue), or both genomes (purple).

Table S12. Summary statistics of pathway prediction for the HOT metagenome and metatranscriptome.

Sample	GenBank SRA	Size (bp)	Reads	ORFs	Annotated ORFs	MetaCyc Reactions	Predicted Pathways
HOT 25m	SRX007372	160,254,663	623,559	405,613	214,149	4,138	864
HOT 75m	SRX007369	164,376,456	673,674	430,689	222,572	4,052	854
HOT 110m	SRX007370	127,754,820	473,166	336,035	165,775	4,133	860
HOT 500m	SRX007371	274,826,172	995,747	714,743	361,193	4,464	949
HOT 25m (cDNA)	SRX016893	139,331,608	561,821	234,404	85,781	3,433	723
HOT 75m (cDNA)	SRX016897	133,294,602	557,718	203,359	66,855	3,208	669
HOT 110m (cDNA)	SRX156384	90,843,408	398,436	135,107	36,912	2,549	532
HOT 500m (cDNA)	SRX156385	127,589,826	479,661	207,465	71,400	3,034	641

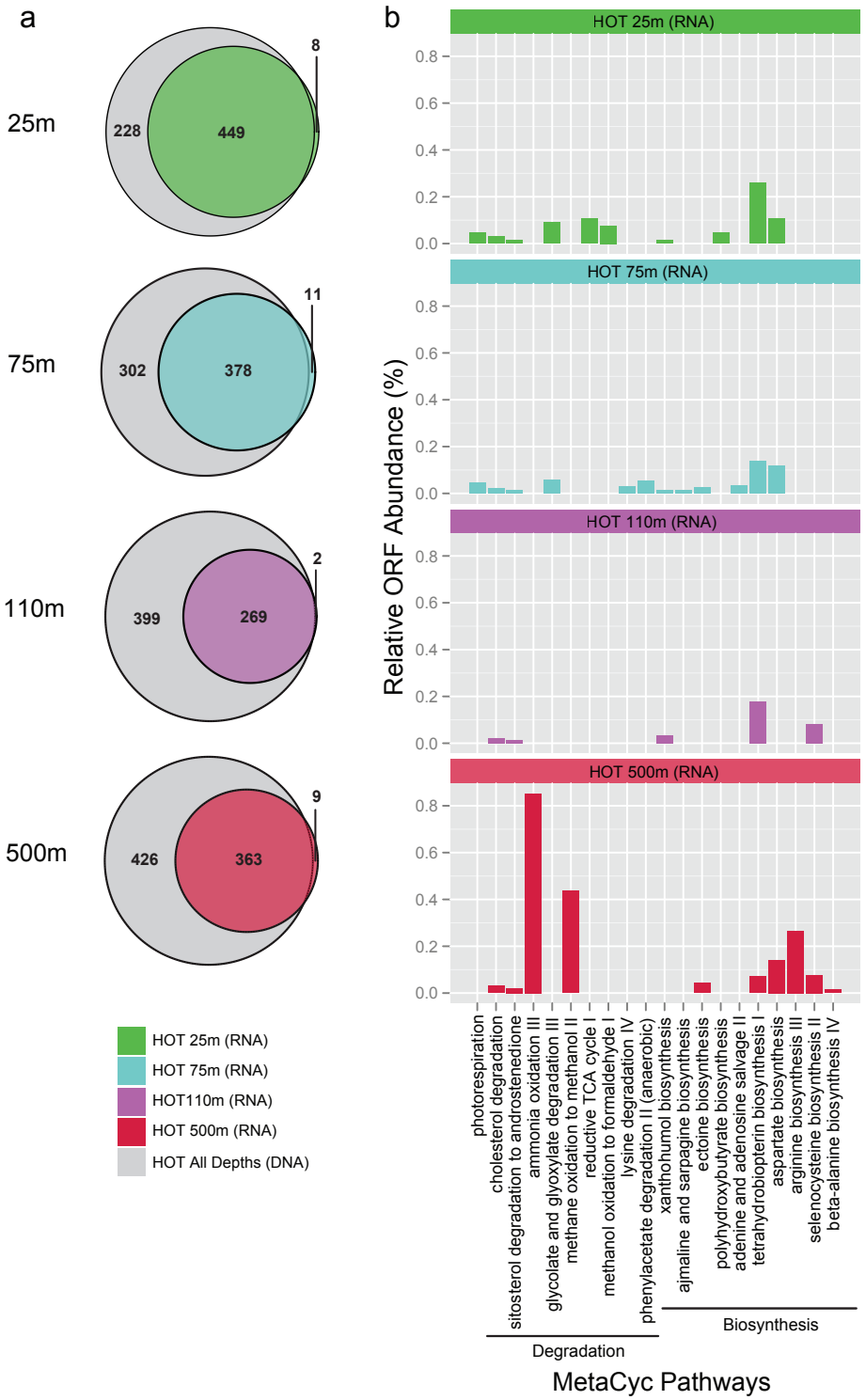


Figure S4. Overview of unique transcriptomic signal. **(a)** Euler diagrams comparing common genomic and transcriptomic pathways for each depth. **(b)** Unique transcriptomic pathways projected to all depths.

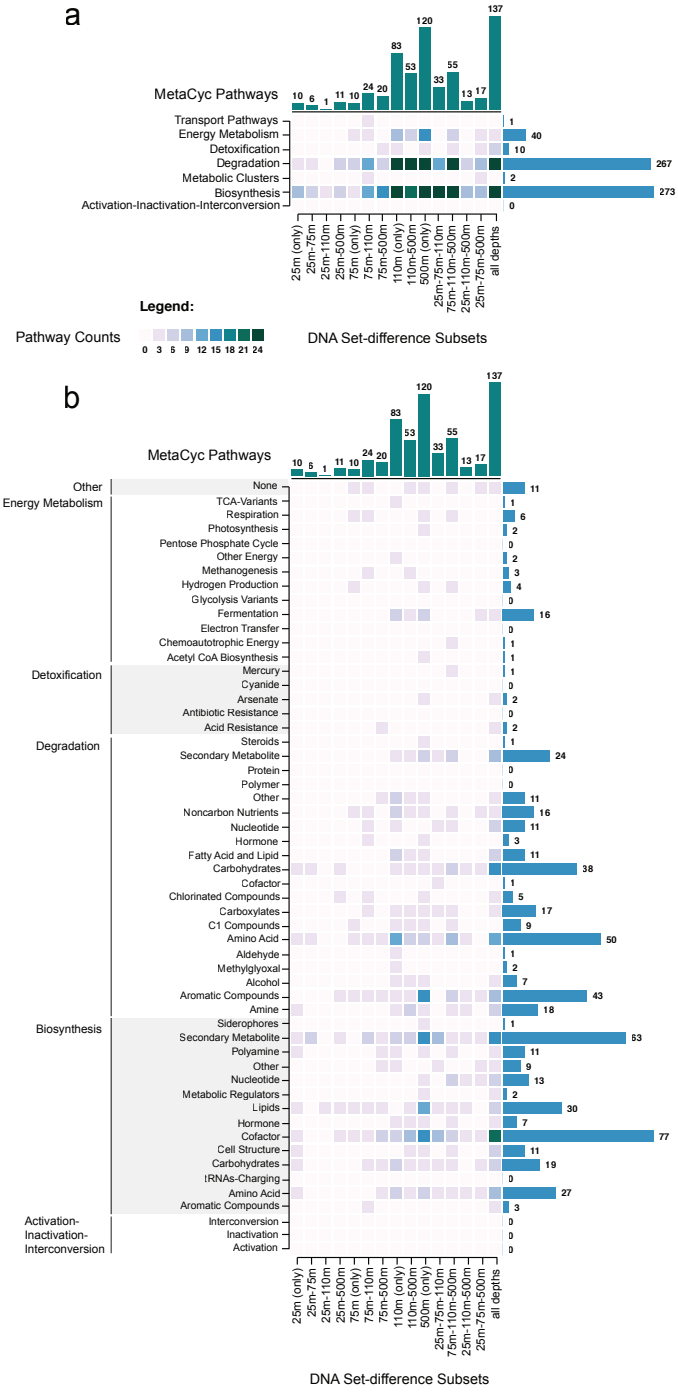


Figure S5. Predicted pathways predicted unique to DNA samples. **(a)** Unique DNA pathways projected at the highest MetaCyc classification. **(b)** Unique DNA pathways projected to the next MetaCyc sub-classification.

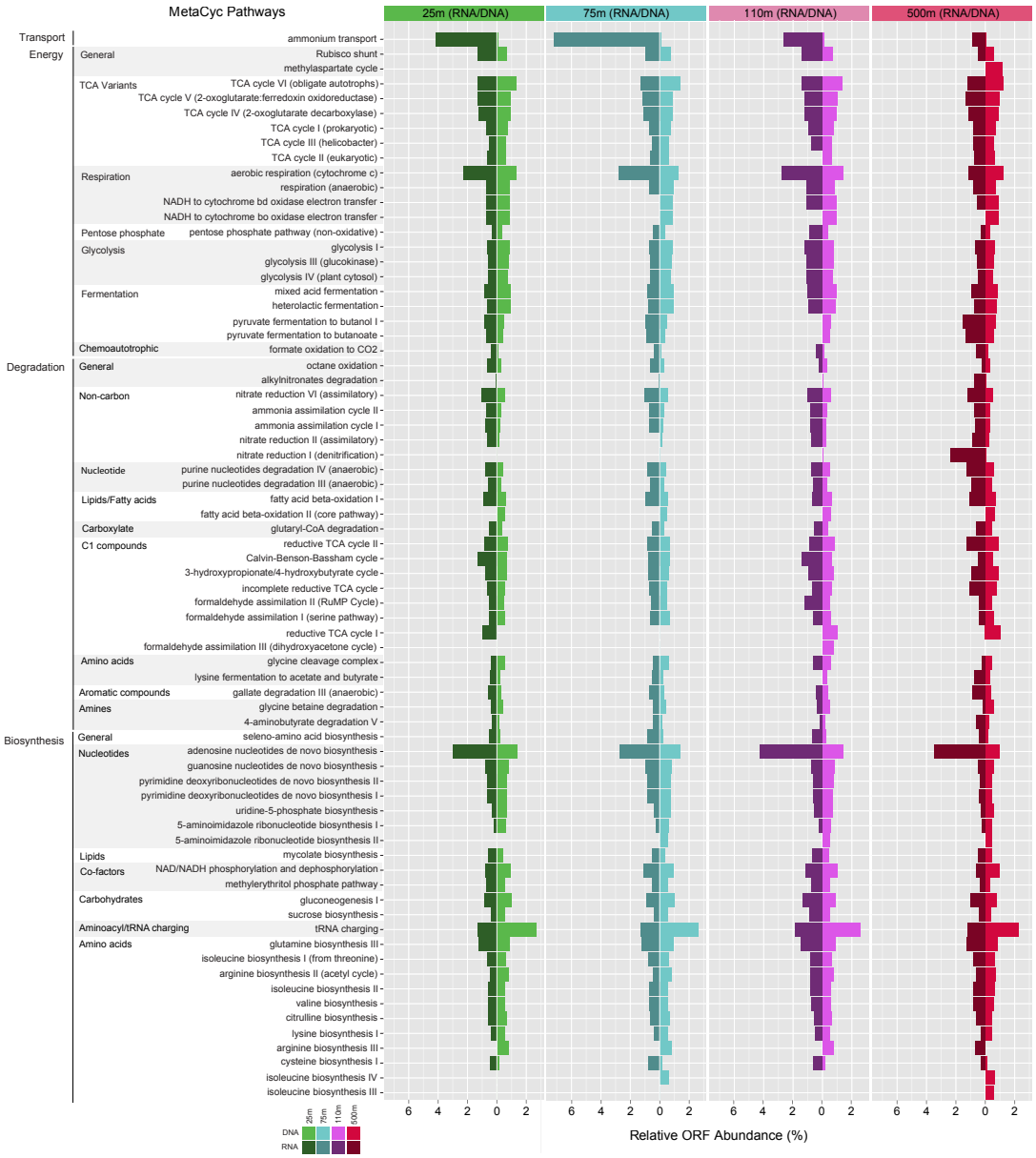


Figure S6. Top-40 predicted pathways based on open reading frame (ORF) and transcript abundance from four HOT depth intervals. The most abundant pathways were largely stable between samples, with the Rubisco shunt, pyruvate fermentation, NADH to cytochrome electron transfer, aerobic respiration, nitrate reduction, and pyruvate fermentation varying between sunlit and dark ocean waters.

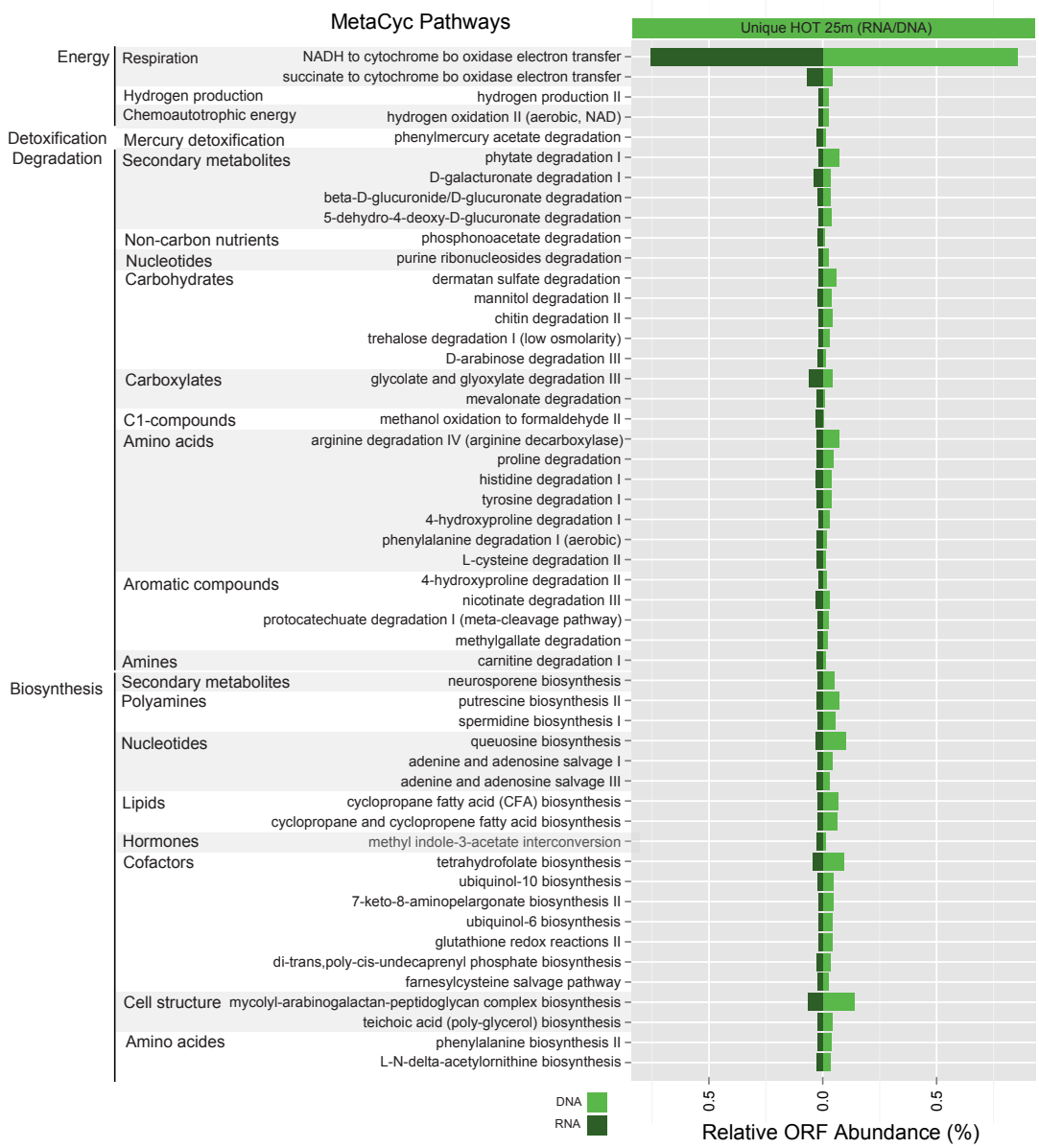


Figure S7. Genomic and transcriptomic signal for pathways unique to the 'near surface' sample (25 m). The predicted presence of cytocrome oxidase electron transfer and reversible hydrogen production and oxidation indicate a strong signal of aerobic growth.

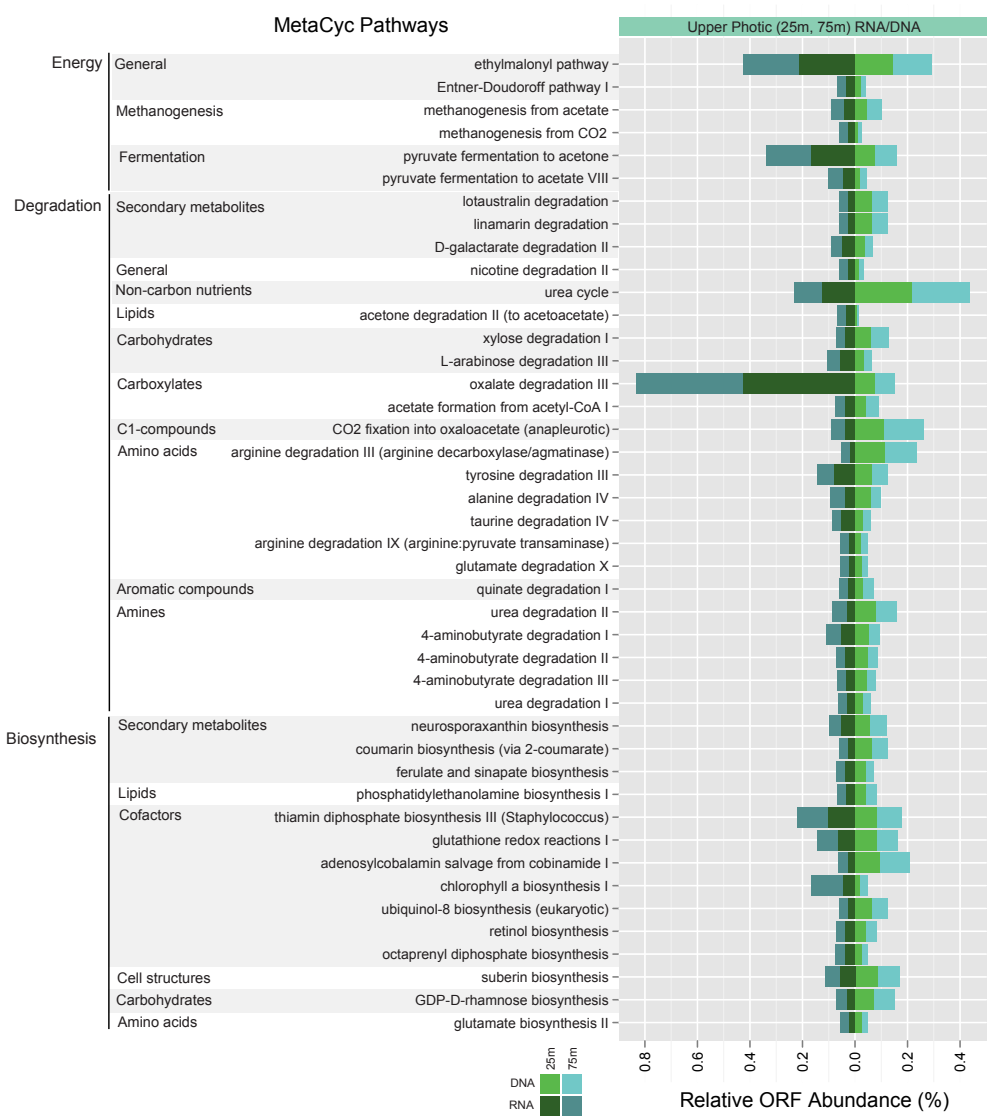


Figure S8. Genomic and transcriptomic signal for pathways unique to the 'upper photic zone' samples (25 m and 75 m). Notable pathways included chlorophyll a biosynthesis, the ethylmalonyl, Entner-Doudoroff, and pyruvate fermentation pathways.

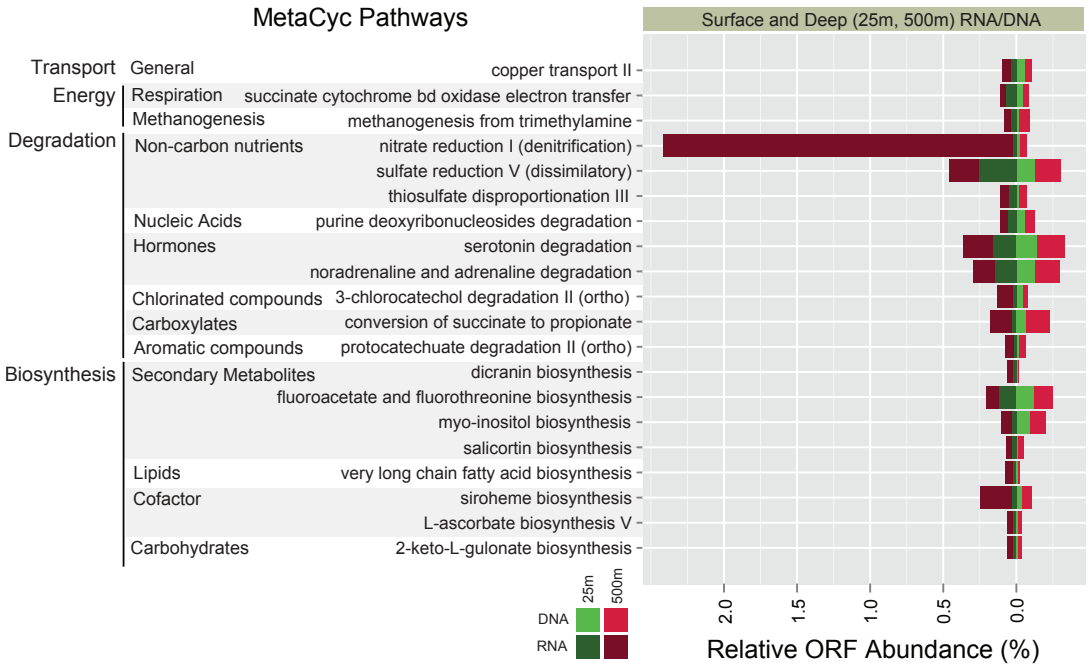


Figure S9. Genomic and transcriptomic signal for pathways unique to the 'surface and deep' samples (25 m and 500 m). There exist a limited number of pathways common to the surface and deep, but note that the largest signal is for nitrate and sulfate reduction, the first steps of sulfur recycling being shared with nitrate reduction.

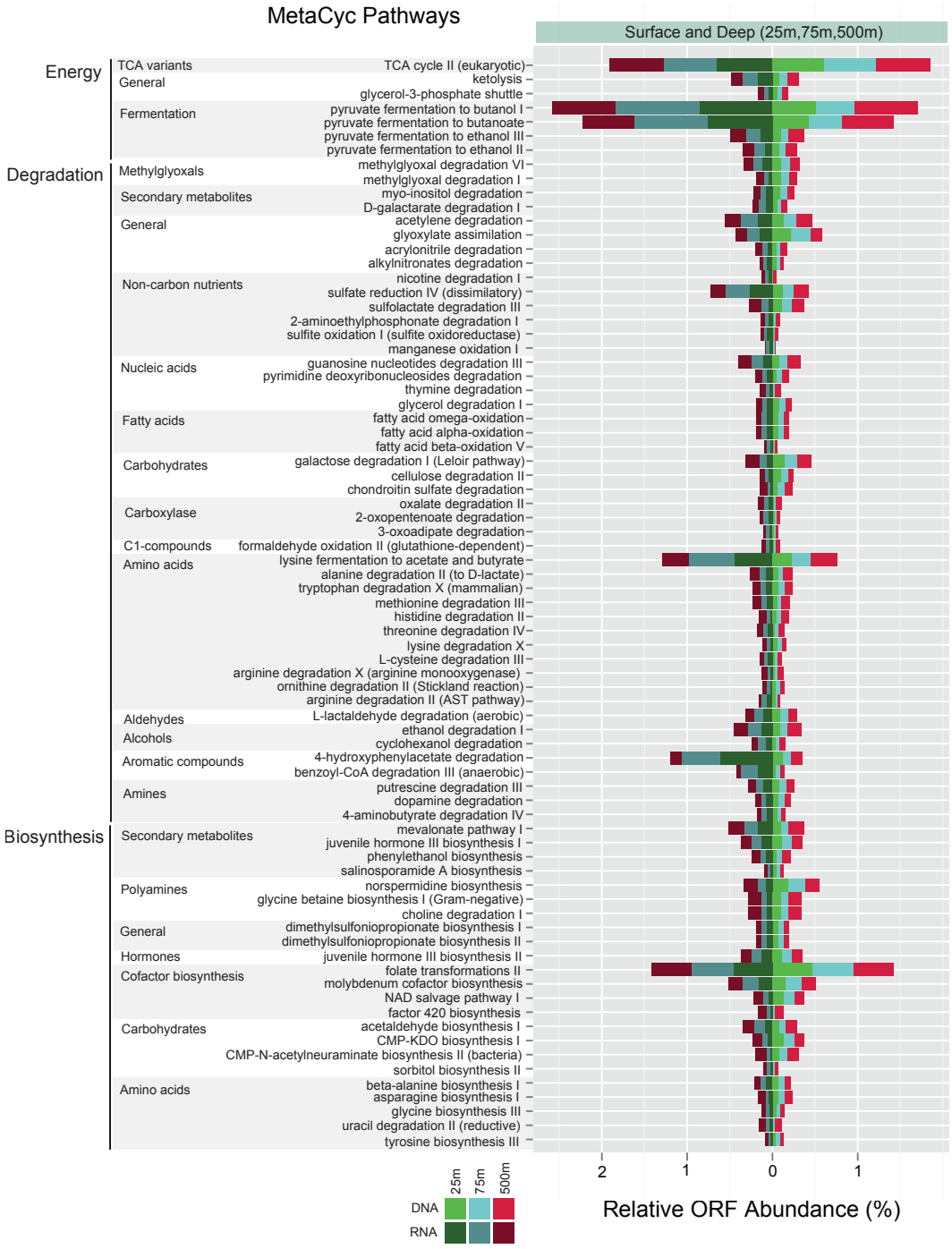


Figure S10. Unique pathways to the 'photic and deep' samples (25 m, 75 m, and 500 m). This set is characterized by a strong signal for the TCA cycle, ketolysis and pyruvate fermentation, as well as a large number of degradation pathways for many major compounds, consistent with the breakdown of biological material falling from the surface waters.

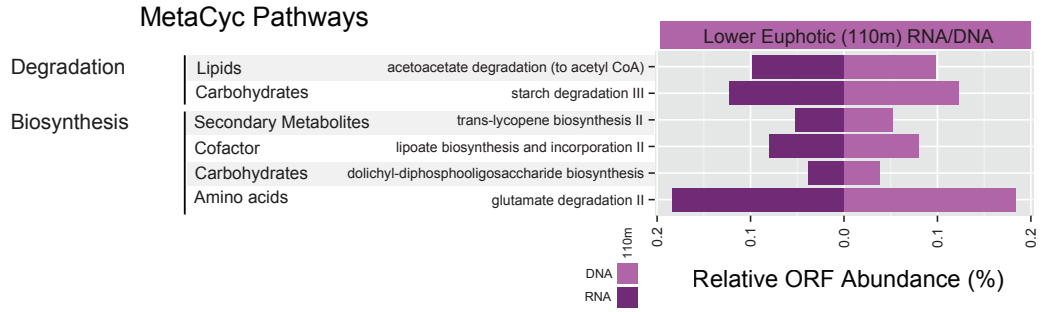


Figure S11. Unique pathways to the lower euphotic 110m sample.

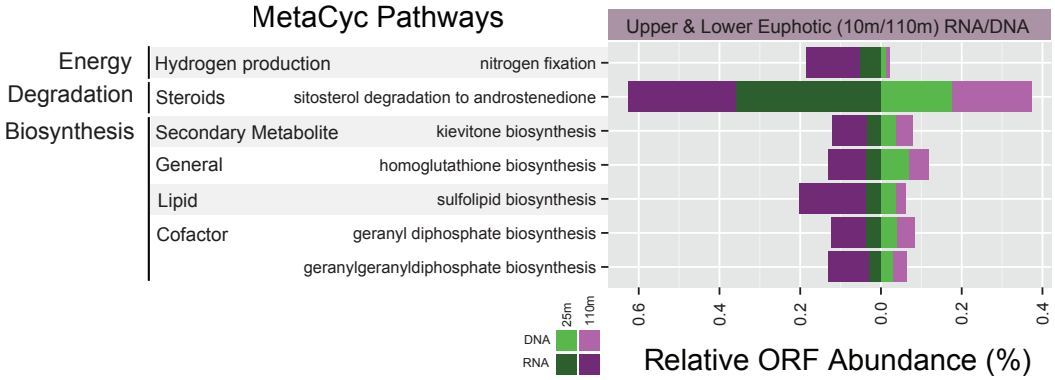


Figure S12. Unique pathways to the upper and lower euphotic 25m and 110m samples.