The following supplement accompanies the article

Winter mixing impacts gene expression in marine microbial populations in the Gulf of Aqaba

D. Miller, U. Pfreundt, S. Hou, S. C. Lott, W. R. Hess, I. Berman-Frank*

*Corresponding author: ilana.berman-frank@biu.ac.il

Aquatic Microbial Ecology 80: 223–242 (2017)

Supplement 1

Supplementary Materials and methods

Analysis of metatranscriptome libraries

Quality control, adapter trimming and computational removal of rRNA reads

The complete analysis pipeline of metatranscriptome libraries is visualized in Fig. S1. Following sequencing, adapter sequences and low quality bases were clipped using cutadapt v1.0 (Martin, 2011). Adapter sequences were removed wherever they appeared in the read ("-b" switch) using a minimum overlap of 10 nt between read and adapter sequence and 0.2 allowed error rate. Minimum quality was set to 20 and all reads shorter than 20 nt were removed. FASTX-toolkit (version 0.0.13 - http://hannonlab.cshl.edu/fastx_toolkit/) was used to convert fastq to fasta while keeping "N" characters in the sequence ("-n"). 100% identical reads were clustered using a custom made perl script and rRNA reads were removed using SortMeRNA v1.9 (Kopylova et al., 2012) with default parameters and the provided rRNA database.

Similarity searches of mRNA sequence reads

Quality filtered, clustered fasta formatted sequences were subjected to similarity searches against the refseq's nt (nucleotide) and nr (non-redundant protein) databases (http://www.ncbi.nlm.nih.gov/refseq/about/). To analyze the taxonomic affiliation of the reads, dc-megablast search against the nt database (updated on 28th, September 2014) was performed using an e-value cutoff of 10^{-5} and without any identity cutoff. The top 25 hits were used for taxonomic classification with MEGAN 5.10.6 (Huson et al., 2011 see next paragraphs). For global functional analysis of the data set, DIAMOND (Buchfink et al., 2015) was used to search the reverse-reads libraries against the NCBI nr database (updated on 28th, September 2014) in blastx mode with default parameters (including an e-value cutoff 10^{-3}). The 25 top hits were selected for further analysis. The forward reads in dRNA-seq libraries mostly correspond to the 5' untranslated end of a transcript (Hou et al., 2016) and cannot be detected in similarity searches against a protein database. This was the reason for applying DIAMOND searches only on reverse read libraries.

Global functional analysis

The results obtained in the DIAMOND similarity search were further analyzed by MEGAN 5 (Huson et al., 2011) using the following last common ancestor (LCA) parameters: min identity score: 50, max e-value: 10^{-5}, top percent: 5, min support percent: 0.0 (off), min

support: 1, LCA Percent: 90, min complexity filter: 0.0 (off) (further details in the MEGAN manual at http://ab.inf.uni-tuebingen.de/data/software/megan5/download/manual.pdf). Reads with a hit in the nr database were counted and used as input for differential expression analysis using NOISeq (Tarazona et al., 2011). The NOISeq-sim method implemented in NOIseq simulates replication to increase statistical significance of differential expression analysis, and is thus suitable for analysis of non-replicated data sets.

MD differential expression plots (plotting mean log2 value M against absolute value of the difference in TMM normalized read counts D) were produced by NOISeq-sim using a five repeats simulation of 20% of the data sets and variability of 2% (pnr = 0.2, nss = 5, v = 0.02). Differential expression is defined as described below (Page 9).

Taxonomic affiliation of mRNA sequence reads

The dc-megablast results were analyzed using MEGAN 5, applying the same LCA parameters as for the analysis of DIAMOND results (see above). Taxonomic analysis of mRNA reads was performed using read counts normalized to the smallest library size. Eukaryotic organisms were omitted from the mRNA taxonomic analysis to enable the comparison of the relative transcript composition with population composition determined using the 16S rRNA gene amplicons. However, eukaryotes comprised18.5%, 8.9% and 9.6% of all reads in 2.5, 45 and 440 m, respectively (see main text,

Selection of reference organisms

We attempted but were not able to appropriately de-novo assemble the metatranscriptomes to transcripts and use them as references (Pfreundt et al., 2014). This was probably due to the fact that dRNA seq enriches the sequence library with reads assigned to the 5' end of transcripts, restricting the localization of reads to a defined transcript region. In addition, assembly of reads emerging from complex communities of metagenomes and metatranscriptomes can generally be seen critically (Charuvaka and Rangwala, 2011; Mende et al., 2012). We hence used for read mapping reference genomes obtained from the RefSeq database (Tatusova et al., 2014) instead of de-novo-assembled references.

The choice of reference genome was performed separately for each taxonomic group. We mapped reads assigned to each one of the examined groups to all available genomes in the NCBI-RefSeq database of each groups (cyanobacteria reads were mapped against all available *Synechococcus* genomes, as they contained a large fraction of *Synechococcus* reads. We included reads from other cyanobacteria to be able to identify reads from genes that are conserved across cyanobacteria and are at least 75% identical with the respective *Synechococcus* allele). For cyanobacteria, most reads were mapped to the CC9605 genome (21 and 44% for 2.5 and 440m), which was chosen as reference for all cyanobacterial reads. In the case of SAR11, the three complete genomes found in the RefSeq database were compared as references for all reads assigned to SAR11 using dc-megablast. *Pelagibacter* sp. HTCC7211 recruited the highest percentage of reads (38 and 43%) and was selected as reference genome. No reference genomes could be found in RefSeq for Euryarchaeota or Thaumarchaeota. To generate a reference for mapping, archaeal reads were first mapped to a set of 997 archaeal contigs from a Mediterranean Sea metagenomics library (Deschamps et al., 2014). Contigs recruiting at least 200 reads were extracted and used as reference. This resulted in a reference composed of 115 contigs ascribed to the Euryarchaeota MG-II/III and a second reference for Thaumarchaeota, composed of 107 contigs. These references recruited 56-95% of all archaeal reads. A separately defined set of MG-II/III genomic bins (Li et al., 2015) did not improve this recruitment. All annotations used were based on the original annotations of the 997 published contigs. *Micromonas* sp. RCC299 genome which recruited 74 and 71% of Mamielalles aligned reads at 2.5 and 440m was selected as a representative of

the Mamielalles group. The mapping of reads to reference genomes was performed as follows: Reads assigned by dc-megablast to a selected taxon were extracted from each depth sample using MEGAN 5 and for prokaryotes mapped to the chosen references using Segemehl (Hoffmann et al., 2009). Low stringency 75% identity was chosen to enable the mapping of reads from related species (otherwise with default parameters). Eukaryotic reads were mapped to the reference genome using Tophat2 (Kim et al., 2013), with the following parameters: --b2-D=20, --b2-R=3, --b2-N=1, --b2-L=20, --b2-i=S,1,0.5, --read-mismatches=10, --read-gap-length=10, --read-edit-dist=15, --read-realign-edit-dist=0, --min-anchor-length=5, --splice-mismatches=1, --max-multihits=20, --microexon-search, --segment-mismatches=2, --segment-length=20, --no-convert-bam. The stringency of mapping was comparable to the 75% set up for segemehl mapping of prokaryotic reads.

Expression levels were defined as the averaged forward and reverse read counts assigned to a feature (coding sequences, non-coding RNA (ncRNA), rRNA and tRNA genes) for which a transcription start site (TSS) was detected (hereafter: TSS expression).

TSS prediction

Following mapping, TSSs were predicted as described (Hou et al., 2016, see also fig S10 in the Supplementary Methods). Due to the noisy character of RNA sequencing results, 5' ends of forward library reads were not only assigned to the exact nucleotide TSS positions: To reduce noise, dRNA-seq Forward Read Start Sites (dFRSS) were calculated as local maximum coverage in a range of 5 nt up- and downstream of the max peak. The number of reads in each cluster was defined as the TSS count. dFRSS were then listed from global highly expressed dFRSS to global lowly expressed dFRSS (see also Mitschke et al., 2011a). dFRSS were primarily classified gTSS, iTSS, aTSS and nTSS according to their genomic context. These correspond to **g**ene, **i**nternal, **a**ntisense, and **n**o-gene TSS, with no-gene referring to TSS found in genomic loci without a known transcribed feature (Mitschke et al., 2011b). gTSS was defined if a TSS was located 0-200 nt upstream from an annotated gene (0-600 nt in the case of eukaryotic chromosomal genes that may have longer 5' UTRs), or – if more than 200 nt upstream from a gene – if the forward or revers reads are overlapping with the gene. Since this study was focusing on protein coding genes and metabolic functions expressed, only gTSS were regarded in the downstream analysis. Noise was further reduced by dividing the TSS count by maximum coverage of forward reads in the 5 nt up- and downstream region used to define the TSS (covRatio parameter – minimum 0.5). Assuming more reads will map to a genomic region downstream of a TSS than upstream, the total coverage (of number of reads mapped) in the 100 nt downstream from the examined TSS was divided by the total coverage 100 nt down- and upstream of it (locCovEnrich parameter. Threshold >0.4-0.65). To pick a TSS from a selection of possible TSS peaks in close vicinity, the locTssEnrich parameter was calculated as the TSS count of the examined TSS divided by the sum of all close optional TSS (since more than one TSS may be attributed to a gene yielding transcripts with different properties, we did not simply select the TSS with the highest value. The threshold for the locTssEnrich parameter was set to 0.4–0.5). Finally, revSumRatio was defined as the TSS count was divided by the sum of reverse reads mapped to the 500 nt downstream region, expecting a small value (if too little reads from the rev library are mapped to the region of the gene compared with the forward reads mapped to the TSS, then forward reads must have been falsely assigned to the TSS). For visual description of all the above mentioned parameters, see (Hou et al., 2016). Threshold values were set after a primary TSS prediction, reflecting the conditions that will identify 95% of dFRSS as TSS. These conditions varied between the different groups of organisms, on which the analysis was applied. The minimal threshold for TSS count to define a TSS differed between representative genomes based on their relative abundance and read distribution along the transcript

(affecting TSS prediction): 13 reads were set as minimum TSS count for *Synechococcus*, 16 reads for SAR11, 11 and 15 reads for Euryarchaeota and Thaumarchaeota, respectively, and 6 and 8 reads for the *Micromonas* RCC299 plastid / mitochondrial and chromosomal genes, respectively. In a second step, TSS prediction was controlled manually by visualizing the coverage of the reads mapped to the reference genome using Artemis genome browser (Rutherford et al., 2000), and following 20 - 25 detected TSS to find their typical peak form in the plot (Fig. S2). If a clear TSS peak was observed in the visualization but no TSS was predicted (or vice versa) the thresholds for TSS prediction were adjusted. Following TSS prediction, all reads assigned to a TSS-associated gene were counted and included. Reads assigned to genes without a predicted TSS were counted as well (and their respective transcripts names begin with the word "None" instead of "gTSS").

After detecting TSS, averaged values of counts of forward and reverse reads assigned to each TSS were used as a measure for expression level of the associated gene. These values were compared between the different depths for differential expression analysis with NOISeq as described below.

Analysis of differential expression

Differential expression (DE) was assessed using NOISeq-sim implemented in the NOISeq package (Tarazona et al., 2011). The calculation of DE was based on TMM normalized TSS expression values (Trimmed mean of M, Robinson and Oshlack, 2010) using five technical repeat simulations, each representing 20% of complete sample size (pnr = 0.2, nss = 5). 2% variability was presumed (v= 0.02) and the probability threshold for DE was set to 0.9 (q = 0.9). The higher the probability, the more likely that the observed difference is due to different experimental condition and not to chance (see NOISeq manual for details). Transcripts with <2 counts per million (CPM) were excluded from the analysis. Since the size of each examined feature (transcript) is known from the reference genome, size correction was also included in the analysis. In some cases, two different TSS of the same gene were differentially expressed, indicating possibly alternatively regulated promoters. Since no data was available on different functions such hypothetical transcripts, they were treated as synonymous and were considered as differentially expressed when the TMM value of an overexpressed transcript in one sample was at least three-fold higher compared to the TMM value of the other transcript, detected as overexpressed in the other sample.

Analysis of KEGG pathways

Protein sequences matching the respective transcripts were used for BlastKoala searches (Kanehisa et al., 2016) against the KEGG database (Kanehisa and Goto, 2000), separately for each of the reference genomes. The resulting KEGG orthologues (KO) were visualized using iPath2.0 (Yamada et al., 2011). The procedure was performed also on lists of differentially expressed KO numbers, and depth specific functional profiles were generated in the form of heat-maps visualizing the number of differentially expressed KOs recruited to KEGG pathways in each depth

Alignment of cyanobacterial reads to carbon transporters

To answer whether cyanobacteria are applying mixotrophic metabolism when subjected to deep mixing conditions, we aimed to expand our search for carbon transporters beyond the genes annotated on the chosen reference genome. To that aim we constructed a blast database using a set of 26 *Synechococcus* gene sequences encoding transporters for organic compounds, carbon dioxide and bicarbonate. To construct the database, we performed a word search on the word "transporter" against the genomes of *Synechococcus sp*. WH 7803 and *Candidatus Synechococcus spongiarum* and extracted all carbon transporters protein

sequences. Seventeen of these were annotated as “multispecies”. Alignment was performed using command line blastx using the following command: blastx -query $QUERY_FILE -db $CUSTOM_MADE_C_TRANSPORTERS_DB -out $OUTPUT_FILE_NAME -evalue 100 -outfmt "6 qseqid qlen sseqid slen qstart qend sstart send length pident evalue bitscore salltitles staxids" -max_target_seqs 25 -num_threads 15. The results were filtered with a custom python script, and only hits with both alignment length and bitscore above 50 were kept. The sequences used for the database and the filtration script are available on demand.

Supplementary Figures

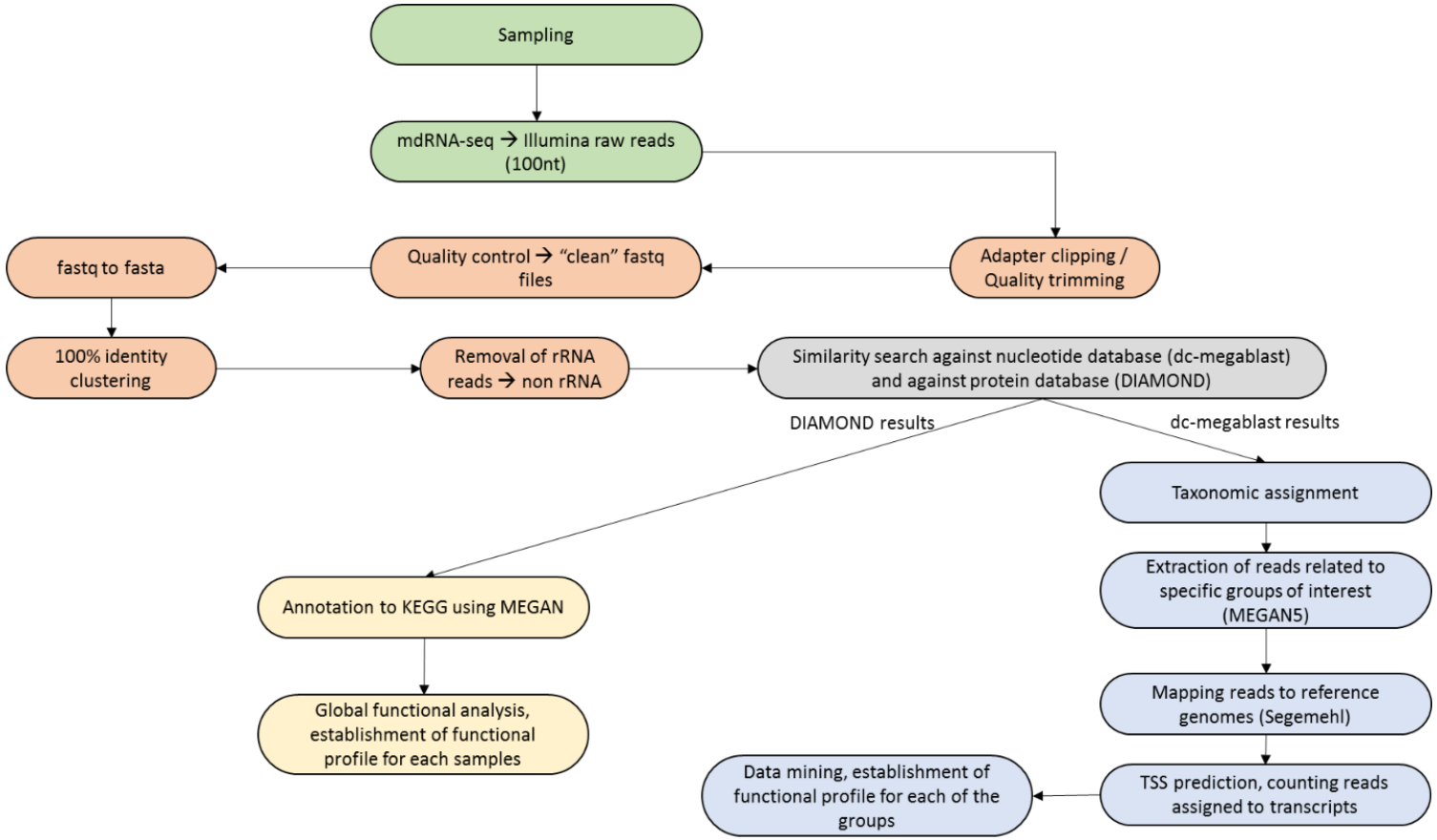

Figure S1: Bioinformatics pipeline for the data analysis performed in this work. Steps preceding bioinformatics analysis are presented with green background. Quality control and cleaning steps have red background, the similarity search step has purple background, and downstream analyses are divided into analysis of searches against nucleotide database (blue background) and protein database (yellow background). Abbreviations: mdRNA-seq: meta-differential RNA sequencing, rRNA: ribosomal RNA, dc-megablast: discontiguous megablast, TSS: transcription start site.

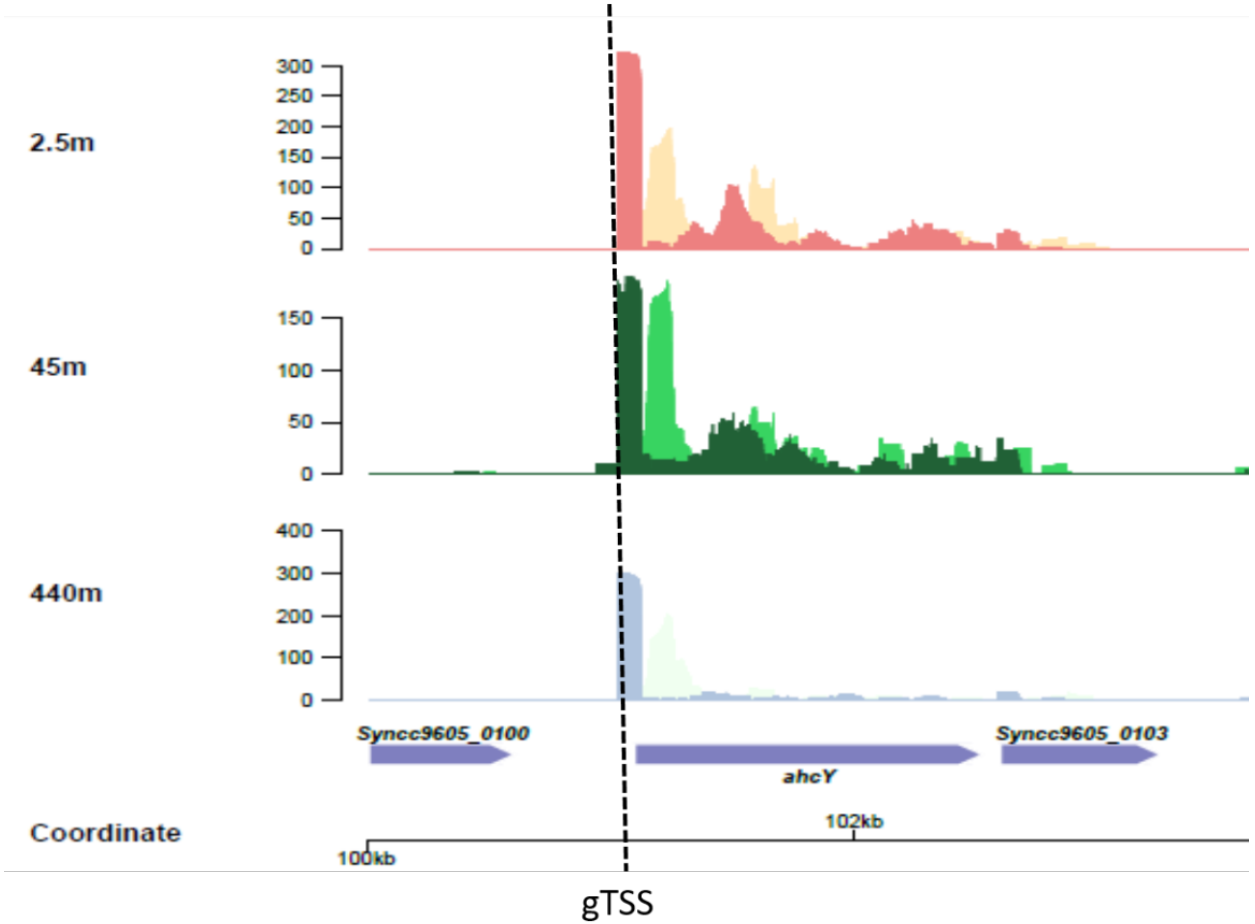

Figure S2*:* Example for a gene TSS (gTSS) and visualization of read coverage for the samples from 2.5, 45 and 440 m in red, green and blue, respectively. Presented is the *ahcY* gene of *Synechococcus* CC9605 encoding for adenosylhomocysteinase. Lighter colors represent the coverage achieved by mapping the reverse reads. The dashed line marks the gTSS at position 101030 of the reference genome.

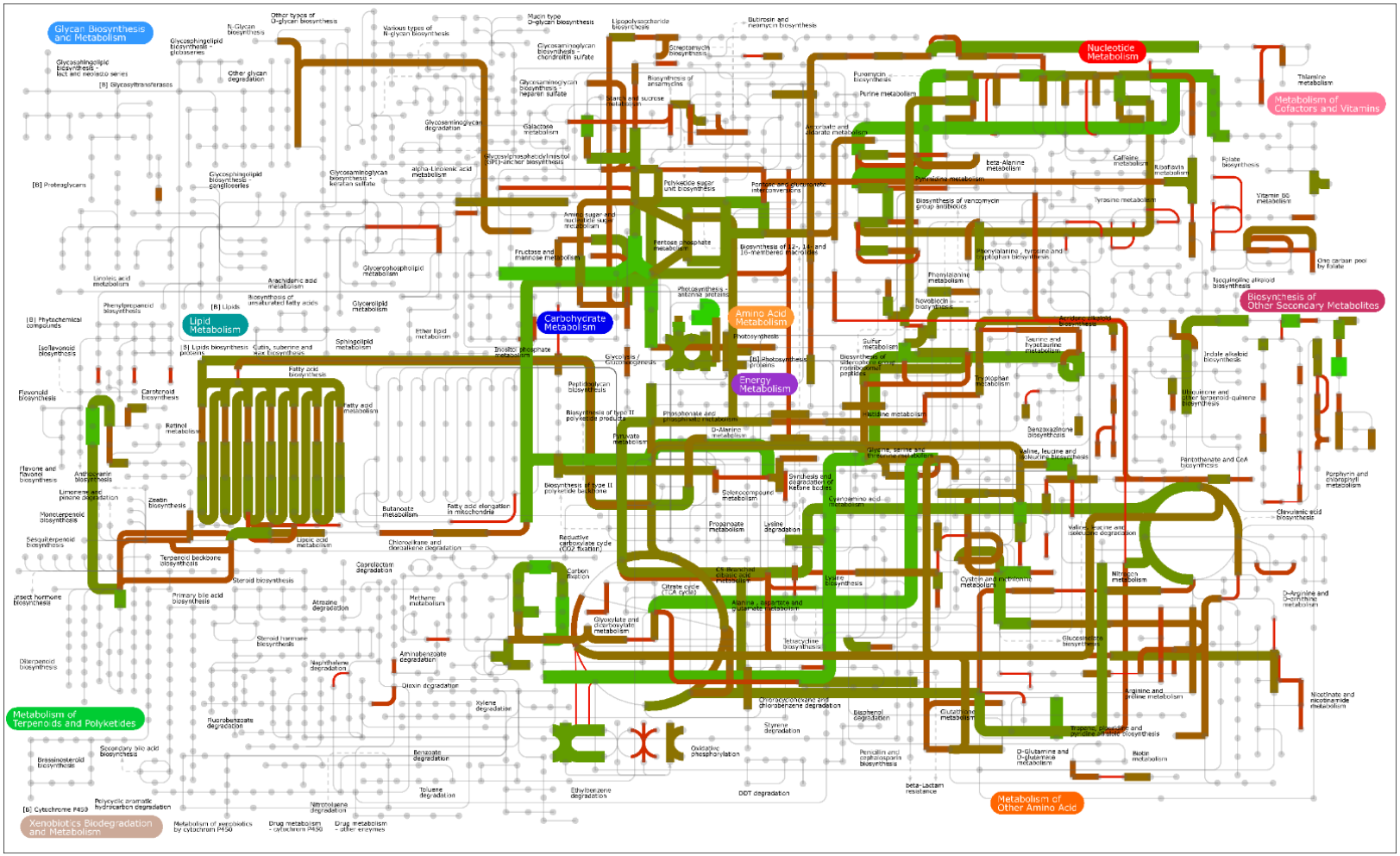

Figure S3: Metabolic map of all KOs found among cyanobacterial reads when mapped to the *Synechococcus* CC9605 genome. Nodes and edges represent metabolites and enzymes catalyzing the reactions that convert the linked metabolites, respectively. Expression level (averaged TMM normalized counts in 2.5 and 440 m) is illustrated as both the width and the color of the edges – the wider and the greener an edge appears, the stronger is the expression of the gene it represents. Map was generated using iPATH2 (http://pathways.embl.de/) – visit the iPATH website for specific details on each node and edge in the graph.

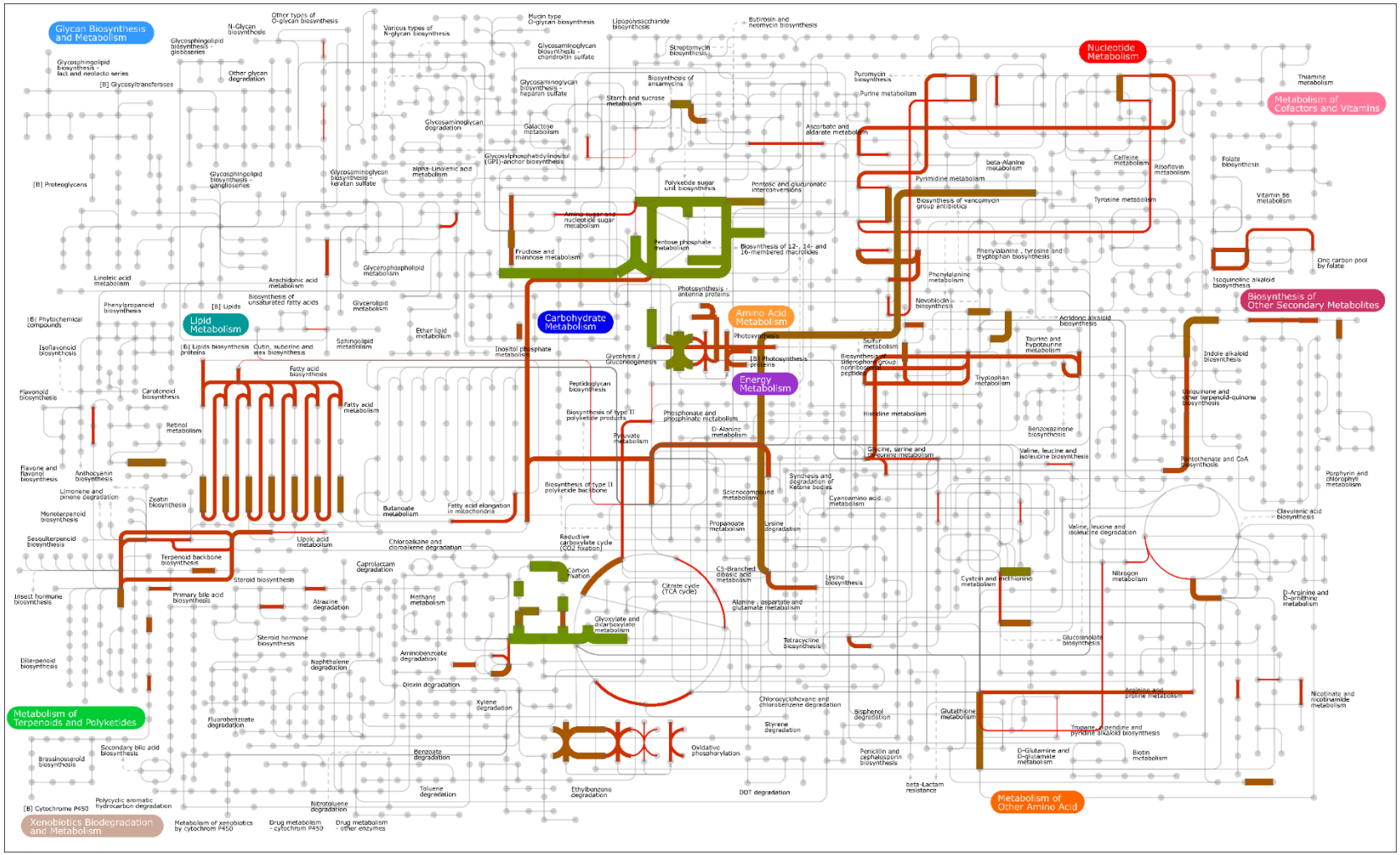

Figure S4: Metabolic map of all KOs found among Mamiellales reads when mapped to the *Micromonas* RCC299 genome. Legends are the same as in figure S4. Map was generated using iPATH2 (http://pathways.embl.de/) – visit the iPATH website for specific details on each node and edge in the graph.

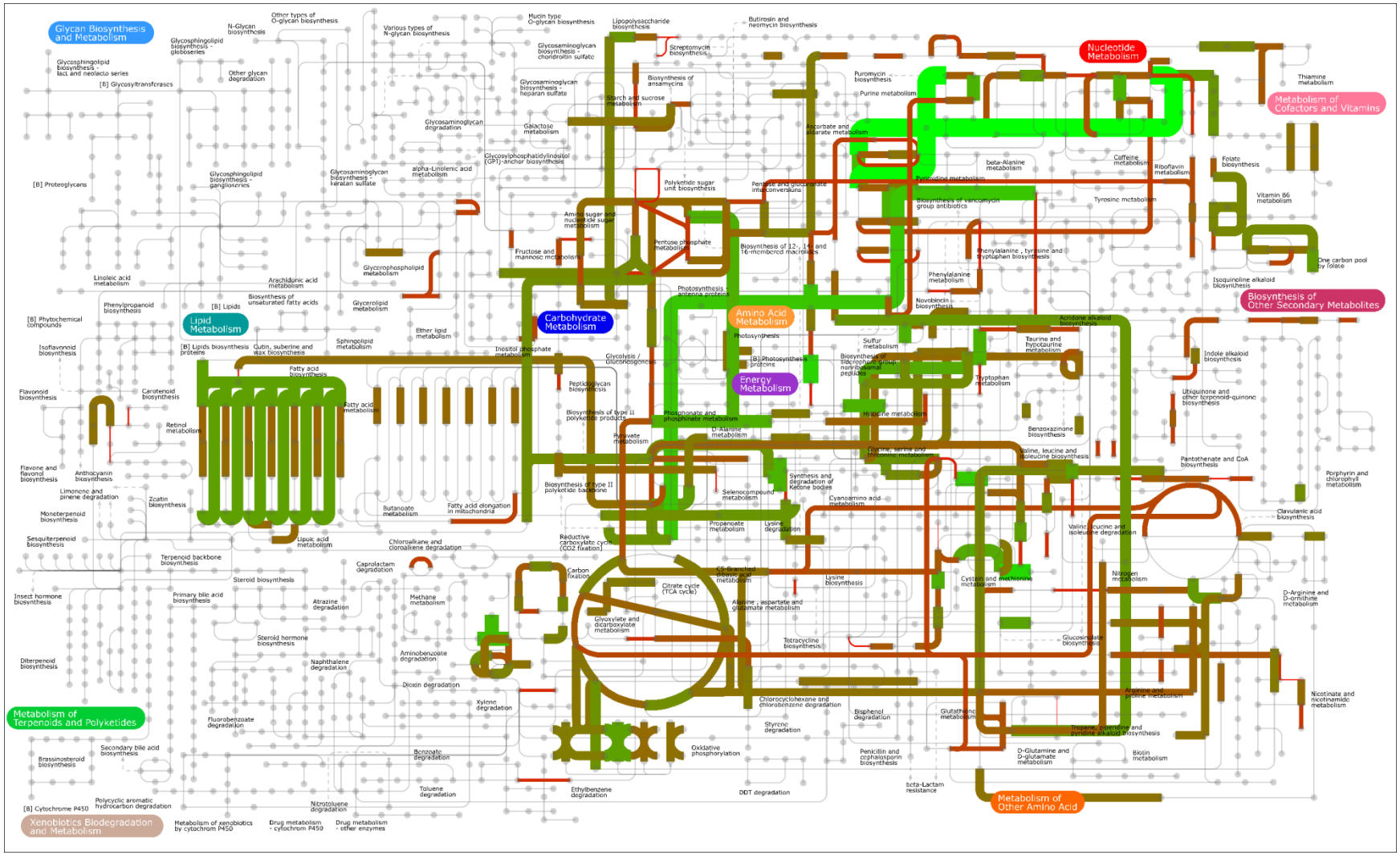

Figure S5*:* Metabolic map of all KOs found among SAR11 reads when mapped to the *Pelagibacter* sp. HTCC7211 genome. Legends are the same as in Figure S4. Map was generated using iPATH2 (http://pathways.embl.de/) – visit the iPATH website for specific details on each node and edge in the graph.

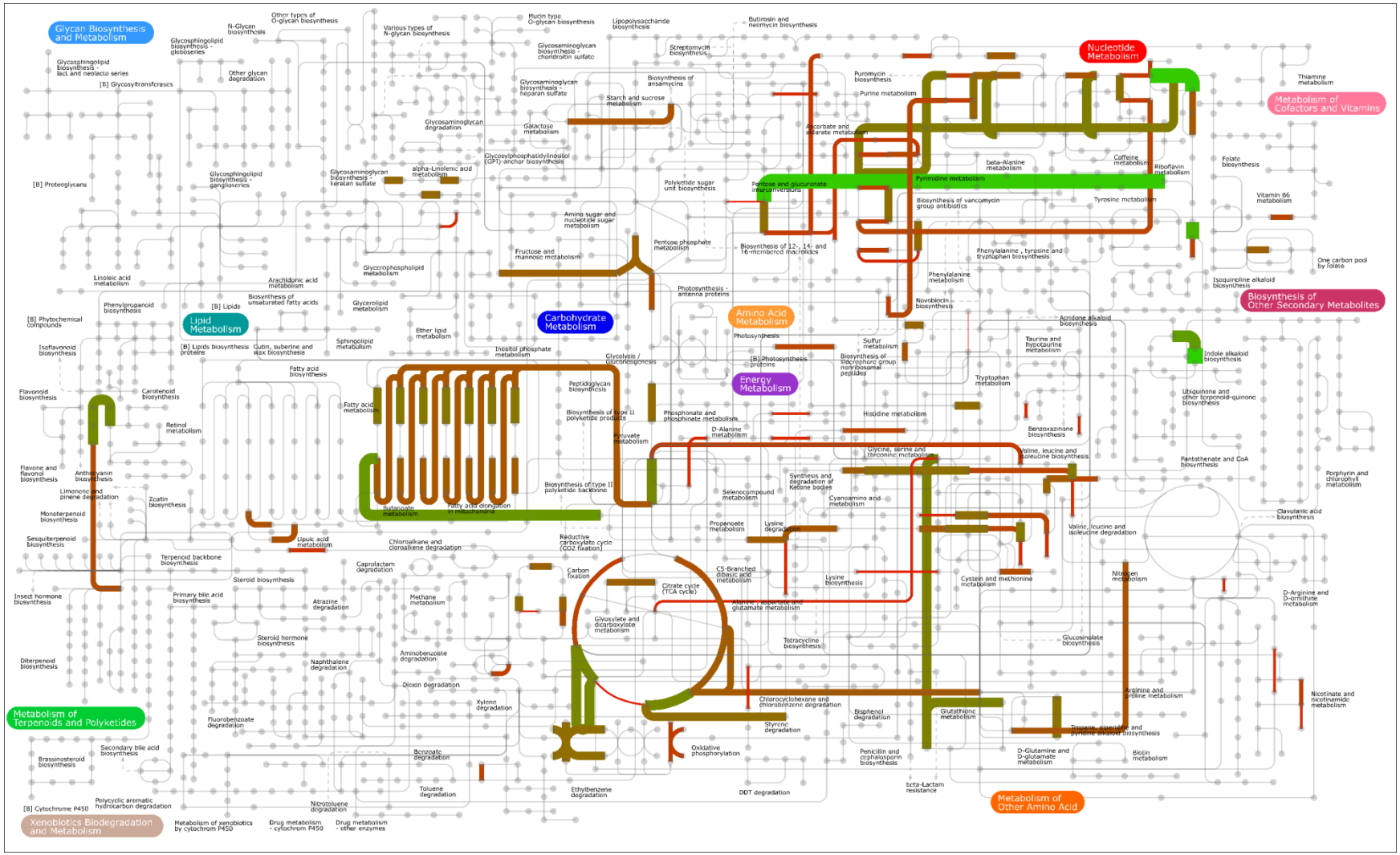

Figure S6*:* Metabolic map of all KOs found among Euryarchaeota reads when mapped to the custom reference. Legends are the same as in figure S4. Map was generated using iPATH2 (http://pathways.embl.de/) – visit the iPATH website for specific details on each node and edge in the graph.

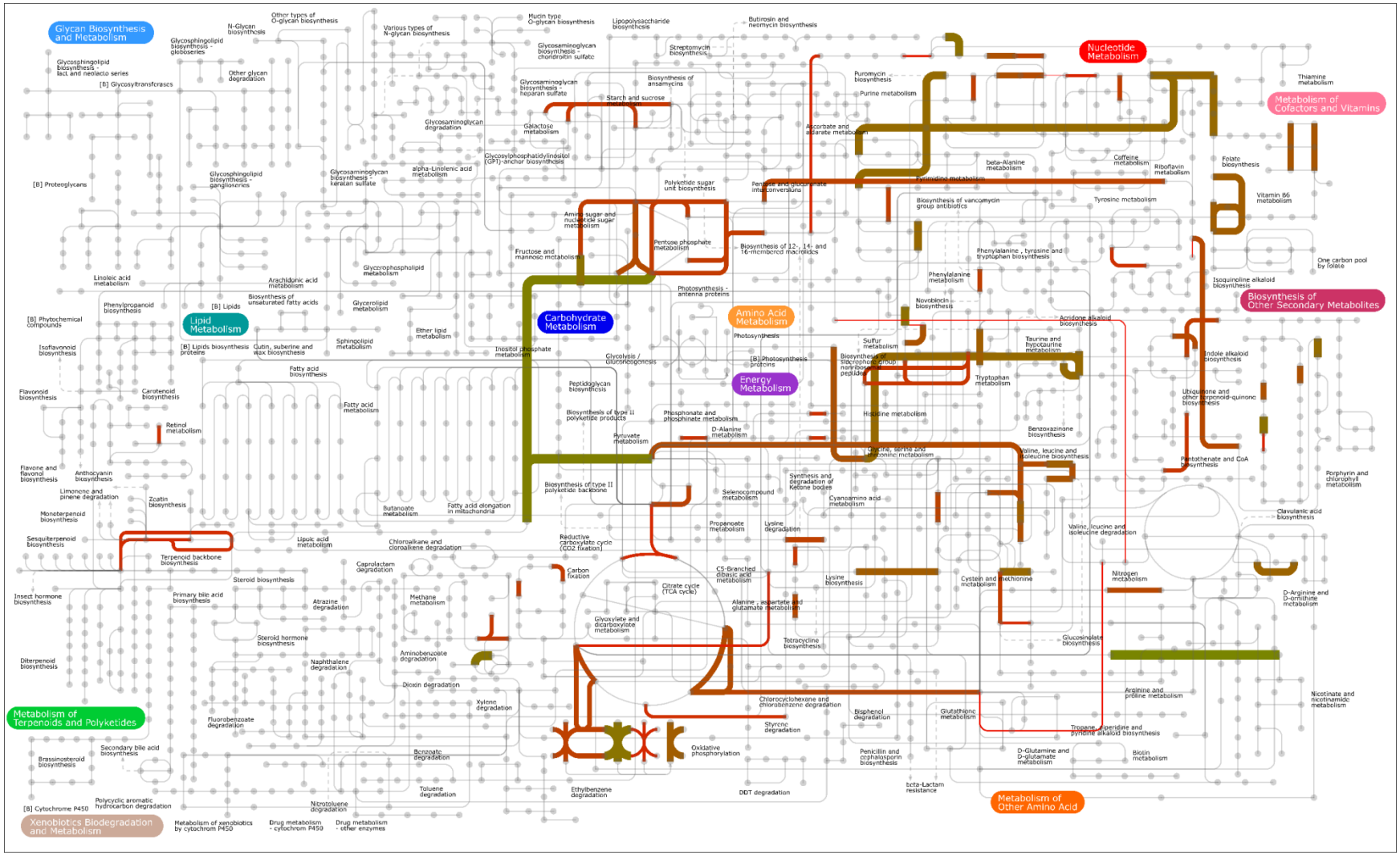

Figure S7: Metabolic map of all KOs found among Thaumarchaeota reads when mapped to the custom reference. Legends are the same as in Figure S4. Map was generated using iPATH2 (http://pathways.embl.de/) – visit the iPATH website for specific details on each node and edge in the graph.

Supplementary References

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60

Charuvaka A, Rangwala H (2011) Evaluation of short read metagenomic assembly. BMC Genomics 12:1–13

Deschamps P, Zivanovic Y, Moreira D, Rodriguez-Valera F, López-García P (2014) Pangenome Evidence for Extensive Interdomain Horizontal Transfer Affecting Lineage Core and Shell Genes in Uncultured Planktonic Thaumarchaeota and Euryarchaeota. Genome Biol Evol 6:1549–1563

Hoffmann S, Otto C, Kurtz S, Sharma CM and others (2009) Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. PLOS Comput Biol 5:e1000502

Hou S, Pfreundt U, Miller D, Berman-Frank I, Hess WR (2016) Differential RNA-Seq analysis of microbial populations (meta-dRNAseq). Sci Rep 6:35470

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560

Kanehisa M, Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*: 27–30.

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

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36

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

Li M, Baker BJ, Anantharaman K, Jain S, Breier JA, Dick GJ (2015) Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. Nat Commun 6:8933

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*: 10–12.

Mende DR, Waller AS, Sunagawa S, Järvelin AI and others (2012) Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. PLOS ONE 7:e31386

Mitschke J, Georg J, Scholz I, Sharma CM and others (2011a) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis sp*. PCC6803. Proc Natl Acad Sci USA 108:2124–2129

Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM (2011b) Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena sp*. PCC7120. Proc Natl Acad Sci USA 108:20130–20135

Pfreundt U, Miller D, Adusumilli L, Stambler N, Berman-Frank I, Hess WR (2014) Depth dependent metatranscriptomes of the marine pico-/nanoplanktonic communities in the Gulf of Aqaba/Eilat during seasonal deep mixing. Mar Genomics 18:93–95

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: A matter of depth. Genome Res 21:2213–2223

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42:D553–D559

Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. Nucleic Acids Res 39:W412–W415