Metagenomic Analysis of the Saanich Inlet at a Depth of 120m

Sahi Hajirawala, Richard Kunze, Diana Lin, Emily Wong, Jessica Zhang

Abstract

Metagenomic and metatranscriptomic data from seawater samples from Saanich inlet collected at a depth of 120 metres on Cruise SI72 were analyzed by genomic binning and subsequent analysis using the generated metagenome-assembled genomes (MAGs). Using pathview on individual MAGS, the transcriptomic activity of the MAGs were mapped to metabolic pathways of interest.

The majority of the organisms in the sample showed signs of methane metabolism (via glycolysis and the TCA cycle), and particular MAGs including *Proteobacteria* and *Rhodospirillales* exhibited nitrogen metabolism distributed throughout the entire nitrogen metabolism pathway. Some other MAGs only expressed nitrogen metabolism transcripts corresponding only to either early, middle, or late stages of nitrogen metabolism. Sulfur metabolism was also seen in several MAGs, correlating with the higher hydrogen sulfide concentration generally seen at deeper depths.

Based on biogeochemical data, the 120 meter depth analyzed in this report is in an oxygen depleted region, bordering the more oxygen rich zones above. Many microbes at this depth are adaptable to chemical changes, as seen in some MAGs, likely using their ability to use alternative terminal electron acceptors at the single-organism level as well as possibly at the level of the whole microbial community. Thus, microbes will be able to adapt to lower oxygen levels by altering their metabolism, which will lead to overall denitrification and significant biogeochemical changes.

Importance

This research furthers the knowledge of underwater microbial community responses, increasing our ability to predict and understand marine responses to large-scale global warming trends and ocean deoxygenation.

Global warming is leading to the increase of ocean deoxygenation and the formation of oxygen-minimum zones (OMZs). Saanich Inlet is a proposed model ecosystem for OMZs, as it is a seasonally anoxic fjord with oxygenation that fluctuates seasonally. The analysis of metagenomic and metatranscriptomic data obtained from Saanich Inlet can help in understanding how these microbes respond to oxygenation changes. The present analysis will

cover a single depth and time, but piecing together the full dataset will help provide a multidimensional model of the microbial community in an oxygen minimum zone.

Introduction

Saanich Inlet

Saanich Inlet, a seasonally-shifting anoxic fjord off the coast of Vancouver Island, British Columbia, is a marine ecosystem that serves as a model for studying biogeochemical microbial responses to ocean deoxygenation. Over the years, geochemical and microbial data from Saanich Inlet water columns has served as a model for studying the relevance of OMZs (1). Causing the presence of OMZs in the Saanich Inlet is the phenomenon of seasonal stratification, followed by renewal. This phenomenon occurs with distinct stratification and anoxic regions at lower depths due to the physical shape of the inlet. Eventually, this stratification dissipates and oxygen levels are restored (1). This transition occurs cyclically and seasonally throughout the year.

At certain depths in OMZs in Saanich Inlet, microbes use alternative terminal electron acceptors (TEAs) for metabolism which can lead to the inadvertent production of nitrous oxide and methane, which are noteworthy greenhouse gases involved in global warming (1). These anoxic, greenhouse-gas producing regions are currently expanding with the progression of global warming. Climate change severely impacts ocean dynamics by increasing stratification and decreasing oxygen solubility in warming waters (2). Furthermore, the hypoxic environment established in these OMZs are known to be toxic to marine creatures like crabs, fish, and other invertebrates (3). This disturbance of marine ecological dynamics in turn threatens the economy of coastal fisheries (3), which are a vital part of British Columbia's economy. This highlights the importance of monitoring geochemical changes and microbial communities in the columns of oxygen-deficient bodies of water (2).

Cultivation-Independent Methods

To effectively study vast amounts of microbial data, most Saanich Inlet studies make use of metagenomic (and metatranscriptomic) data. Although cultivation-dependent methods to study microbes provide concrete experimental data, studying enormous and diverse microbial communities can be difficult. For that reason, cultivation-independent techniques have been used to study microbial community responses (4). Cultivation-independent methods also account for 'unculturable' bacteria and archaea that are limited by the laboratory media used. Likewise, it is very difficult to replicate the exact environmental conditions that microbes are naturally found in, thus a cultivation-independent approach would sample a somewhat accurate depiction of microbial species in their most 'natural' state. This is evident in a study where soil bacteria identified in cultured lab media was significantly different (and less) than that identified by analysis of 16S ribosomal ribonucleic acid (16S rRNA) genes by culture/cultivation-independent methods (4). In general, cultivation-independent methods

provide a broad view of the microbial community in an ecosystem but to prove inter-species interactions and/or specific bacterial functions and mechanisms, a cultivation-dependent approach may be required. For this project, our team analyzed metagenomic data at a 120 metre depth in Saanich Inlet to discover any significant pathways and/or microbial species at that specific depth.

Metagenomics and Metagenomic Binning

Metagenomics is a cultivation-independent approach that involves sequencing gathered DNA from a bulk sample to analyze the diverse genomes of microbes present in an environmental sample (5). First, the environmental sample to be studied must be collected for the DNA to be extracted and subsequently sequenced. Then, all of the fragments of genomic DNA are assembled into what is called the 'metagenome'. These metagenome assemblies (or metagenome-assemble genomes = MAGs for short) are highly fragmented and contain thousands of 'contigs', which are sets of overlapping segments that conglomerate into a consensus region of DNA (6). These contigs are subsequently grouped into 'scaffolds'. To establish microbial species from these fragmented genomes, metagenomic binning is required to group the contigs. For metagenomic binning, either a supervised or an unsupervised method can be used. A supervised method uses already sequenced genomes from online databases to label the contigs into taxonomic groups, whereas unsupervised methods detect naturally-forming groups within the data itself (6). Regardless of the method used, both require a metric that defines the similarity between given contigs and bins, as well as an algorithm that converts the previously-mentioned similarities into assignments (6). Using *k*-mer frequencies as a metric is a way to identify intrinsic nucleotide patterns to group contigs into discrete sequencing bins (6). This contig clustering method is statistically best used with tetramers. Grouping contigs follows the principle that contigs originating from the same genome will likely have similar coverage values within each metagenome (6). After all the metagenomic data has been processed, the assemblies can be taxonomically and functionally annotated for further taxonomic and metabolic profiling.

Pros, Cons, and Alternatives to Metagenomics

This technique can be useful in providing information about the taxonomic diversity and in providing some insight on the physiology/function of microorganisms present in that environment. Like other cultivation-independent methods, it takes into account the total microbial community, even including unculturable organisms. This surmounts any limitations from cultivation-dependent studies where some unculturable organisms are disregarded. However, due to the expansive data gathered from metagenomics, it is more difficult to interpret the complex data and requires complicated computational methods. Likewise, processing terabytes of data requires sufficient computational hardware, which can be very expensive. Trying to assemble large amounts of sequenced data from highly diverse communities can also lead to chimeric contigs and polymorphisms (genetic variants), which complicate microbial community analysis. Another problem of analyzing a highly diverse microbial community is that it may be difficult to study the less-abundant members of the community. One alternative

cultivation-independent method is single-cell genomics, in which only the genomic data of individual cells are assembled and analyzed (not the whole community). Although single-cell genomics is often complementary to metagenomics, it does have its own advantages. For example, single-cell genomics provides information on the expression levels of each gene per genome across thousands of individual cells (7). This can assist the discovery of novel genes and pathways that dictate a single cell's function and other modalities (7).

For our project, we analyzed and interpreted metagenomic data derived from the microbial and geochemical components at a 120-m depth in Saanich Inlet. Previous research indicates that this depth is a noteworthy transitional region (8) between oxic and anoxic environments. We predict that both aerobic and anaerobic metabolism will be present and decided to investigate the nitrogen, sulfur, and methane metabolism pathways of various microbial species present at this depth.

Methods

Overview

The workflow for this project is summarized in *Figure 1*, with *Table 1* depicting the GitHub repositories and publications of all bioinformatics tools used. All computations were run on the provided ORCA (9) server, in our projects directory */projects/micb405/project1/Team10/project2/*. All scripts can be found in the following repository forked from EDUCE-UBC/MICB405: https://github.com/dy-lin/MICB405/tree/v1.0. In the *Project2/R* directory are the R scripts (and input/output datasets) and bash scripts used for data processing, and in the *Project2* directory are the bash scripts used to run the tools. All scripts will be hyperlinked to our forked GitHub repository.

Getting the Reads

To get the reads from the Sequence Reads Archive (SRA) (10), the SRA toolkit (11) was used, with the command *fastq-dump* (*--split-files --gzip)*. The metagenomic sample was taken from the Saanich Inlet, at depths of 10m, 100m, 120m, 135m, 150m, 165m, and 200m, by research cruises SI042, SI048, SI072, SI073, SI074, and SI075. For the analysis portion of this project, we will be focusing on the data at a depth of 120m taken by research cruise SI072.

The input for *fastq-dump* are SRA accession numbers, and the output files are the raw reads.

Processing the Reads

Next, the reads needed to be processed. The low-quality bases located at the ends of the reads were trimmed using Trimmomatic (12) (*ILLUMINACLIP:/usr/local/share/Trimmomatic-0.35/adapters/TruSeq3-PE.fa:2:3:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36*). At the same time, entire reads of low Phred

quality scores were discarded. This process ensures that only high quality reads remain in the dataset.

The full Trimmomatic command with all parameter specifications can be found in lines 123-134 in the bash script *launch_MEGAHIT.sh*.

The input files for Trimmomatic are raw reads, and the output files are the trimmed reads.

Metagenomic Assembly

Using MEGAHIT (13) (*--k-min 27 --k-step 10 --memory 0.30 --min-contig-len 500*), the remaining high quality reads were assembled into metagenomic contigs. MEGAHIT was chosen for this particular project because it uses low amounts of memory, has a fast run-time, yet still yields high-quality results.

The full MEGAHIT command with all parameter specifications can be found in lines 198-208 in the bash script *launch_MEGAHIT.sh*.

The input files for MEGAHIT are reads, and the output files are the assembled contigs.

Deduplication

In order to build representative bins, all the metagenomic assemblies (MAGs) were merged into one FASTA file, where length and identity filtering occurs. At this stage, contigs shorter than 1500bp were discarded, as well as sequences that shared high identities, using the bash script *dedupe.sh* (*--Xmx20g --minidentity=99*) of the toolkit BBTools (https://jgi.doe.gov/data-and-tools/bbtools/). With this script, for each group of sequences that shared >=99% identity, the longest representative sequence was kept while the rest of each group was discarded. Reducing redundancy in the contigs helps reduce the amount of memory and run-time required for assembly, as well as helps decrease the complexity of the assembly graphs.

This deduplication process can be found in lines 83-101 in the bash script *bin_multi.sh*.

The input file to *dedupe.sh* is contigs resulting from MEGAHIT, and the output files are the ‘deduplicated’ contigs.

Contig Binning

Using MetaBAT2 (14), the contigs were binned to represent their operational taxonomic unit (OTU) (15). In MetaBAT2, the tetranucleotide frequencies (TNF) of each contig and differential abundance values are taken into account when binning. This process ensures that the

downstream metagenomic assembly of each bin will be representative of a group of closely related organisms.

The full MetaBAT2 command with all parameter specifications can be found in lines 156-161 in the bash script *bin_multi.sh*.

The input files for MetaBAT2 are the contig files resulting from *dedupe.sh*, and the output files are the binned contigs.

Then, RPKM values were calculated for all the binned contigs using BWA (16) and RPKM (17), with the metagenomic reads and the metatranscriptomic reads. These are then plotted using the R script *rpkm.R* using *ggplot2* and *dplyr* of the tidyverse R package (18). The MAGs in these RPKM files were renamed using the bash script *rename-mags.sh*.

The full RPKM command with all parameter specifications can be found in lines 251-281 in the bash script *bin_multi.sh* (and the BWA command with all parameter specifications can be found in the previous lines 259-273).

The input files for RPKM are the binned contigs resulting from MetaBAT2 and the alignment of the reads against the binned contigs (SAM file), and the output file is a CSV file containing the RPKM values for each binned contig.

Assembly of the Bins

To assemble MAGs of closely related organisms (resulting from contig binning), each of the bins was reassembled using minimus2 (19) (*--D REFCOUNT=0 -D OVERLAP=200 -D MINID=95*) of the toolkit AMOS (A Modular Open-Source Assembler) (20).

The full re-assembly commands with all parameter specifications can be found in lines 176-178 in the bash script *bin_multi.sh*.

The input files for minimus2 are the binned contigs resulting from MetaBAT2 and the output files are the reassembled MAGs.

Quality Control

To determine the completeness and contamination of our MAGs, checkM (21) was used, allowing the classification of each MAG into a low, medium or high quality MAG. Low quality MAGs are located in the *LowQ_MAGs.tar.gz* file, whereas the medium and high quality MAGs are located in the *MedQPlus_MAGs* directory. The high quality MAGs were then filtered out of this set by thresholding at >90% completeness and <5% contamination using the bash script *find-hq.sh*.

The full checkM command with all parameter specifications can be found in lines 206-212 in the bash script *bin_multi.sh*. The separation of low quality MAGs from medium and high quality MAGs can be found in lines 284-298 in the bash script *bin_multi.sh*. The checkM output files also had the MAGs renamed using the bash script *rename-mags.sh*.

The input files for checkM are MAGs resulting from minimus2, and the output files are two TSV files containing contamination and completeness percentages, one for *Bacteria* and one for *Archaea*.

The contamination vs completeness plots (*Figures 2-3)* were made with *ggplot2*, after data wrangling with *dplyr*, both from the tidyverse R package, with the R script *QC.R*.

Taxonomy Assignment

The GTDB-Tk (22), the Genome Taxonomy Database Tool Kit, was used to assign taxonomy to each medium quality or higher MAG. In this case, each MAG was assigned a kingdom of *Bacteria* or *Archaea*.

The full GTDB-Tk command with all parameter specifications can be found in line 305 in the bash script *bin_multi.sh*.

The input files for GTDB-Tk are the high quality MAGs filtered out by checkM metrics and the output file is a TSV file with taxonomic lineages for each MAG.

ORF Prediction

Then, for each medium quality or higher MAG at our assigned depth of 120m, open reading frames (ORFs) were predicted using Prokka (23) (*--kingdom $kingdom --cpus 8 --locustag $mag*). Since the *--kingdom* parameter needs to be specified, the Bacteria and Archaea MAGs (classified by GTDB-Tk), were run separately with Prokka. The parameter *--locustag* was added so the MAG in which each ORF originates is clear for downstream analysis, and to be consistent with the renamed MAGs in the RPKM files.

The full Prokka command with all parameter specifications can be found in the bash script *run-Prokka.sh*.

The input files for Prokka are the medium and high quality MAGs filtered by checkM metrics and the output files are annotated ORFs for each MAG.

Upset plots *(Figures 4-9)* showing the overlap of Prokka-annotated genes across high quality MAGs were generated using the R package UpsetR (24), with the R script *plot-upset.R*. The data wrangling to put the data in the correct format for UpsetR was done using the bash script *get-hq-genes.sh*. The bash script *summarize-prokka.sh* was used to extract all genomic features

of the high quality MAGs from the Prokka-generated GFF files. The summary plots and tables of the Prokka output were generated using the R script *prokka-summary.R*.

Metabolic Reconstruction

The resulting ORFs from each medium quality or higher MAG were then uploaded to the KEGG Automatic Annotation Server, KAAS (25), using default parameters (*--search-program GHOSTX --query-sequences file-upload --dataset GENES --assignment SBH)*, to determine the associated KEGG (26) orthology (KO) numbers with each ORF.

The input files for KAAS are the annotated ORFs from Prokka, and the output file is a TSV file containing a KO number for each ORF.

The output file *SaanichInlet_HQ_MAG120m_ORFs_ko.txt* was then processed for R using the bash script *process-KAAS.sh*, into *SaanichInlet_HQ_MAG120m_ORFs_ko.cleaned.txt.* The same process was repeated for medium quality MAGs, yielding *SaanichInlet_MedQ_MAG120m_ORFs_ko.txt* and *SaanichInlet_MedQ_MAG120m_ORFs_ko.cleaned.txt*.

The upset plots (*Figures 10-11)* showing the overlap of KO numbers across high quality MAGs was generated using the R script *plot-upset.R*. The data wrangling to put the data in the correct format for UpsetR was done using the bash script *gen-upset.sh*. To get the KOs that were unique to each MAG, the bash script *find-KO-MAG-unique.sh* was used.

Metatranscriptomics

The metatranscriptomic reads were aligned to the high quality MAGs using BWA (16) (*bwa mem -t 8 -p*). Then the RPKM value was calculated using RPKM (17).

The full BWA command with all parameter specifications can be found in the bash script *run-bwa.sh*. The full RPKM command with all parameter specifications can be found in the bash script *run-rpkm.sh*.

The input files for BWA are the transcript sequence corresponding to each annotated ORF from Prokka and the reads, and the output file is an alignment (SAM) file. The input files for RPKM were the same transcripts and the BWA resulting SAM file, and the output file is a CSV file containing RPKM values for each nucleotide sequence.

Pathview

The processed KAAS results and the metatranscriptomic RPKM values were fed into Pathview (27), using the R script *prepPathView.R*, to visualize the metabolic pathways. We use the script *generate-prokka-mag-map.sh* to generate a CSV file that maps the MAG to the appropriate Prokka output.

After analysis of our 11 high quality MAGs, we then incorporated our 70 medium quality MAGs into our analysis. Using the metadata given in *MetaBAT2_SaanichInlet_120m_min1500_checkM_stdout.tsv*, we extracted the metadata on the 81 high and medium quality MAGs and grouped them by their marker lineage. Lastly, a script *group_rpkm_orfs.sh* was run to group ORF and RPKM data respectively into each marker lineage present in our MAGs.

The input files for Pathview were the TSV file resulting for KAAS and the TSV file resulting from RPKM, and the output files were the pathway visualization images.

Results

CheckM Results

Using CheckM, we were able to determine which MAGs were of low, medium and high quality. The contamination and completion thresholds for each stratum are delineated in *Table 2*, and the results for each high quality MAG are in *Table 3*. The contamination vs completion comparison, along with RPKM values and taxonomy, is illustrated for every MAG in *Figure 2*.

The main checkM output file is *MetaBAT2_SaanichInlet_120m_min1500_checkM_stdout.tsv*.

GTDB-Tk Results

Using GTDB-Tk, we assigned each high quality MAG a taxonomy classification. All high quality MAGs have been classified up to, and including order, as shown in *Table 4*. A taxonomic classification for family and genus were not assigned for all high quality MAGs.

The main two output files for GTDB-Tk are two TSV files for the domains *Bacteria* and *Archaea* respectively: *gtdbtk.bac120.classification_pplacer.tsv* and *gtdbtk.ar122.classification_pplacer.tsv*.

RPKM Results

In *Figures 12-13*, comparing the RPKM coverages between the metatranscriptomic and metagenomic reads, it is shown that there are some instances where the metatranscriptomic RPKM value is greater than the metagenomic RPKM value, and vice versa. This illustrates that expression levels from the MAGs do not necessarily correspond with genomic abundance.

The main output files for RPKM of the MAGs can be found in *rpkm_output/.*

Prokka Results

Using Prokka, we annotated the high quality MAGs with 25340 coding sequences (CDS), 8 ribosomal RNA (rRNA) genes, 8 transfer-messenger RNA (tmRNA) genes, and 386 transfer RNA (tRNA) genes, for a total of 25750 genes. For the breakdown of annotated genes per MAG, see *Table 5* and *Figure 14*, made using *dplyr* and *ggplot2* of the tidyverse R package.

In order to see what genes were shared amongst the high quality MAGs, upset plots were generated. Upset plots were generated to see the overlaps in CDS (*Figures 4-5*), rRNA *(Figure 6)*, tRNA *(Figures 7-8)*, and tmRNA *(Figure 9)* genes. In *Figures 4-5*, it is shown that all high quality MAGs share 90 CDS genes, but the largest intersections in the plot are actually CDS genes unique to a single MAG. In *Figure 6*, it is interesting to see that there are two MAGs that do not contain any rRNA genes. In *Figures 7-8*, it is interesting to see that there are so many different tRNAs, and within each tRNA there are many different anticodons, yet a wide range of overlaps among the high quality MAGs. In *Figure 9*, it is shown that there is only one tmRNA gene among all high quality MAGs, vastly different numbers when compared to the number of rRNA genes and tRNA genes. Overall, there are very few genes that are not possessed by one MAG or another, demonstrating that the MAGs complement one another.

The output files for Prokka are located on ORCA in the directory *Prokka_output/HQ_MAGs/* for the high quality MAGs and *Prokka_output/MedQ_MAGs/* for the medium quality MAGs.

KAAS Results

Using the KAAS results (KO number for each annotated gene), upset plots (*Figures 10-11*) were generated. In these upset plots, it is evident that while all high quality MAGs share 189 KO numbers, there are also some KO numbers that are unique to only one out of the eleven high quality MAGs. The metabolic pathways that these KO numbers correspond to are depicted in *Table 6*. Some notable pathways include sulfur, carbon, and nitrogen metabolism among others. All of these are potential metabolic pathways of the high quality MAGs.

Additionally, since the high quality MAGs share 189 KO numbers but 90 CDS genes, it can be concluded that there are some genes that are not shared that share the same KO numbers, therefore sharing similar functions.

The output files for KAAS are located on ORCA in the projects directory under the names *SaanichInlet_HQ_MAG_ORFs_ko.cleaned.txt* and *SaanichInlet_MedQ_MAG_ORFs_ko.cleaned.txt.*

Pathview Results

Using pathview, we first visualized nitrogen, sulfur, and methane metabolism for only the high quality (HQ) MAGs (*Figure 15*, *Figure 16*, *Figure 17*). We notice the general spread of gene

expression in methane metabolism, a focus in assimilatory sulfate reduction, and expression of primarily early nitrate reductases.

Our analysis of medium quality (MQ) and HQ MAGs shows strong, wide expression of glycolysis and TCA pathways in methane metabolism for almost all organisms, especially *Archaea*, *Flavobacteriaceae*, *Rhodospirillales*, and *Proteobacteria* (*Figure 18*, *Figure 19*, *Figure 20*, *Figure 21*), in addition to *Betaproteobacteria* in the Ribulose-P pathway (*Figure 22*), and *Deltaproteobacteria* in Coenzyme-B production (*Figure 23*). MAG133, *Gracillibacteria* shows no expression in methane metabolism.

In nitrogen metabolism, we note that in particular *Proteobacteria* and *Rhodospirillales* (*Figure 24*, *Figure 25*) exhibit widespread gene expression for nitrogen metabolism. *Alphaproteobacteria*, *Euryarchaeota*, and *Betaproteobacteria* show higher expression of early genes in nitrate reduction pathways (*Figure 26*, *Figure 27*, *Figure 28*), and *Deltaproteobacteria* expresses genes in the middle to late stages of nitrogen metabolism pathways (*Figure 29*). Furthermore, a group of *Marinisomatia* appear to strongly express nitrogen metabolism genes, in particular nitrate reductases involved in early nitrate reduction pathways and genes involved in glutamate synthesis and ammonia production (*Figure 30*). Overall, nitrogen metabolism expression is mainly concentrated on nitrate reduction, denitrification, and glutamate synthesis.

In sulfate metabolism, we note *Alphaproteobacteria*, some *Gammaproteobacteria*, and *Rhodobacteraceae* express genes throughout sulfate reduction, with a slight emphasis on assimilatory sulfate reduction (*Figure 31*, *Figure 32*, *Figure 33*). *Deltaproteobacteria* exhibits involvement in early sulfate reduction (*Figure 34*), and *Proteobacteria* appears to primarily use the SOX system (*Figure 35*). We note that *Rickettsiales* expresses virtually no genes in nitrogen or sulfate metabolism, and expressed specific genes in methane metabolism (*Figure 36*). Expression of sulfur metabolism genes is present but not as widespread as in methane or nitrogen metabolism pathways.

We note that not all Pathview images are included in our results, but can all be found on our GitHub repository: https://github.com/dy-lin/MICB405/tree/master/Project2/PathviewResults_HQMQ, grouped into folders by the label given by checkM results.

Discussion

Biological Significance of Results

Saanich Inlet is a seasonally anoxic fjord. For most of the year, the deep waters are anoxic, owing to the Saanich Inlet sill restricting the exchange of water in and out of the inlet. This dissipates during the late summer and early autumn when oxygen levels are renewed into the fjord. This unique feature creates great temporal diversity in microbial species and the metabolic

pathways they utilize. Because of this, the inlet serves as a great model marine ecosystem for studying microbial diversity, particularly in response to shifting oxygen levels.

While there is no specific question that this study attempts to answer, we aimed to attain a better understanding of the microbial communities present at our assigned depth in Saanich Inlet. This study utilized metagenomics to analyze the diversity, abundance, and activity of microbial communities present at our assigned depth of 120 meters. From the metagenomic data, several high-quality MAGs were assembled. Following quality control using CheckM software, the MAGs were classified into low, medium, and high qualities based on completeness and contamination. Taxonomies were assigned to the MAGs using GTDB-Tk, revealing the presence of several bacterial species, including species in the order *Pseudomonodales* and *Microtrichales*. Overall, bacteria in the phylum *Proteobacteria* and the class *Gammaproteobacteria* were most prevalent.

From the metatranscriptomic data, we observe enzymes necessary for denitrification and sulfate reduction metabolic pathways. Only partial metabolic pathways for both denitrification and sulfate reduction were encoded in the metatranscriptomic data.

From pathview assessment with both high quality and medium quality MAGs, we have discovered that methane metabolism, via glycolysis and the TCA cycle, are still significantly expressed by the majority of our predicted organisms, including *Flavobacteriaceae*, *Rhodospirillales*, *Betaproteobacteria*, *Deltaproteobacteria*, and *Proteobacteria*. However, we also note substantial nitrogen metabolic activity, via nitrate reduction pathways, denitrification, and glutamate synthesis, as well as emerging expression in sulfur metabolism, primarily through sulfur reduction pathways and the SOX system. Because many of the species that use methane metabolism for energy also exhibit genes in nitrogen metabolism, we note the possibility that at 120m, the shift between an aerobic and anaerobic environment can be often and fragile, resulting in the presence of mainly species that are adaptable in the way they acquire energy. The emerging sulfur metabolism is consistent with the presence of hydrogen sulfide at deeper depths, and the few species that appear to be specialists in this area, such as a group of *Proteobacteria* utilizing the SOX System, seem to be filling a new “niche” for potential energy metabolism at this depth.

Methods Rationale

For the most part, we followed the recommended workflow (see *Methods* section). However, after generating the Pathview figures, it was evident that there were still some blanks in the metabolic pathways. There were two major issues: through the diagram, it was not evident which MAG was contributing which enzymes, and there were still some enzymes completely missing from the pathways. To resolve these issues, Pathview was re-run using only one MAG at a time, and and the missing enzymes that could reside in the medium quality MAGs were added to Pathview as well. Therefore, our methodology for Prokka, KAAS, RPKM and Pathview

were repeated to include the medium quality MAGs in addition to the high quality MAGs if the data was for medium quality MAGs was available.

Challenges

One of the main challenges encountered in this project relates to the fact that we were picking up someone else's work-- picking it up where they left off. This proved more difficult than originally thought. We were given many output files of upstream progress to work with. This posed a challenge where the data we wanted wasn't always available to us.

Specifically, the taxonomy classification using GTDB-Tk was only conducted on the high quality MAGs, whereas for some of the analysis, it would have been interesting to see the taxonomy classification for the medium quality MAGs as well. Additionally, GTDB-Tk was not available on ORCA, so running it ourselves on the medium quality MAGs was not possible.

Along with the checkM and GTDB-Tk outputs, we were also given various RPKM files to work with: RPKM values for metagenomic read alignment against the MetaBAT2 binned contigs and RPKM values for metatranscriptomic read alignment against MetaBAT2 binned contigs. The metatranscriptomic RPKM file also only contained RPKM values for high quality MAGs, limiting our analysis by omitting medium quality MAGs.

Future Directions

Additional questions we believe warrant further investigation involve comparing metatranscriptomic and metagenomic data over a range of depths and seasons. How do microbial communities and metabolic pathways change with depth and oxygen concentration? How do these communities change throughout the year as Saanich Inlet shifts between oxygenated and anoxic? Data pertaining to the dynamics of these microbial communities seem to be of more value, and could be related to the global phenomenon of ocean deoxygenation as a result of climate change.

Figures and Tables

Figure 1: Bioinformatics Workflow

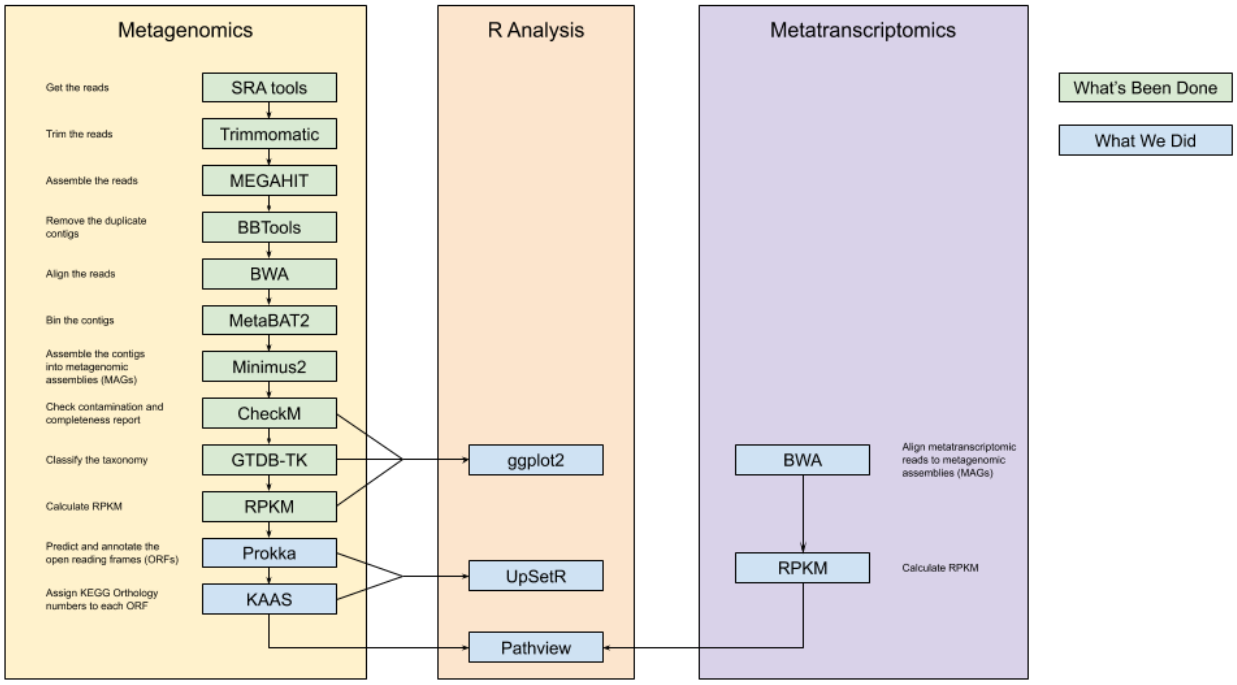

Figure 1 depicts the 3 separate workflows deployed in this project: Metagenomic, metatranscriptomic, and R analysis. Tools coloured in green are parts of the workflow that were pre-done, whereas blue shows parts we did.

Table 1: Bioinformatics Tools

Tool	GitHub	Publication
ORCA	https://github.com/bcgsc/orca	(9)
SRA-tools	http://ncbi.github.io/sra-tools	(11)
Trimmomatic	https://github.com/timflutre/trimmomatic	(12)
MEGAHIT	https://github.com/voutcn/megahit	(13)
BBTools	https://jgi.doe.gov/data-and-tools/bbtools/	—
MetaBAT 2	https://bitbucket.org/berkeleylab/metabat/src/master/	(14)
BWA	https://github.com/lh3/bwa	(16)
RPKM	/projects/micb405/resources/project_2/2019/rpkm	(17)

Minimus2	http://amos.sourceforge.net/wiki/index.php/Minimus2	(19)
CheckM	https://github.com/Ecogenomics/CheckM	(21)
GTDB-Tk	https://github.com/Ecogenomics/GTDBTk	(22)
Prokka	https://github.com/tseemann/prokka	(23)
KAAS	https://github.com/PEHGP/kaas	(25)
Pathview	https://github.com/rforge/pathview	(27)

Table 1 shows the list of bioinformatics tools used in the pipeline, as well as the downstream analysis. Both the publication of the tool as well as the GitHub repository are shown.

Figure 2: Quality Control: Contamination vs Completeness

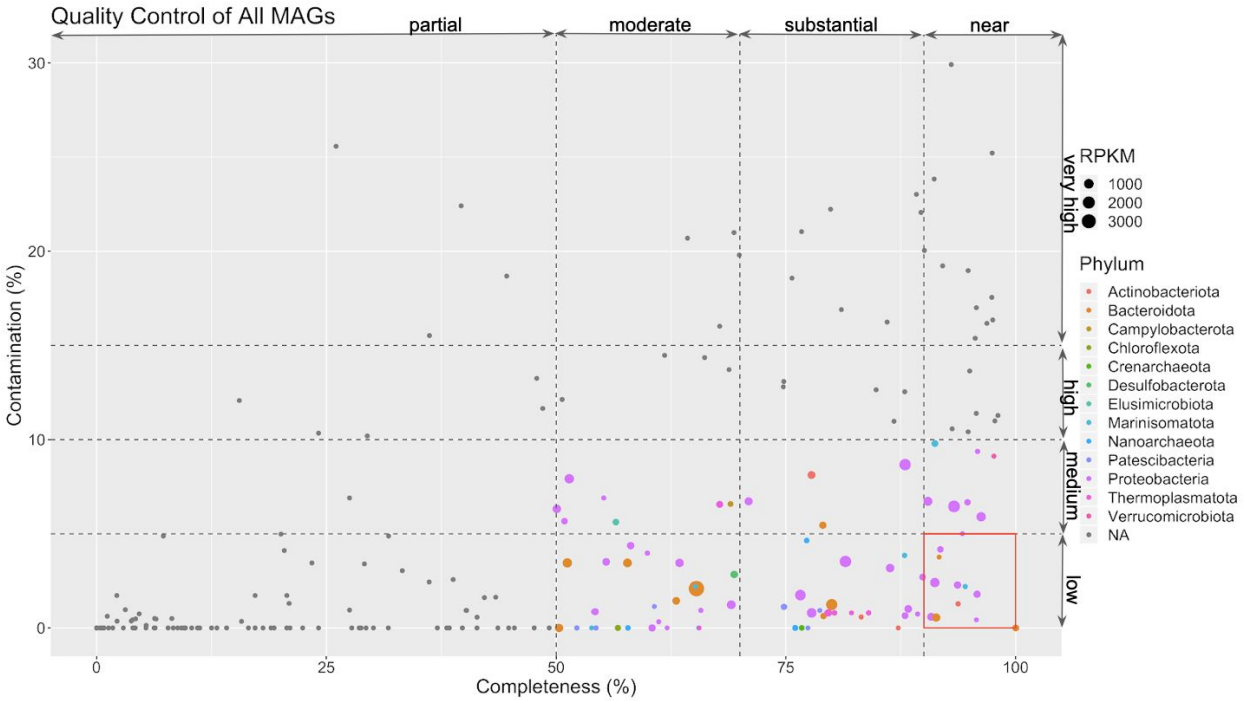

Figure 2 is a scatter plot, comparing the contamination percentage and the completeness percentage between the MAGs. The medium quality and high quality MAGs are colour coded by phylum, and have corresponding RPKM values. The high quality MAGs are contained in the red box.

Figure 3: Quality Control: High Quality MAGs

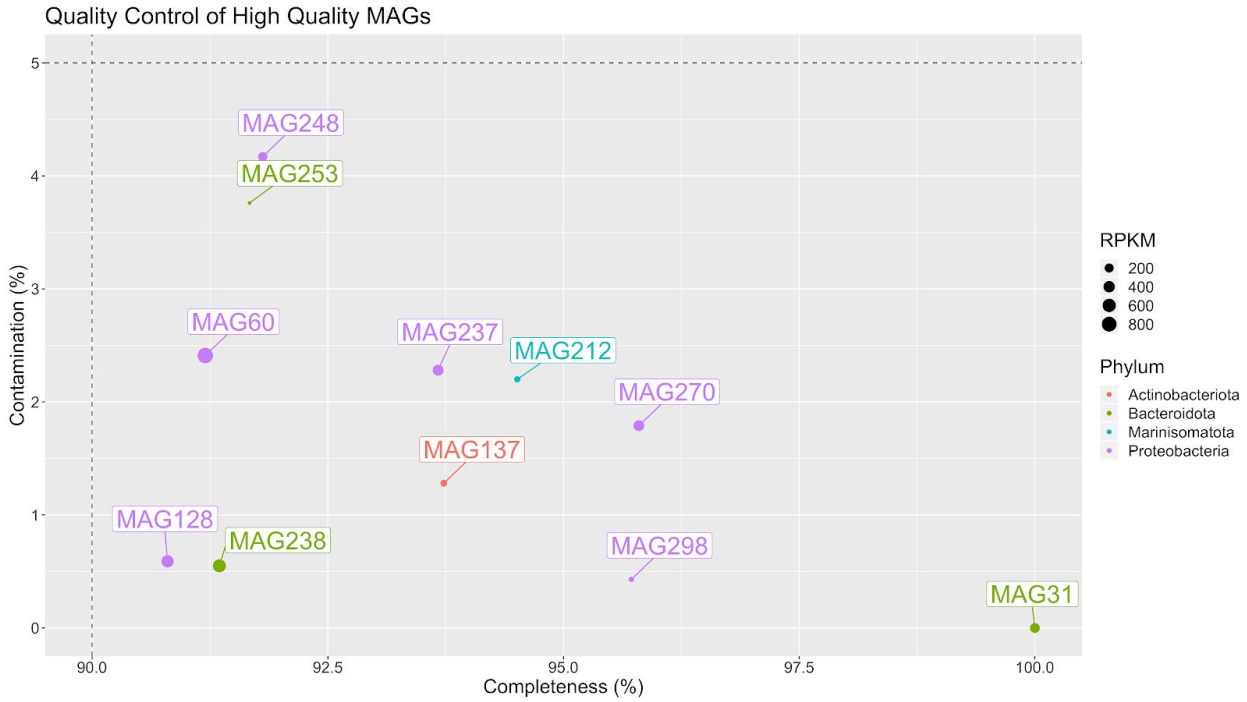

Figure 3 shows a close-up of the contents of the red box in Figure 2, containing all the high quality MAGs.

Figure 4: CDS Overlaps in High Quality MAGs (by frequency)

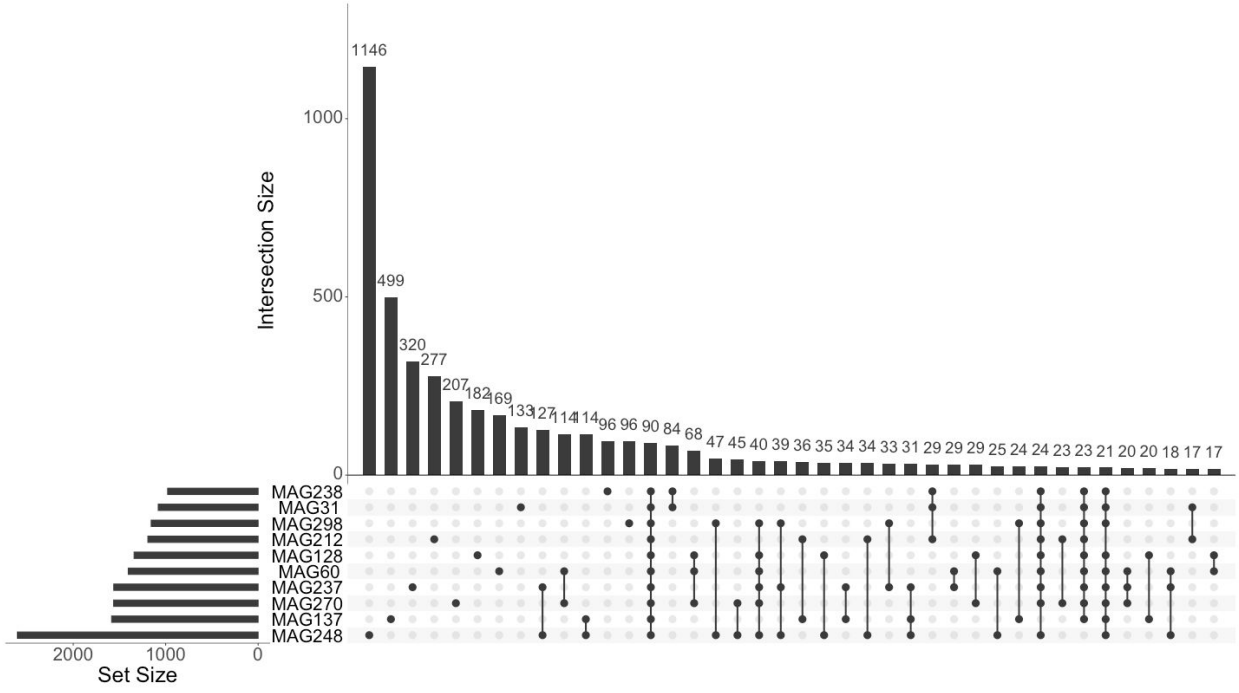

Figure 4 is an upset plot of the Prokka-annotated CDS genes. This plot shows the first 40 intersections by frequency (highest to lowest bar).

Figure 5: CDS Overlaps in High Quality MAGs (by degree)

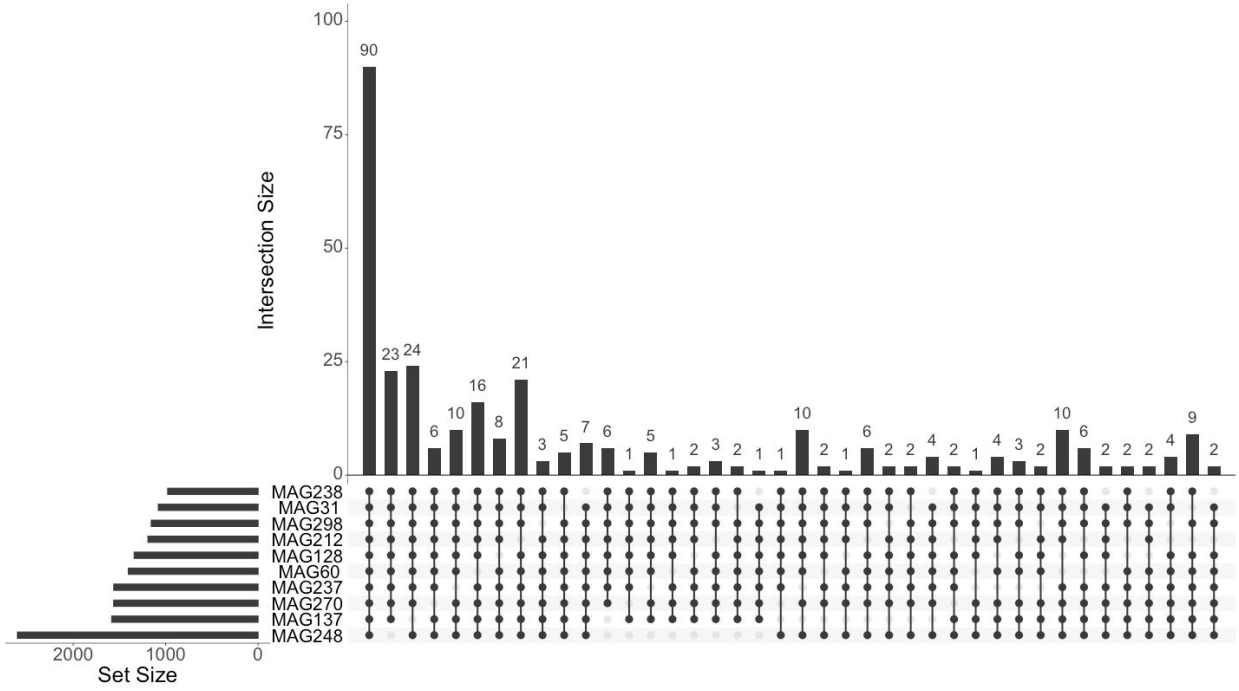

Figure 5 is an upset plot of the Prokka-annotated CDS genes. This plot shows the first 40 overlaps by degree (shared by more MAGs).

Figure 6: rRNA Overlaps in High Quality MAGs

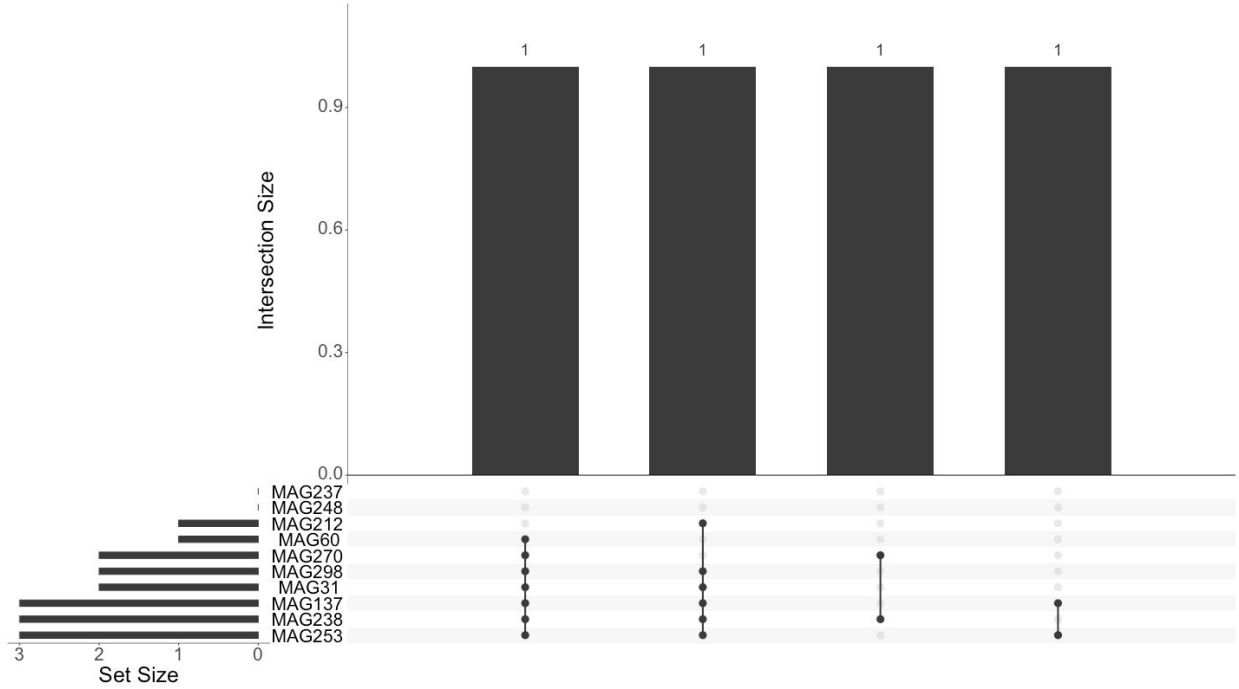

Figure 6 is an upset plot of the Prokka-annotated rRNA genes.

Figure 7: tRNA Overlaps in High Quality MAGs (by frequency)

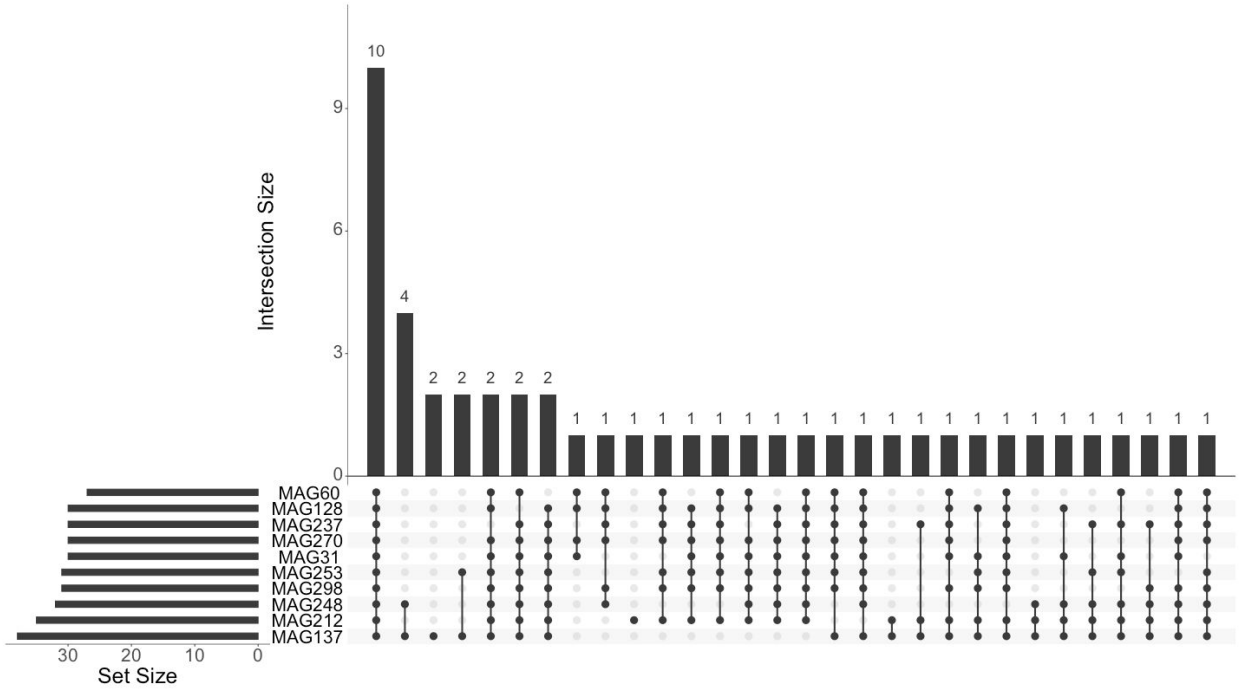

Figure 7 is an upset plot of Prokka-annotated tRNA genes. This plot shows the first 40 intersections by frequency (highest to lowest bars). tRNAs with different anticodons are depicted as separate tRNA genes.

Figure 8: tRNA Overlaps in High Quality MAGs (by degree)

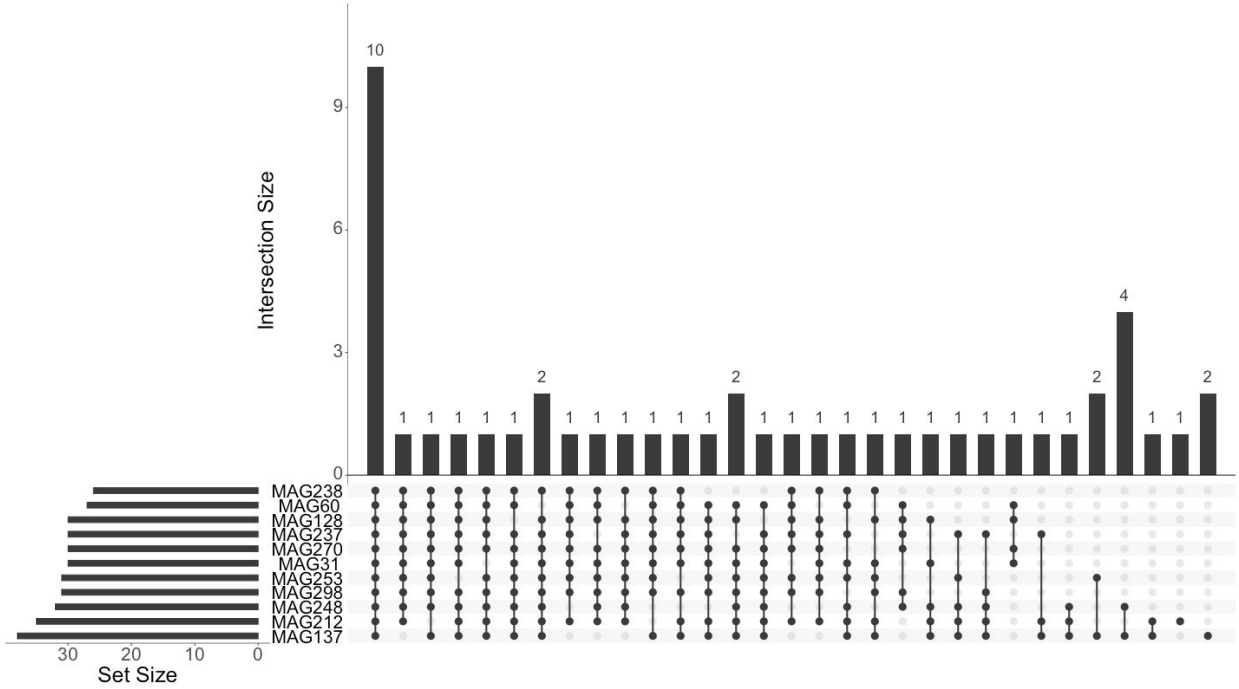

Figure 8 is an upset plot of Prokka-annotated tRNA genes. This plot shows the first 40 overlaps by degree (shared by more MAGs). tRNAs with different anticodons are depicted as separate tRNA genes.

Figure 9: tmRNA Overlaps in High Quality MAGs

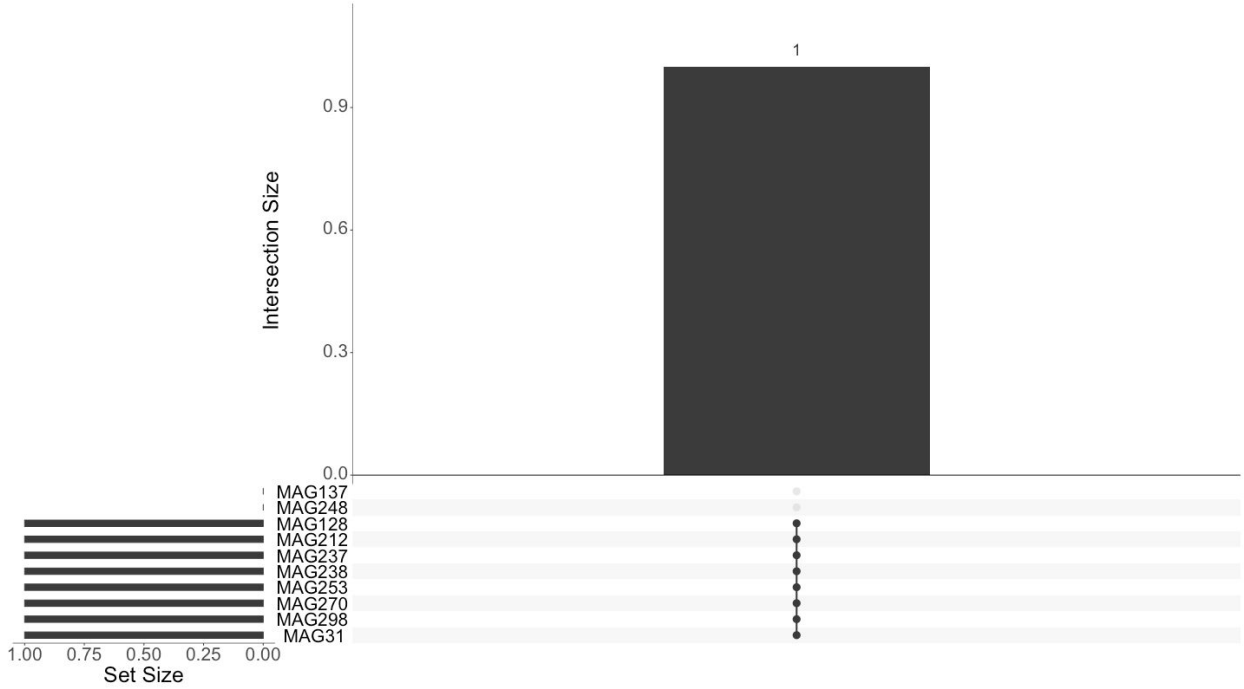

Figure 9 is an upset plot of the Prokka-annotated tmRNAs.

Figure 10: KEGG Orthology Overlaps in High Quality MAGs (by frequency)

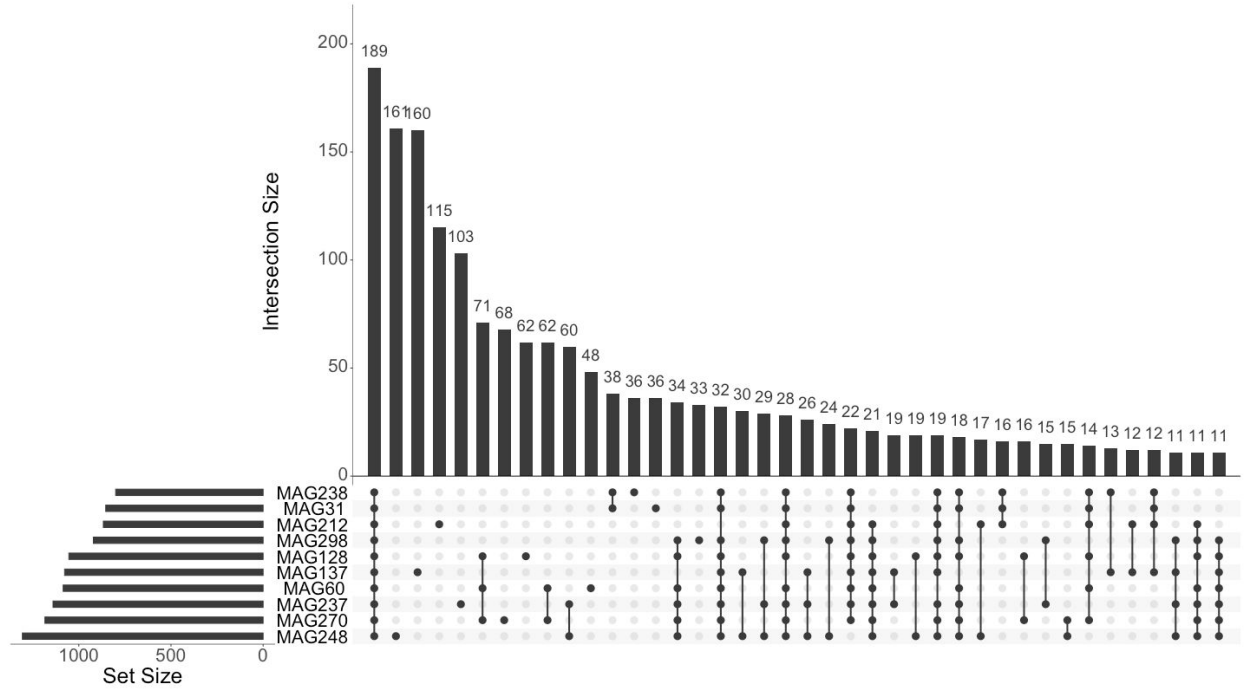

Figure 10 is an upset plot showing the overlaps in KEGG Orthology numbers among the high quality MAGs. This plot shows the first 40 intersections by frequency (highest to lowest bars).

Figure 11: KEGG Orthology Overlaps in High Quality MAGs (by degree)

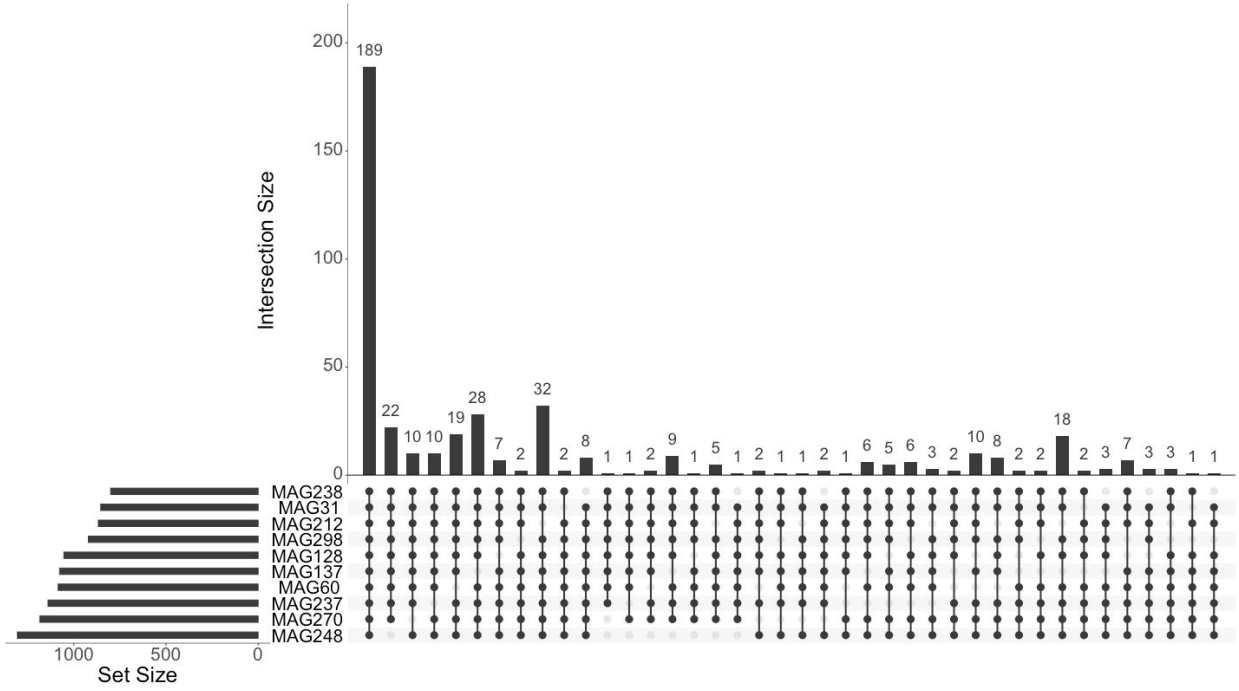

Figure 11 is an upset plot showing the overlaps in KEGG Orthology numbers between the high quality MAGs. This plot shows the first 40 intersections by degree (shared by more MAGs).

Figure 12: Metagenomic vs Metatranscriptomic Coverage of Medium Quality and above MAGs

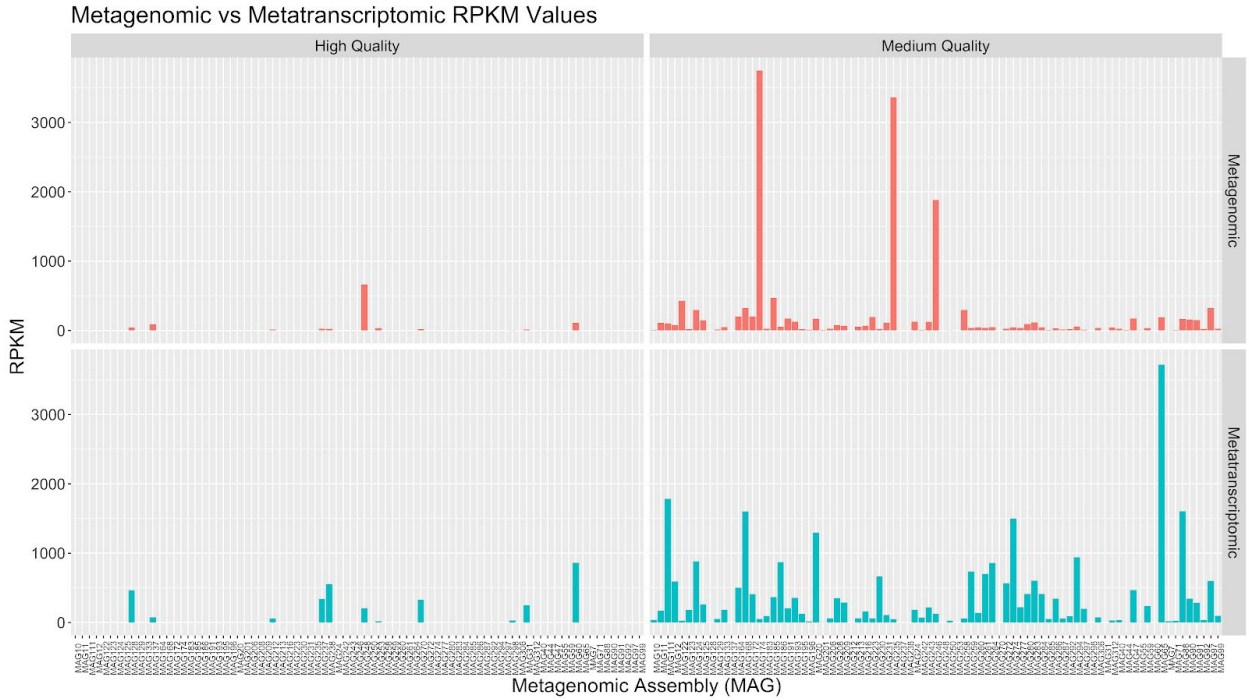

Figure 12 depicts the metatranscriptomic and metagenomic RPKM values across medium and high quality MAGs.

Figure 13: Metagenomic vs Metatranscriptomic Coverage of High Quality MAGs

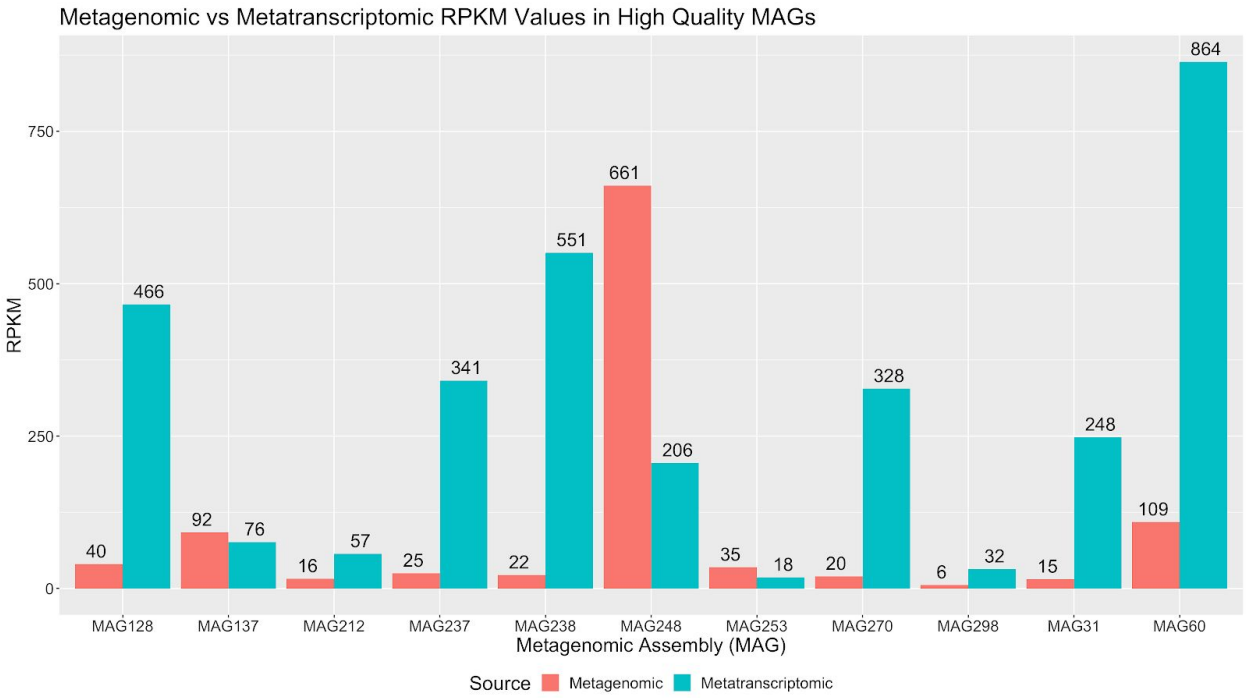

Figure 13 depicts the metatranscriptomic and metagenomic RPKM values across high quality MAGs.

Table 2: MAG Quality Control Classification

Quality of the MAG	Contamination	Completeness
Low	> 10%	< 50%
Medium	5% ≤ *c* ≤ 10%	90% ≥ *c* ≥ 50%
High	< 5%	> 90%

Table 2 shows the contamination and completeness cut offs for each stratum of quality.

Table 3: Contamination and Completeness of High Quality MAGs

MAG	Contamination (%)	Completeness (%)
MAG128	0.59	90.8
MAG137	1.28	93.73

MAG212	2.2	94.51
MAG237	2.28	93.67
MAG238	0.55	91.35
MAG248	4.17	91.81
MAG253	3.76	91.67
MAG270	1.79	95.8
MAG298	0.43	95.72
MAG31	0	100
MAG60	2.41	91.2

Table 3 shows the contamination and completeness percentages of each high quality MAG.

Table 4: High Quality MAGs Taxonomy

MAG	Domain	Phylum	Class	Order
MAG128	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales
MAG137	Bacteria	Actinobacteriota	Acidimicrobiia	Microtrichales
MAG212	Bacteria	Marinisomatota	Marinisomatia	Marinisomatales
MAG237	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales
MAG238	Bacteria	Bacteroidota	Bacteroidia	Flavobacteriales
MAG248	Bacteria	Proteobacteria	Alphaproteobacteria	Sneathiellales
MAG253	Bacteria	Bacteroidota	Bacteroidia	Flavobacteriales
MAG270	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales
MAG298	Bacteria	Proteobacteria	Gammaproteobacteria	Parvibaculales
MAG31	Bacteria	Bacteroidota	Alphaproteobacteria	Flavobacteriales
MAG60	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales

Table 4 shows the taxonomic classification (up to and including Order) for each high quality MAG.

Table 5: Prokka Annotation Breakdown

MAG	# of CDS	# of rRNA	# of tmRNA	# of tRNA	# of genes
MAG31	1813	2	1	34	1850
MAG60	2131	1	0	38	2170
MAG128	2050	0	1	31	2082
MAG137	2991	4	0	39	3034
MAG212	2244	1	1	40	2286
MAG237	2312	0	1	37	2350
MAG238	1626	2	1	29	1658
MAG248	4542	0	0	33	4575
MAG253	1482	3	1	34	1520
MAG270	2417	1	1	34	2453
MAG298	1732	2	1	37	1772
Total	25340	16	8	386	25750

Table 5 shows the total number of predicted genes and for each of the following features: the coding sequence (CDS), ribosomal RNA (rRNA), transfer-messenger RNA (tmRNA), transfer RNA (tRNA).

Figure 14: Prokka Summary

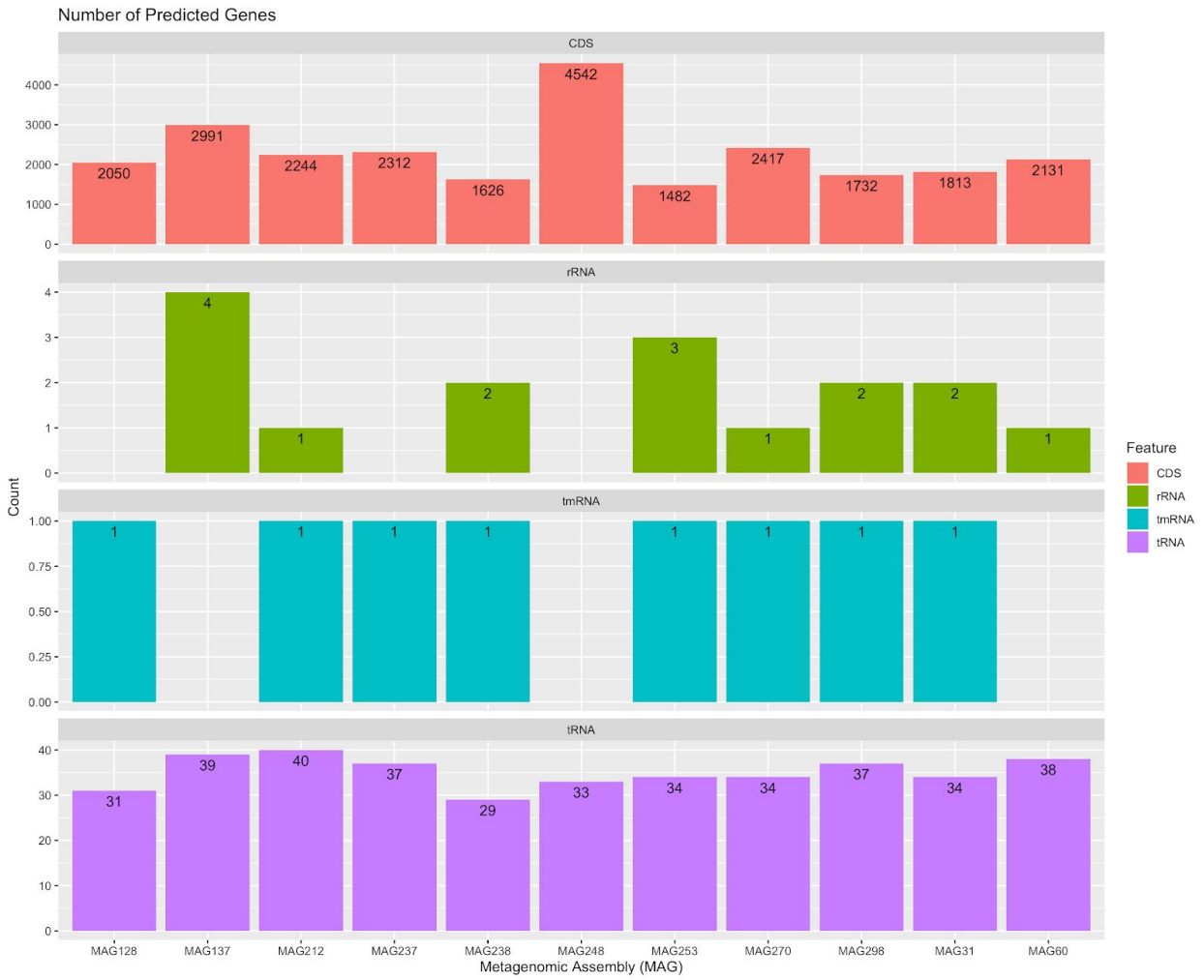

Figure 14 shows the composition of the genes predicted by Prokka.

Table 6: KEGG Orthology Numbers Unique to a Single HQ MAG

MAG	# KOs	# Pathways	Pathways
MAG 128	32	20	Metabolic pathways (6) Two-component system (4) ABC transporters (2) Biosynthesis of secondary metabolites (2) Sulfur metabolism (2) Biosynthesis of secondary metabolites (2) Microbial metabolism in diverse environments (2)

			ABC transporters (2) Purine metabolism (2) PPAR signalling pathway (1) Biosynthesis of unsaturated fatty acids (1) Ascorbate and aldarate metabolism (1) Fatty acid metabolism (1) Longevity regulating pathway - worm (1) AMPK signaling pathway (1) Nicotinate and nicotinamide metabolism (1) Porphyrin and chlorophyll metabolism (1) Cationic antimicrobial peptide (CAMP) resistance (1) Sulfur relay system (1) Axon regeneration (1)
MAG 137	68	39	Metabolic pathways (25) Biosynthesis of secondary metabolites (13) Two-component system (6) Oxidative phosphorylation (4) Microbial metabolism in diverse environments (4) Porphyrin and chlorophyll metabolism (4) Amino sugar and other terpenoid-quinone biosynthesis (2) Inositol phosphate metabolism (2) Glycerolipid metabolism (2) Starch and sucrose metabolism (2) Galactose metabolism (2) Carotenoid biosynthesis (2) Pyruvate metabolism (1) Base excision repair (1) Histidine metabolism (1) Peptidoglycan biosynthesis (1) Biofilm formation - Escherichia coli (1) Caffeine metabolism (1) Glycine, serine and threonine metabolism (1) Alanine, aspartate and glutamate metabolism (1) Drug metabolism - other enzymes (1) Chemical carcinogenesis (1) Riboflavin metabolism (1) RNA degradation (1) Pathways in cancer (1) Valine, leucine and isoleucine degradation (1) Prolactin signaling pathway (1) Lipoarabinomannan (LAM) biosynthesis (1) Hepatocellular carcinoma (1) Steroid biosynthesis (1)
MAG 212	36	55	Metabolic pathways (27) Biosynthesis of secondary metabolites (9)

			Microbial metabolism in diverse environments (8) Two-component system (5) Biosynthesis of antibiotics (4) Butanoate metabolism (4) Nitrogen metabolism (3) ABC transporters (3) Glycolysis / Gluconeogenesis (3) Porphyrin and chlorophyll metabolism (3) Carbon metabolism (2) Inositol phosphate metabolism (2) Phenylalanine, tyrosine and tryptophan biosynthesis (2) Folate biosynthesis (2) Biosynthesis of amino acids (2) Pyruvate metabolism (2) Nicotinate and nicotinamide metabolism (2) Glyoxylate and dicarboxylate metabolism (1) Circadian entrainment (1) Purine metabolism (1) Phenylalanine metabolism (1) Fatty acid degradation (1) Degradation of aromatic compounds (1) Starch and sucrose metabolism (1) Quorum sensing (1) Long-term depression (1) Regulation of lipolysis in adipocytes (1) Naphthalene degradation (1) Photosynthesis (1) Fructose and mannose metabolism (1) cGMP-PKG signaling pathway (1) Arginine biosynthesis (1) Terpenoid backbone biosynthesis (1) Thiamine metabolism (1) Salivary secretion (1) Amoebiasis (1) Tyrosine metabolism (1) Gap junction (1) Carbon fixation pathways in prokaryotes (1) Glycerophospholipid metabolism (1) Alanine, aspartate and glutamate metabolism (1) Citrate cycle (TCA cycle) (1) Riboflavin metabolism (1) Pentose and glucuronate interconversions (1)

			Glycerolipid metabolism (1) Galactose metabolism (1) Selenocompound metabolism (1) Lipopolysaccharide biosynthesis (1) Thermogenesis (1) Vascular smooth muscle contraction (1) Olfactory transduction (1) Peroxisome (1) Platelet activation (1) Glycosaminoglycan biosynthesis - heparan sulfate / heparin (1) Chloroalkane and chloroalkene degradation (1)
MAG 237	35	27	Metabolic pathways (14) Biosynthesis of secondary metabolites (5) ABC transporters (3) Two-component system (3) Microbial metabolism in diverse environments (3) Porphyrin and chlorophyll metabolism (3) Inositol phosphate metabolism (2) Pyruvate metabolism (1) Glycolysis / Gluconeogenesis (1) Glycerophospholipid metabolism (1) Naphthalene degradation (1) Galactose metabolism (1) Biosynthesis of antibiotics (1) Tyrosine metabolism (1) Quorum sensing (1) Riboflavin metabolism (1) Folate biosynthesis (1) Glyoxylate and dicarboxylate metabolism (1) Butanoate metabolism (1) Chloroalkane and chloroalkene degradation (1) Glycerolipid metabolism (1) Pentose and glucuronate interconversions (1) Fatty acid degradation (1) Fructose and mannose metabolism (1) Nicotinate and nicotinamide metabolism (1) Degradation of aromatic compounds (1) Amoebiasis (1)

MAG 238	12	12	Metabolic pathways (2) Peroxisome (1) Two-component system (1) beta-Alanine metabolism (1) Histidine metabolism (1) Oxidative phosphorylation (1) Amyotrophic lateral sclerosis (ALS) (1) Longevity regulating pathway - multiple species (1) ABC transporters (1) Prion diseases (1) Huntington disease (1) Arginine and proline metabolism (1)
MAG 248	77	57	Metabolic pathways (32) Microbial metabolism in diverse environments (11) Carbon metabolism (8) Biosynthesis of secondary metabolites (7) Biosynthesis of antibiotics (5) Two-component system (5) Methane metabolism (4) Oxidative phosphorylation (4) Fatty acid metabolism (3) Bacterial chemotaxis (3) Pentose phosphate pathway (2) Pyruvate metabolism (2) Fatty acid degradation (2) Retrograde endocannabinoid signaling (2) Glycerophospholipid metabolism (2) Propanoate metabolism (2) Butanoate metabolism (2) Citrate cycle (TCA cycle) (2) 2-Oxocarboxylic acid metabolism (2) Nitrogen metabolism (2) Glyoxylate and dicarboxylate metabolism (2) Parkinson disease (2) Biosynthesis of amino acids (2) ABC transporters (2) Valine, leucine and isoleucine degradation (2) Thermogenesis (2) Huntington disease (1) Peroxisome (1) Sulfur metabolism (1)

			Degradation of aromatic compounds (1) Porphyrin and chlorophyll metabolism (1) Nicotinate and nicotinamide metabolism (1) Glycine, serine and threonine metabolism (1) Thiamine metabolism (1) Carbon fixation pathways in prokaryotes (1) Lysine biosynthesis (1) MicroRNAs in cancer (1) Riboflavin metabolism (1) Dioxin degradation (1) Pyrimidine metabolism (1) Fatty acid elongation (1) Gastric cancer (1) Xylene degradation (1) Benzoate degradation (1) One carbon pool by folate (1) Phenylalanine metabolism (1) Lipopolysaccharide biosynthesis (1) alpha-Linolenic acid metabolism (1) Alzheimer disease (1) Non-alcoholic fatty liver disease (NAFLD) (1) Pentose and glucuronate interconversions (1) Valine, leucine and isoleucine biosynthesis (1) Fatty acid biosynthesis (1) Bile secretion (1) Secondary bile acid biosynthesis (1) Pantothenate and CoA biosynthesis (1) Tryptophan metabolism (1)
MAG 253	11	8	Glycosaminoglycan biosynthesis - heparan sulfate / heparin (1) Human T-cell leukemia virus 1 infection (1) Metabolic pathways (1) Biofilm formation - Vibrio cholerae (1) Microbial metabolism in diverse environments (1) Axon guidance (1) Nitrogen metabolism (1) Peptidoglycan biosynthesis (1)
MAG 270	21	5	Metabolic pathways (2) Two-component system (2) Fructose and mannose metabolism (1) Nitrogen metabolism (1)

			Microbial metabolism in diverse environments (1)
MAG 298	15	28	Metabolic pathways (7) Biosynthesis of secondary metabolites (4) Glycine, serine and threonine metabolism (2) Carbon metabolism (2) Amino sugar and nucleotide sugar metabolism (2) Microbial metabolism in diverse environments (2) Biosynthesis of antibiotics (2) Fructose and mannose metabolism (2) Glycosphingolipid biosynthesis - lacto and neolacto series (1) Oxidative phosphorylation (1) Thermogenesis (1) Huntington disease (1) Citrate cycle (TCA cycle) (1) Pyruvate metabolism (1) Isoquinoline alkaloid biosynthesis (1) Phenylalanine metabolism (1) Tropane, piperidine and pyridine alkaloid biosynthesis (1) Ether lipid metabolism (1) Alzheimer disease (1) beta-Alanine metabolism (1) Carotenoid biosynthesis (1) Tyrosine metabolism (1) ABC transporters (1) Glycerolipid metabolism (1) Glyoxylate and dicarboxylate metabolism (1) Non-alcoholic fatty liver disease (NAFLD) (1) Phosphotransferase system (PTS) (1) Parkinson disease (1)
MAG 31	13	14	Metabolic pathways (6) Two-component system (3) Microbial metabolism in diverse environments (2) Fructose and mannose metabolism (2) Phosphotransferase system (PTS) (2) Vancomycin resistance (1) Pertussis (1) Inositol phosphate metabolism (1) Glycerolipid metabolism (1) Amyotrophic lateral sclerosis (ALS) (1) Propanoate metabolism (1)

			Peptidoglycan biosynthesis (1) Sulfur metabolism (1) Ascorbate and aldarate metabolism (1)
MAG 60	22	17	Metabolic pathways (4) Biofilm formation - Vibrio cholerae (3) Bacterial secretion system (3) Flagellar assembly (2) Two-component system (2) Cationic antimicrobial peptide (CAMP) resistance (1) Quorum sensing (1) Amino sugar and nucleotide sugar metabolism (1) Sulfur metabolism (1) Propanoate metabolism (1) Valine, leucine and isoleucine degradation (1) Microbial metabolism in diverse environments (1) Protein export (1) Biosynthesis of antibiotics (1) Pyrimidine metabolism (1) Biofilm formation - Escherichia coli (1) Biosynthesis of secondary metabolites (1)

Table 6 shows the KEGG Pathways in which the KEGG Orthology numbers unique to a single high quality MAG fall. Highlighted in yellow are some interesting pathways.

Figure 15: Nitrogen Metabolism (High Quality MAGs only)

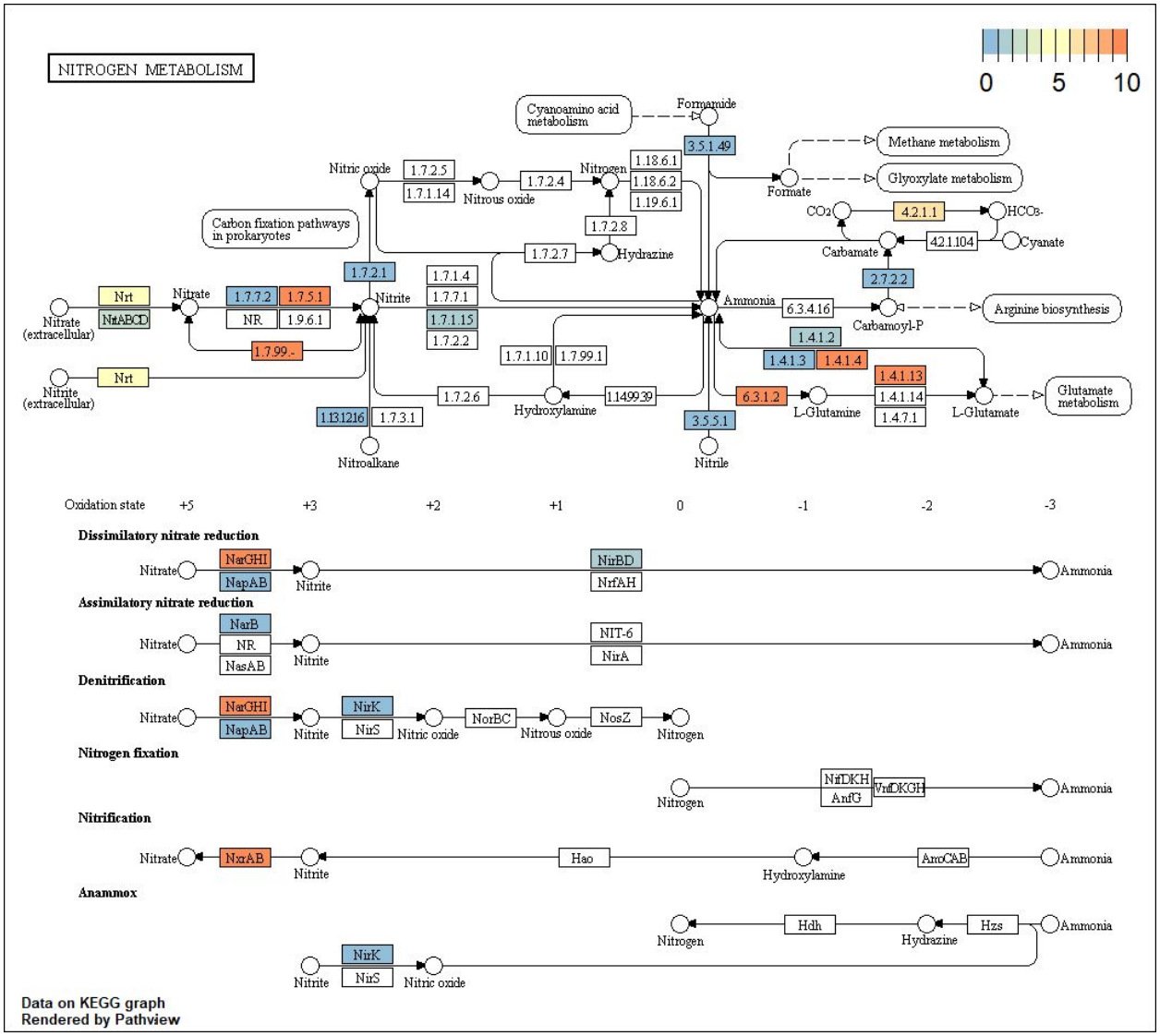

Figure 15 shows nitrogen metabolism for all high quality MAGs.

Figure 16: Sulfur Metabolism (High Quality MAGs only)

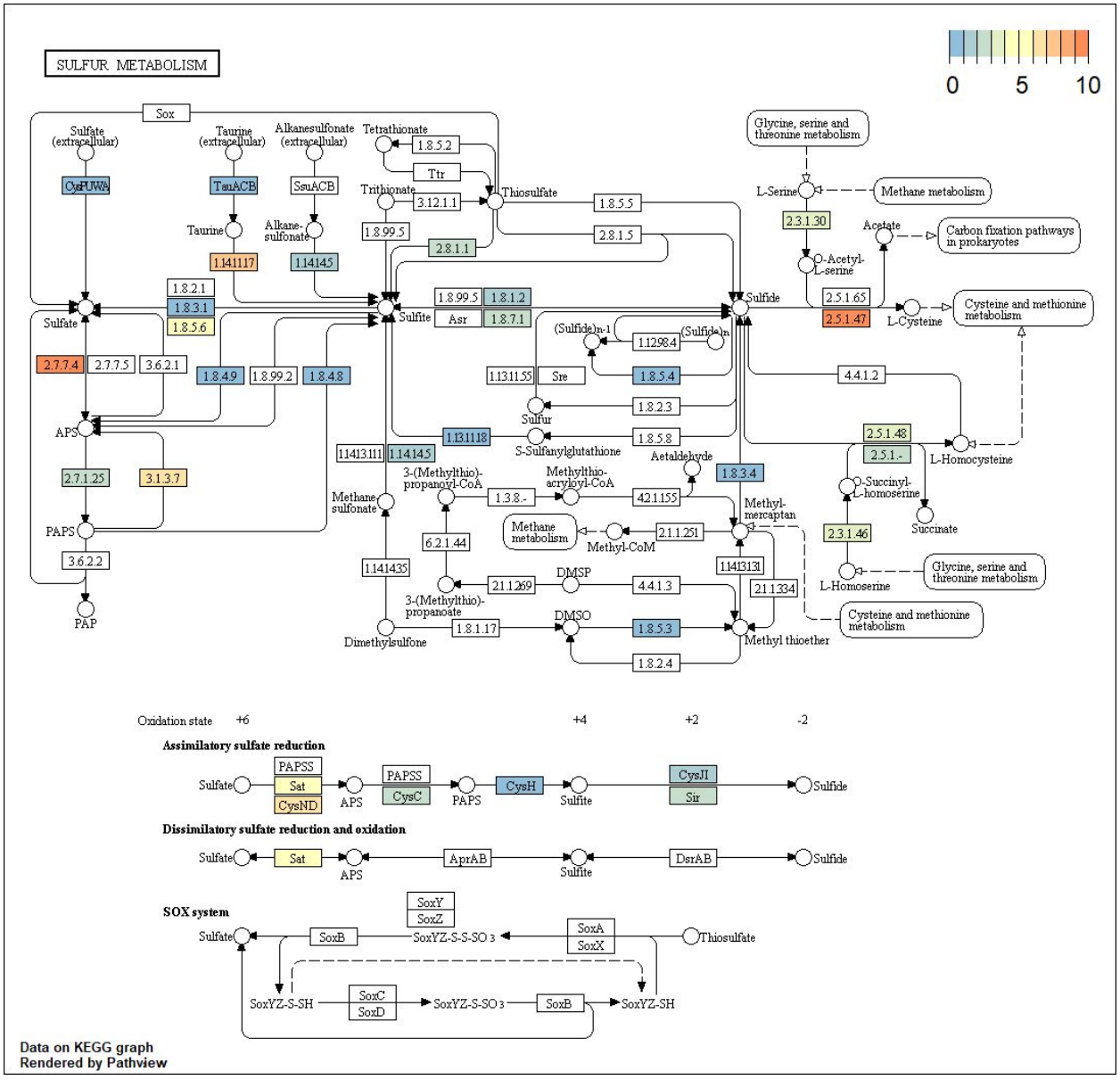

Figure 16 shows sulfur metabolism for all high quality MAGs.

Figure 17: Methane Metabolism (High Quality MAGs only)

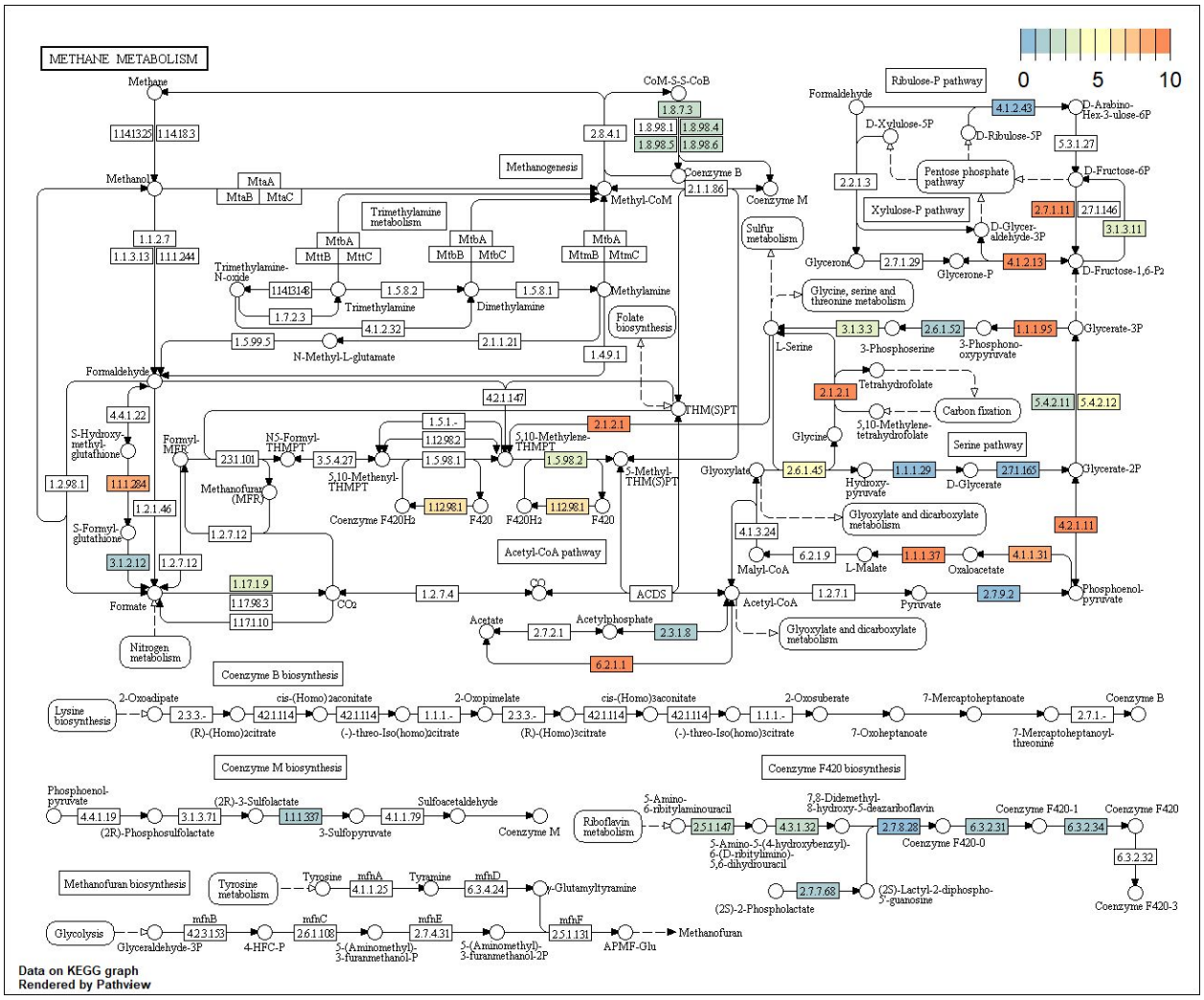

Figure 17 shows methane metabolism for all high quality MAGs.

Figure 18: *Archaea*, Methane Metabolism

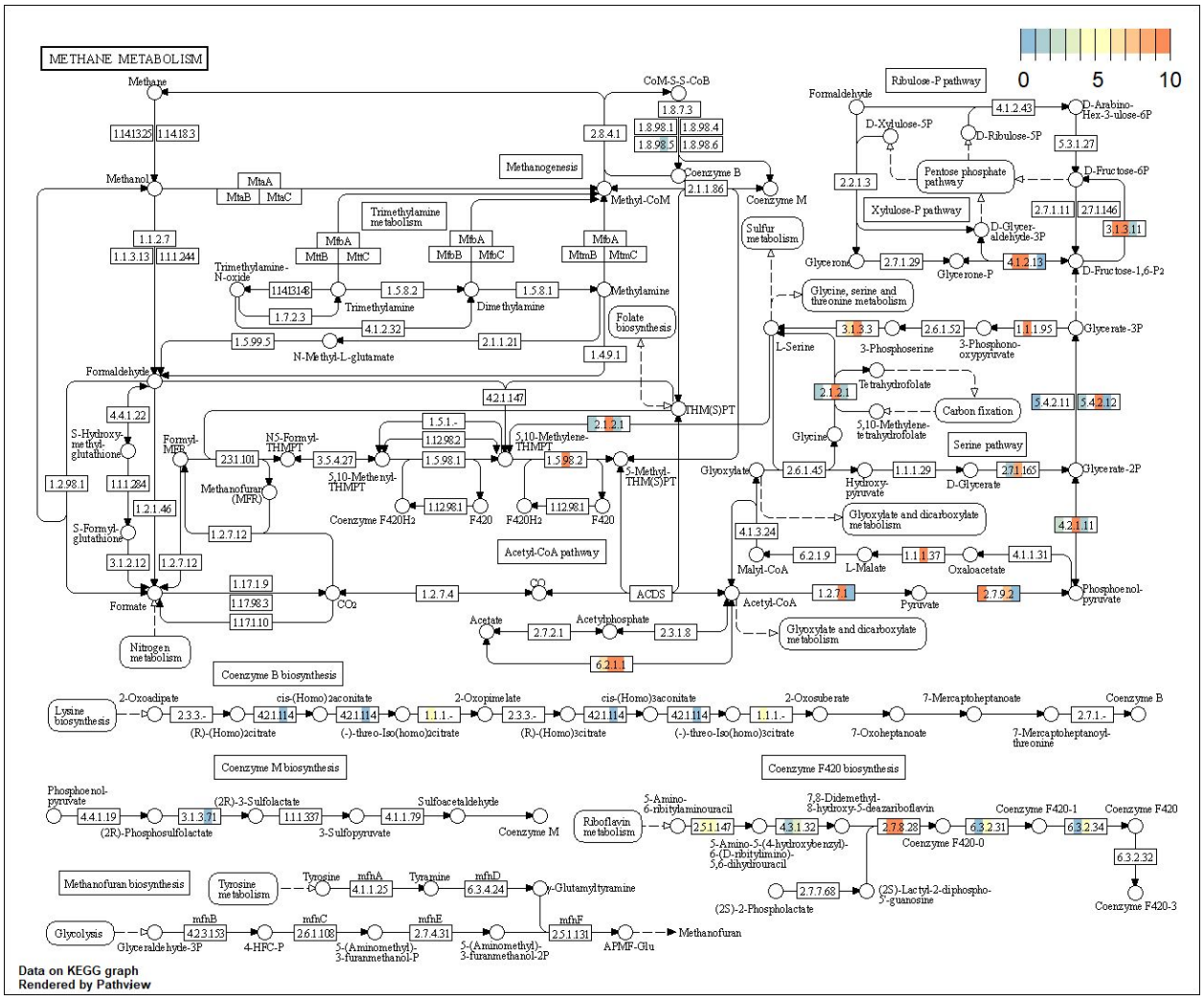

Figure 18 shows methane metabolism for all high and medium quality MAGs labelled as *Archaea* in the checkM results.

Figure 19: *Flavobacteriaceae*, Methane Metabolism

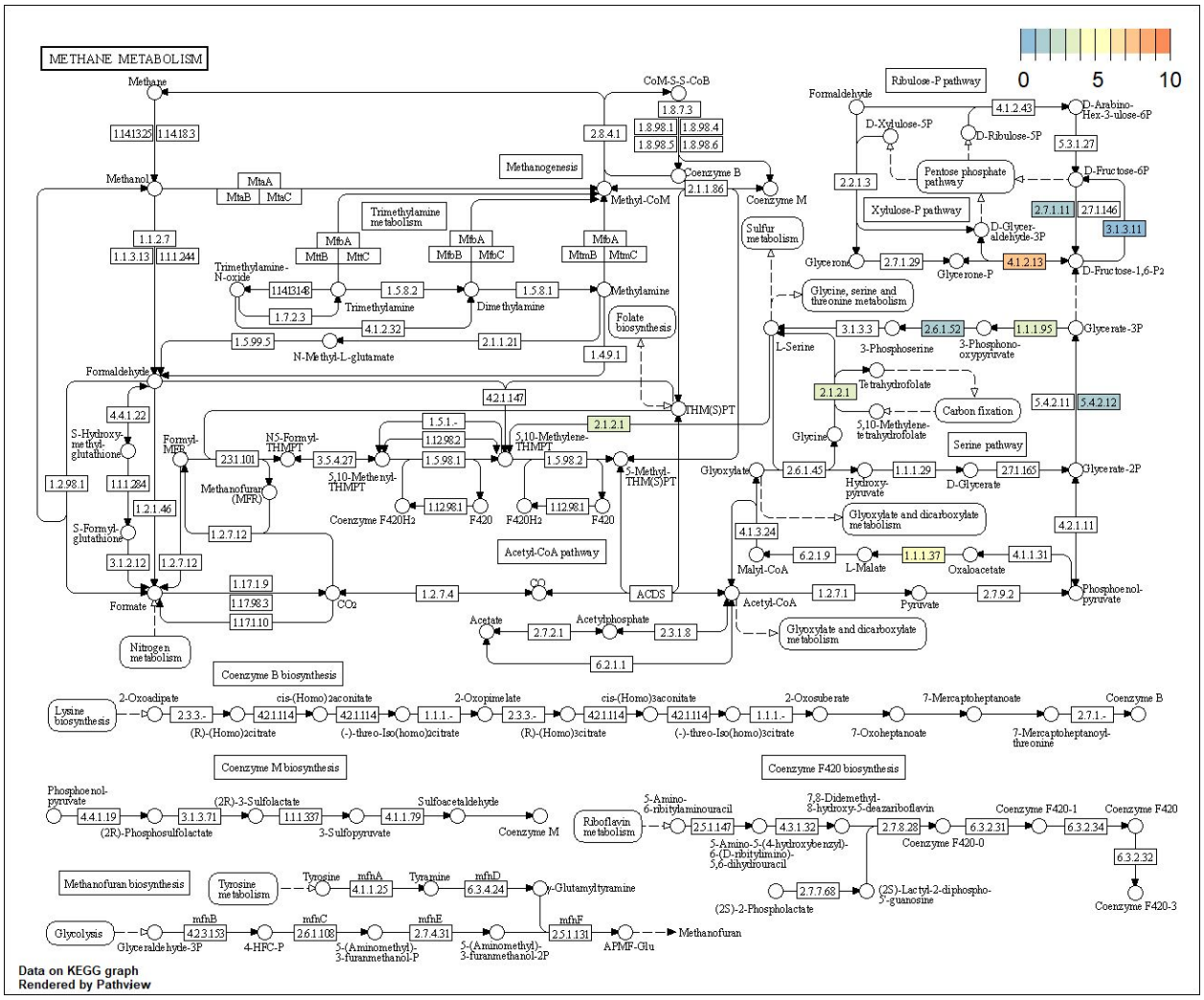

Figure 19 shows methane metabolism for all high and medium quality MAGs labelled as *Flavobacteriaceae* in the checkM results.

Figure 20: *Rhodospirillales*, Methane Metabolism

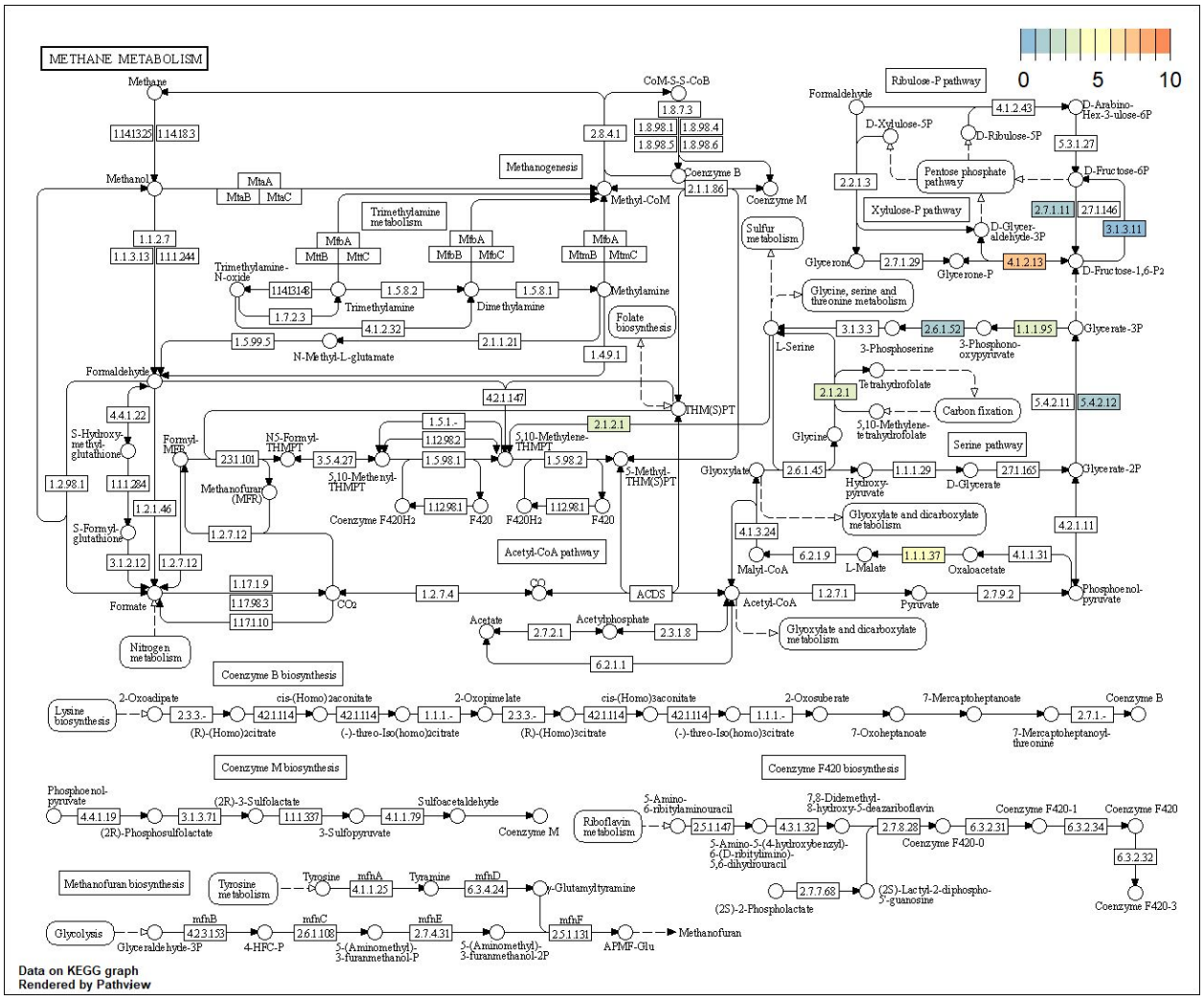

Figure 20 shows methane metabolism for all high and medium quality MAGs labelled as *Rhodospirillales* in the checkM results.

Figure 21: *Proteobacteria*, Methane Metabolism

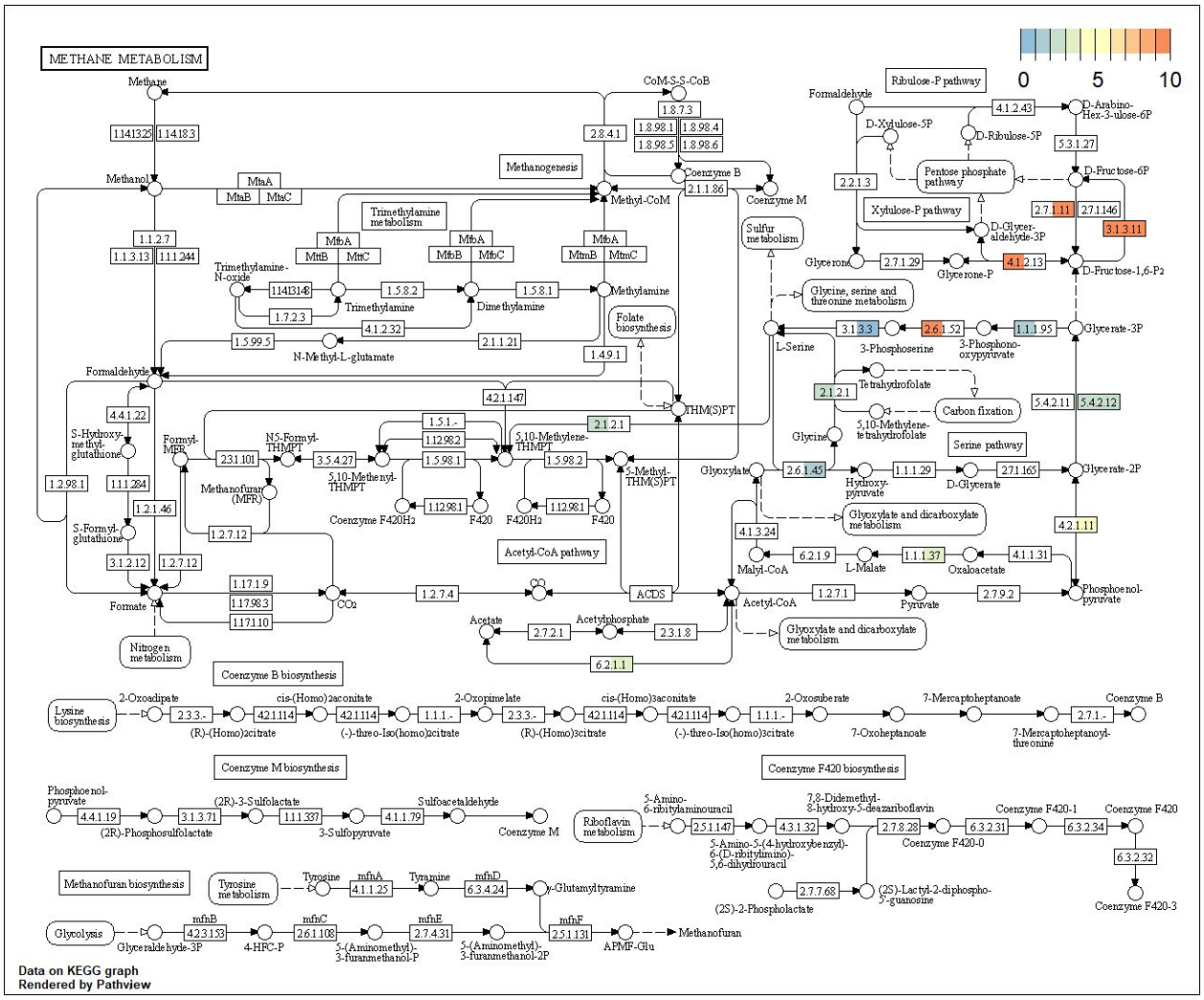

Figure 21 shows methane metabolism for all high and medium quality MAGs labelled as *Proteobacteria* in the checkM results.

Figure 22: *Betaproteobacteria*, Methane Metabolism

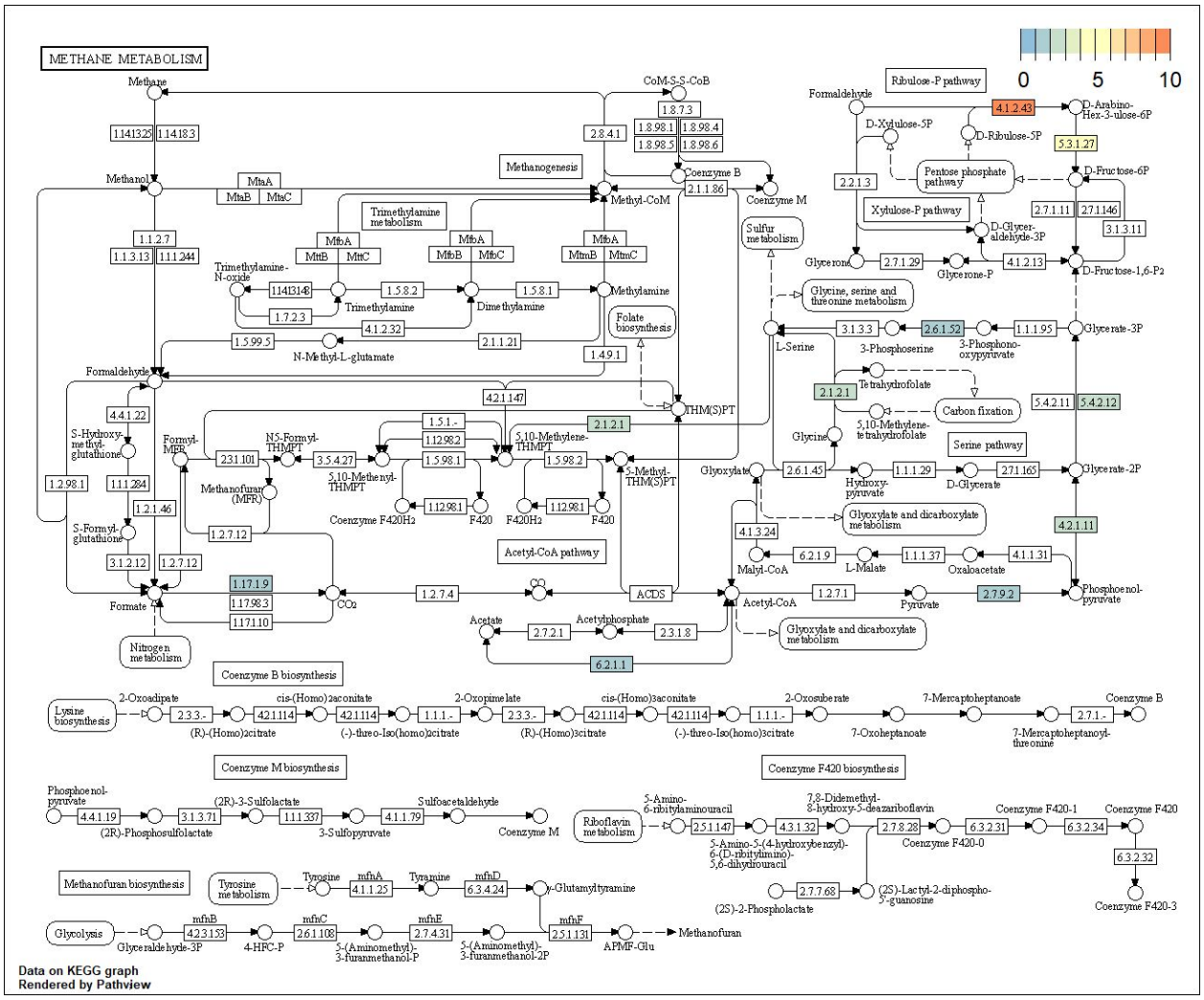

Figure 22 shows methane metabolism for all high and medium quality MAGs labelled as *Betaproteobacteria* in the checkM results.

Figure 23: *Deltaproteobacteria*, Methane Metabolism

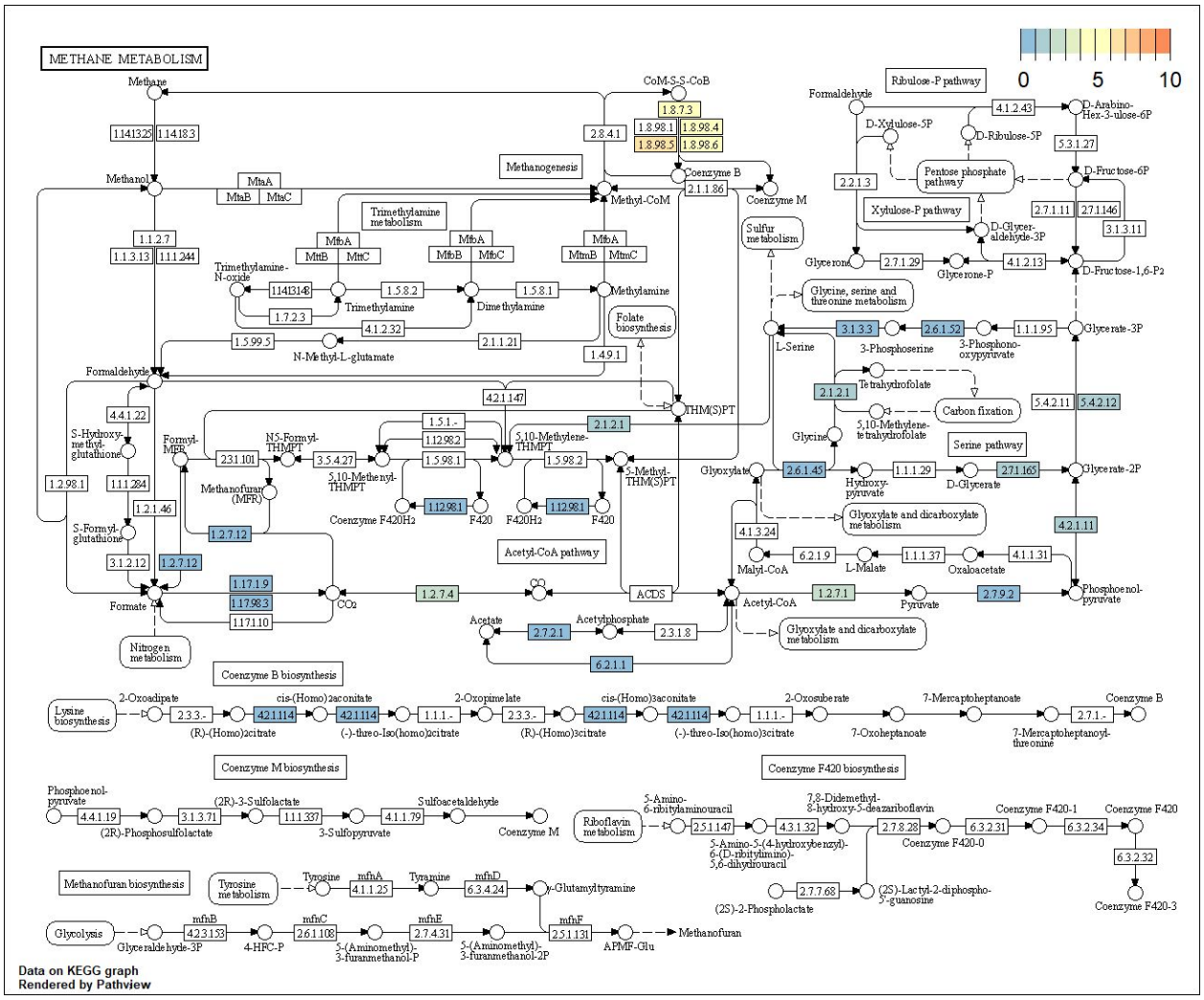

Figure 23 shows methane metabolism for all high and medium quality MAGs labelled as *Deltaproteobacteria* in the checkM results.

Figure 24: *Proteobacteria*, Nitrogen Metabolism

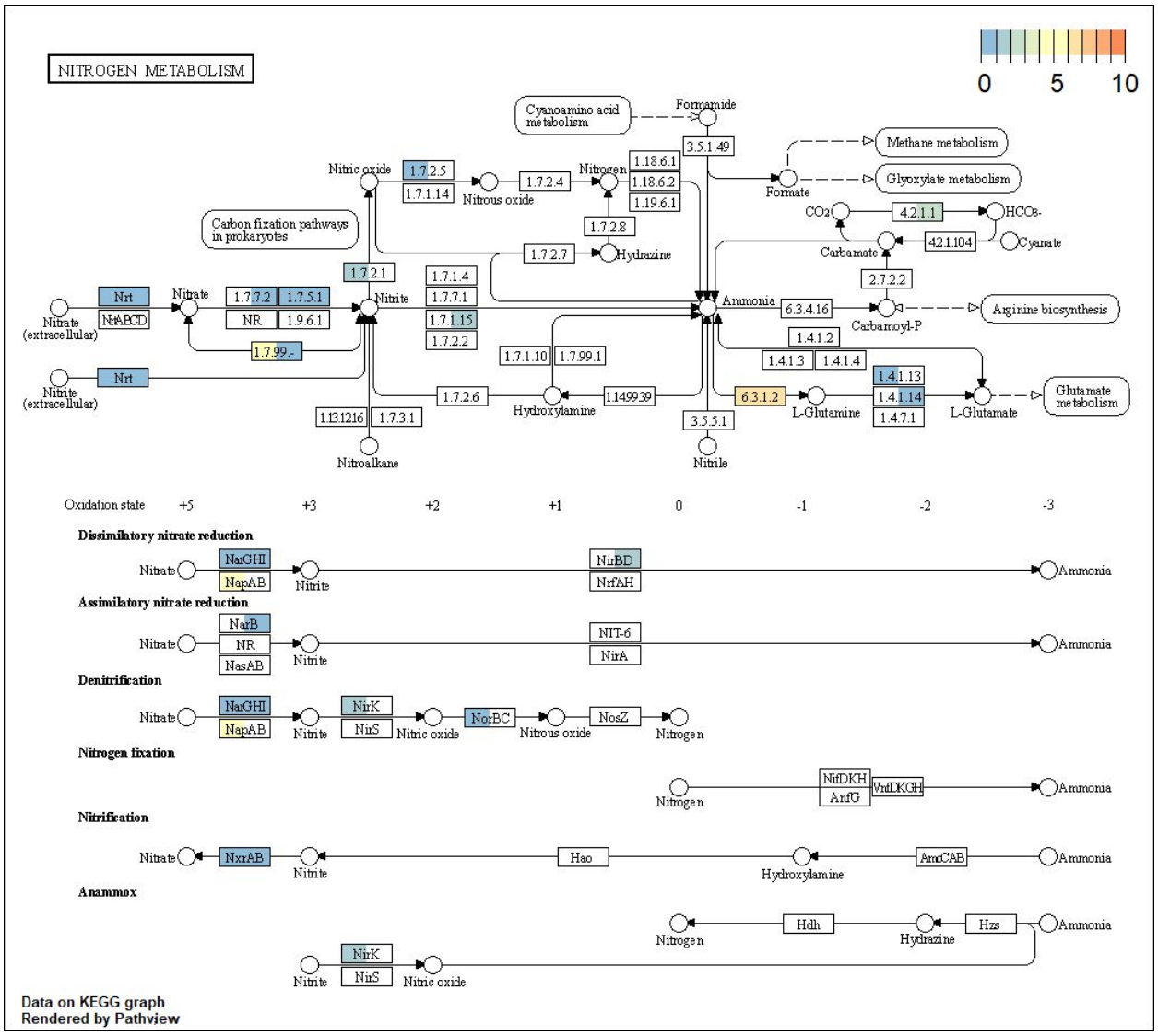

Figure 24 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Proteobacteria* in the checkM results.

Figure 25: *Rhodospirillales*, Nitrogen Metabolism

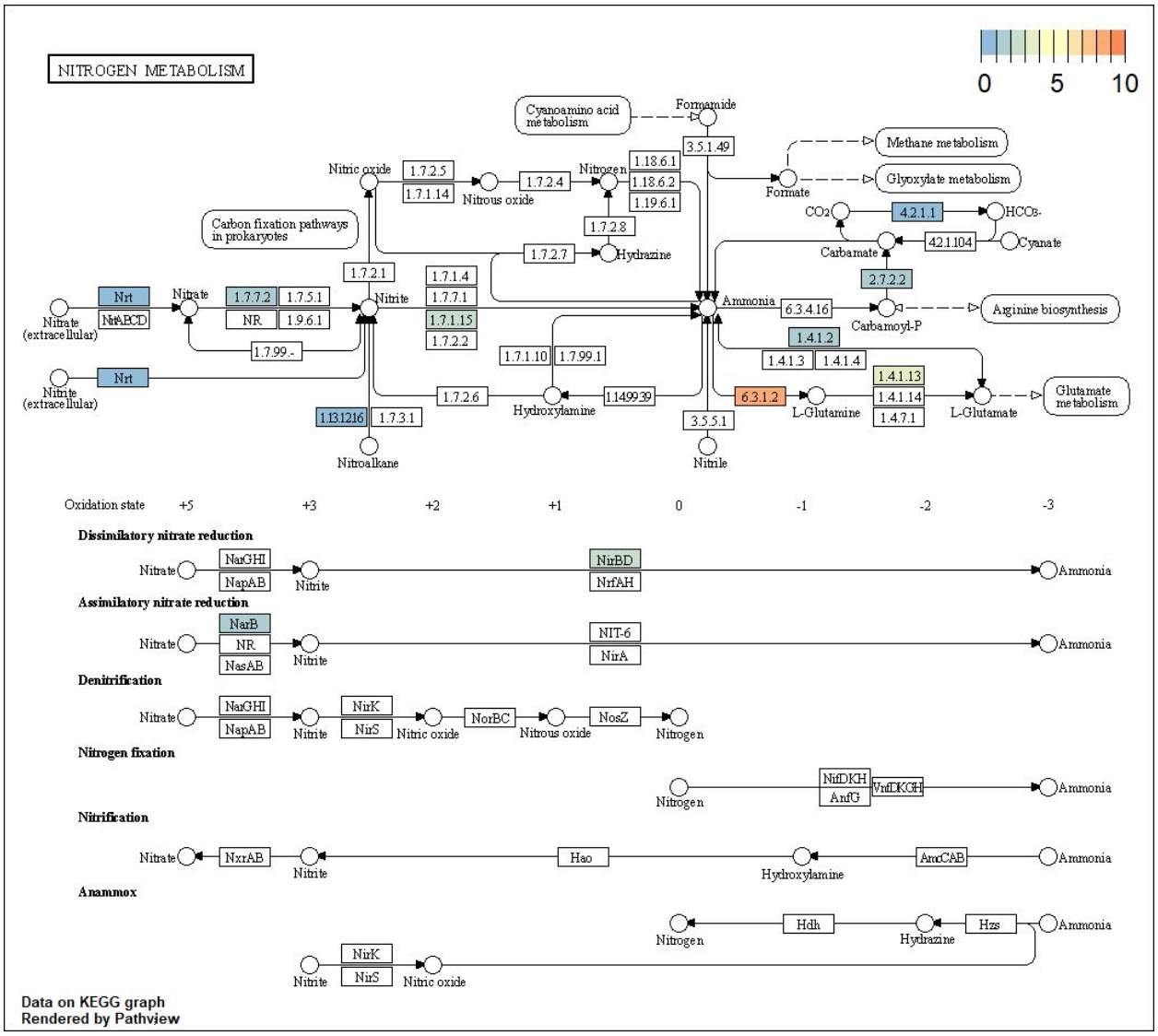

Figure 25 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Rhodospirillales* in the checkM results.

Figure 26: *Alphaproteobacteria*, Nitrogen Metabolism

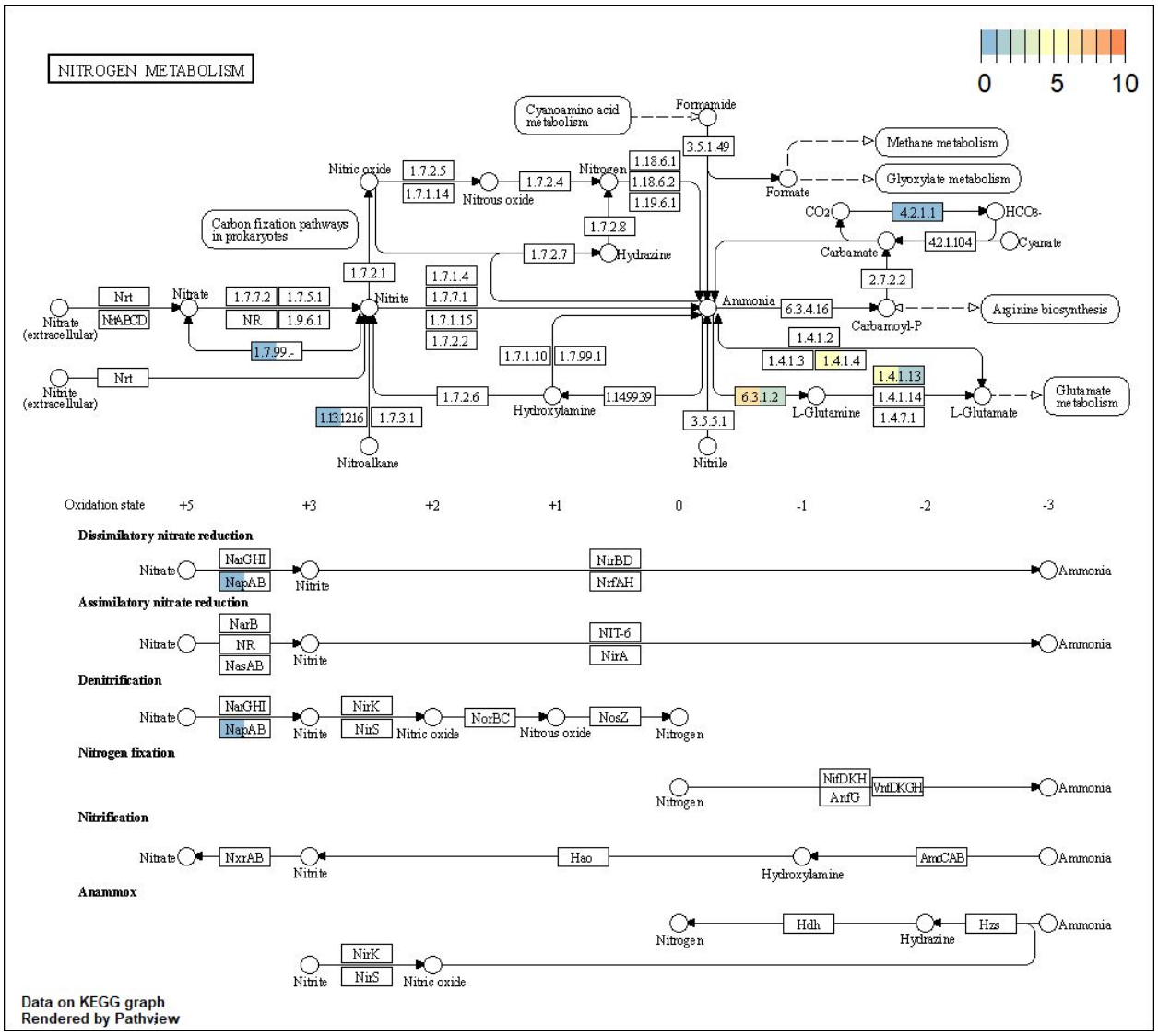

Figure 26 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Alphaproteobacteria* in the checkM results.

Figure 27: *Euryarchaeota*, Nitrogen Metabolism

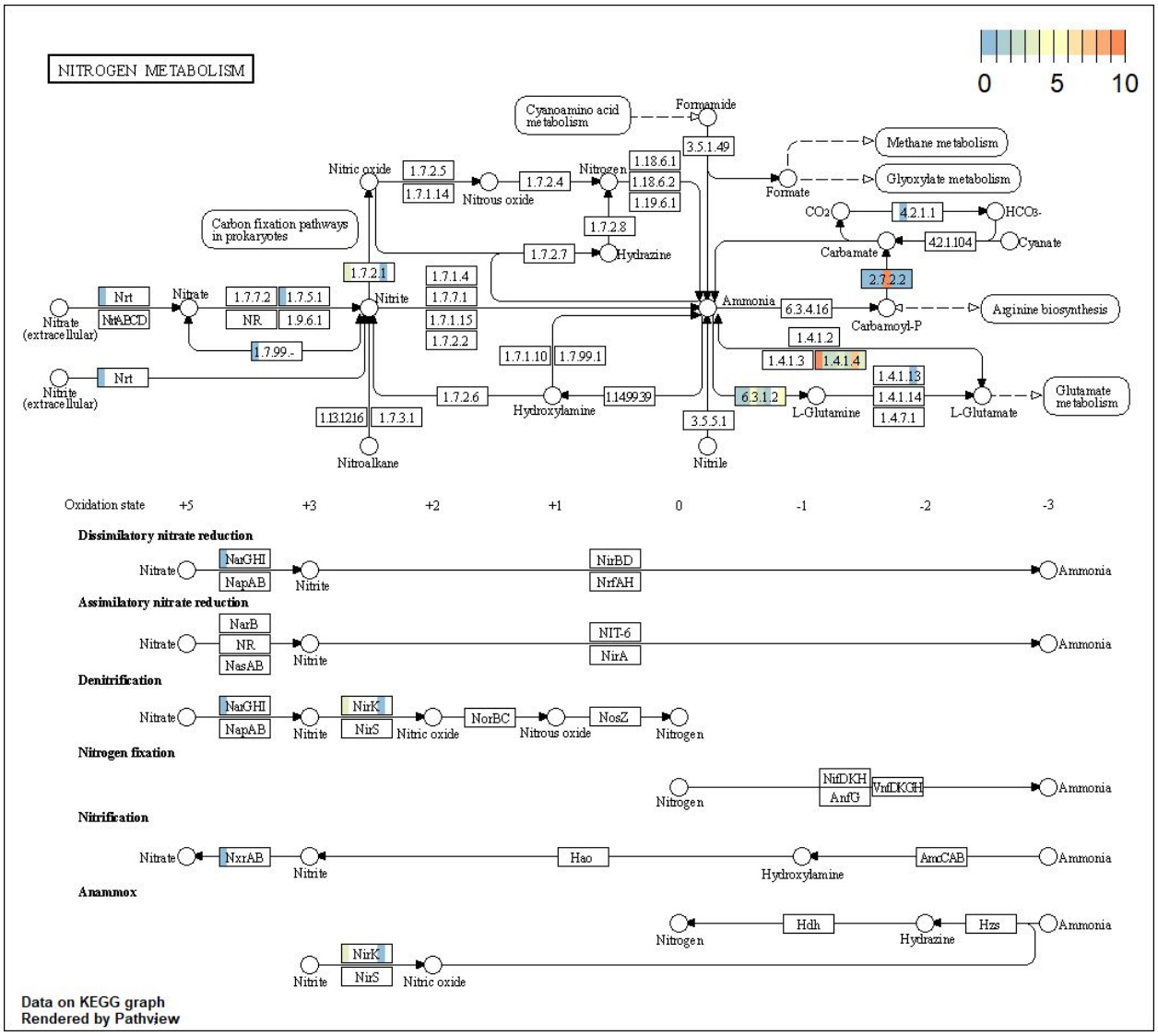

Figure 27 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Euryarchaeota* in the checkM results.

Figure 28: *Betaproteobacteria*, Nitrogen Metabolism

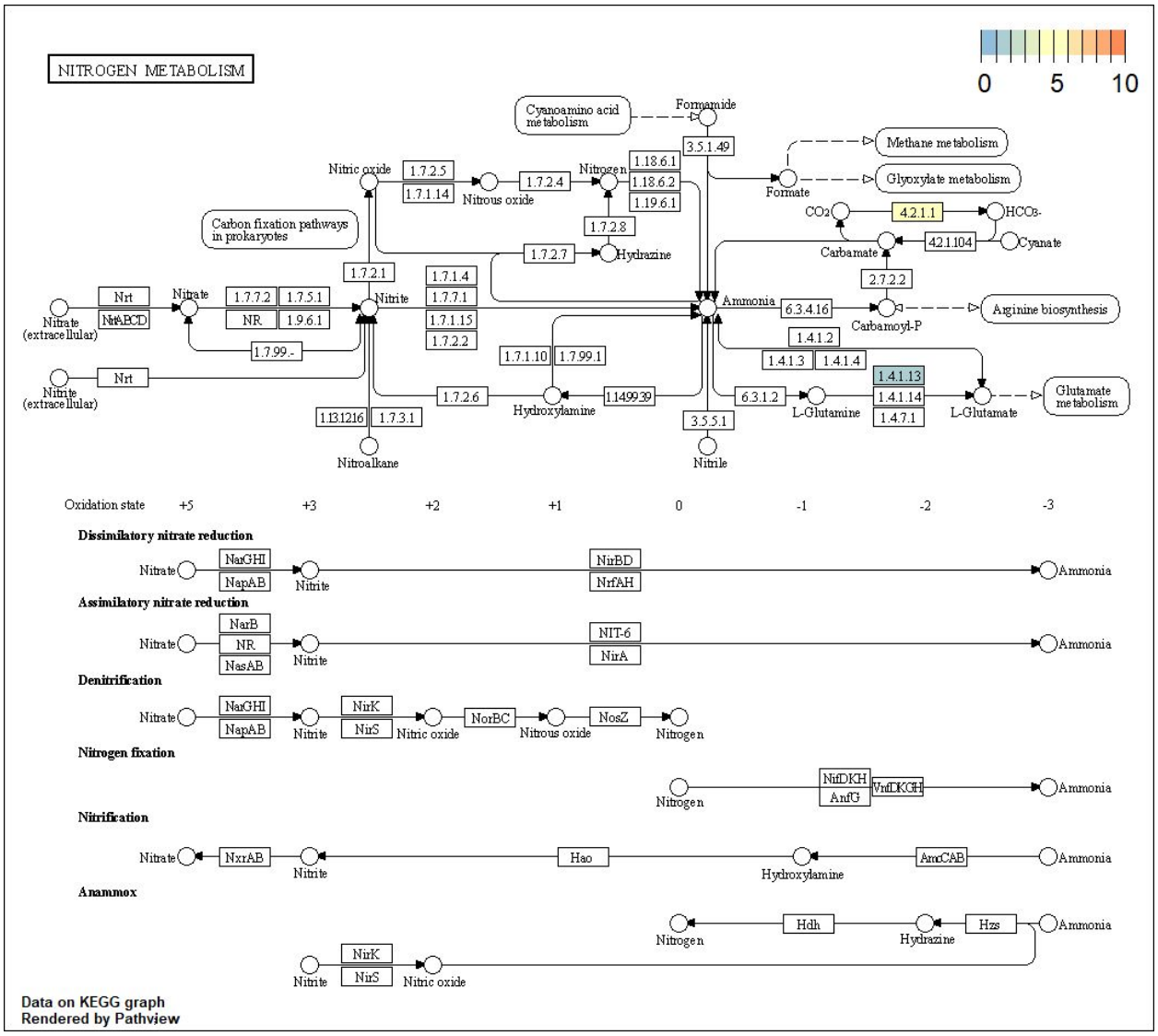

Figure 28 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Betaproteobacteria* in the checkM results.

Figure 29: *Deltaproteobacteria*, Nitrogen Metabolism

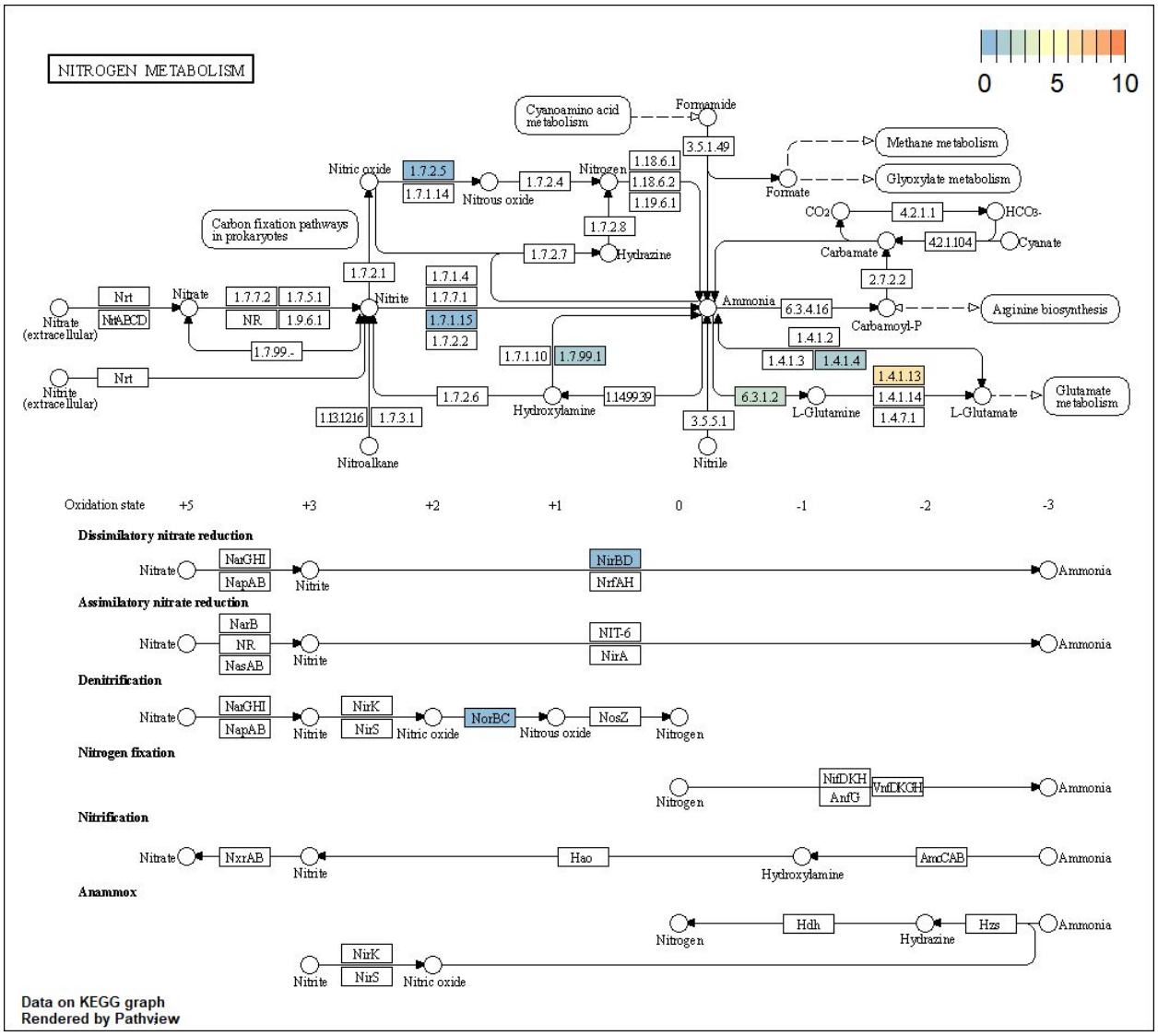

Figure 29 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Deltaproteobacteria* in the checkM results.

Figure 30: *Marinisomatia*, Nitrogen Metabolism

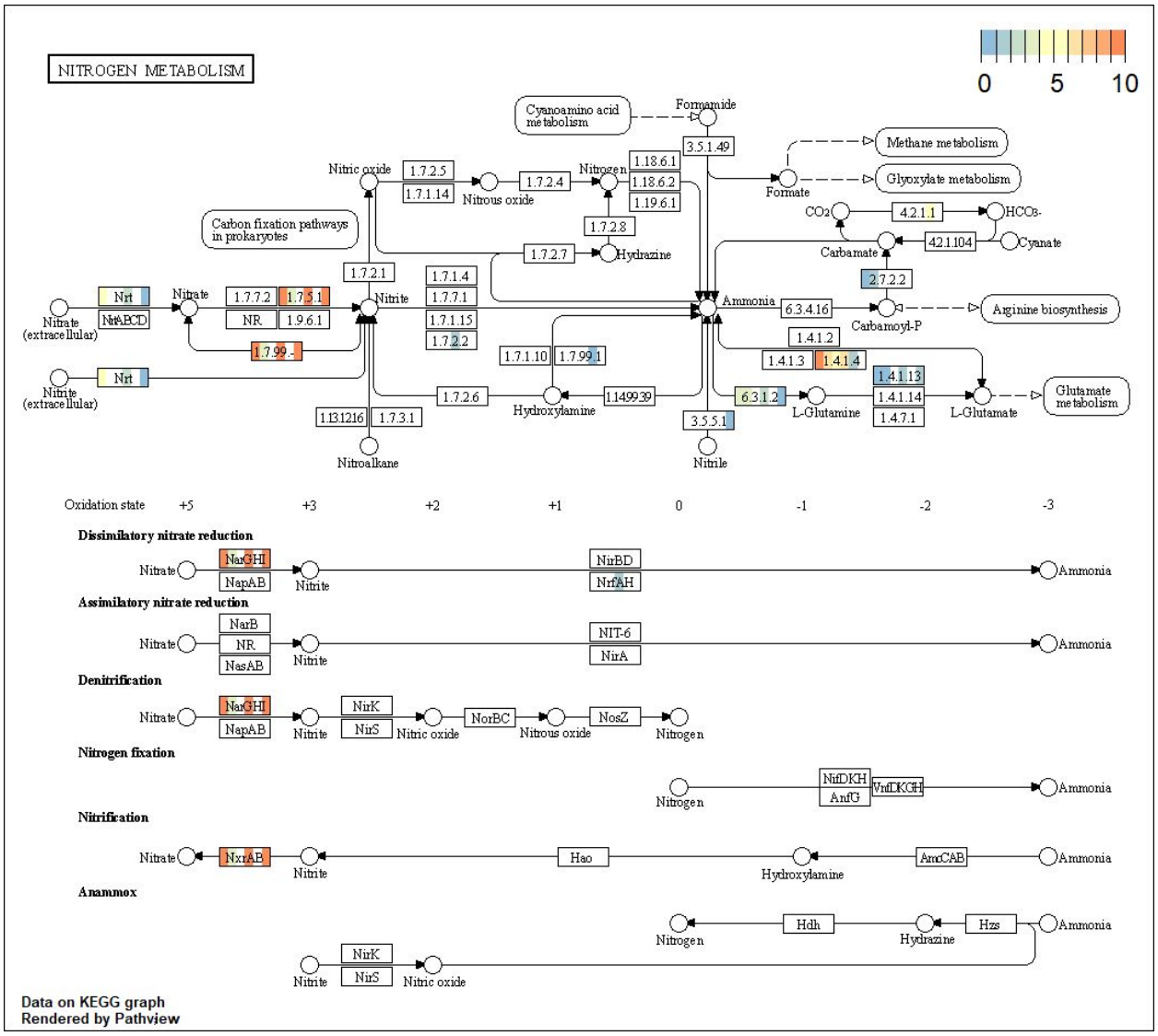

Figure 30 shows nitrogen metabolism for all high and medium quality MAGs labelled as *Marinisomatia* in the checkM results.

Figure 31: *Alphaproteobacteria*, Sulfur Metabolism

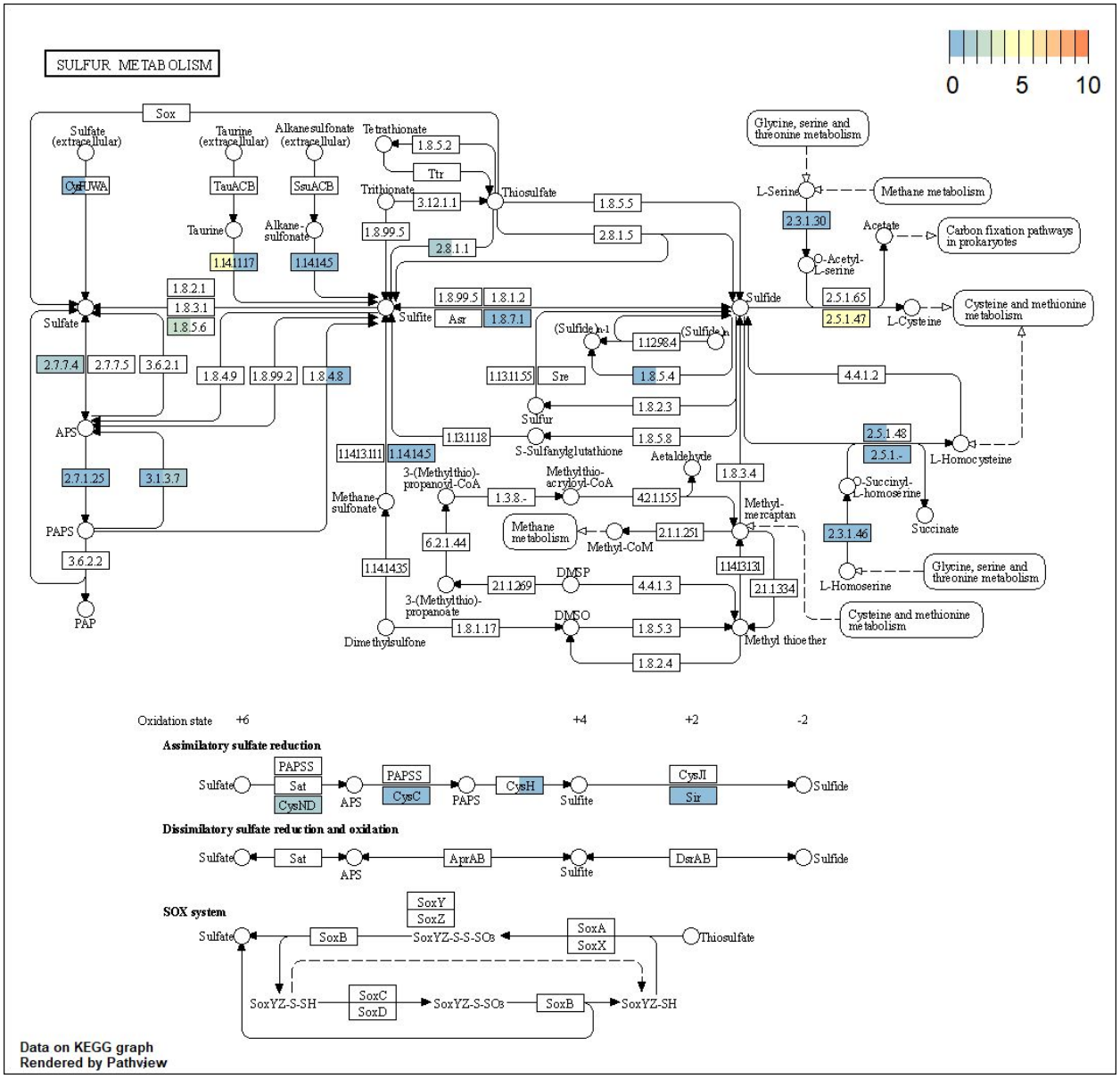

Figure 31 shows sulfur metabolism for all high and medium quality MAGs labelled as *Alphaproteobacteria* in the checkM results.

Figure 32: *Gammaproteobacteria*, Sulfur Metabolism

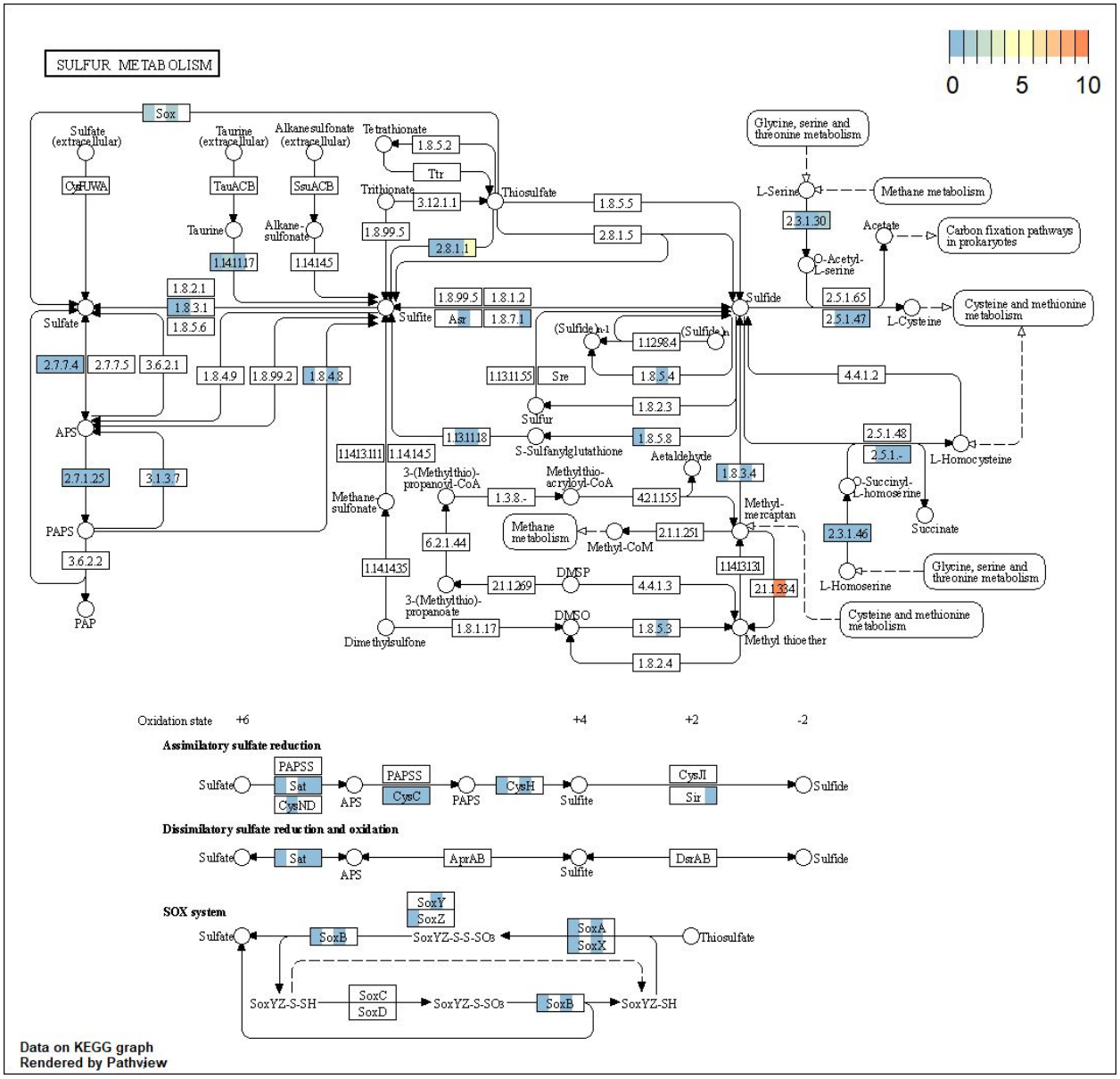

Figure 32 shows sulfur metabolism for all high and medium quality MAGs labelled as *Gammaproteobacteria* in the checkM results.

Figure 33: *Rhodobacteraceae*, Sulfur Metabolism

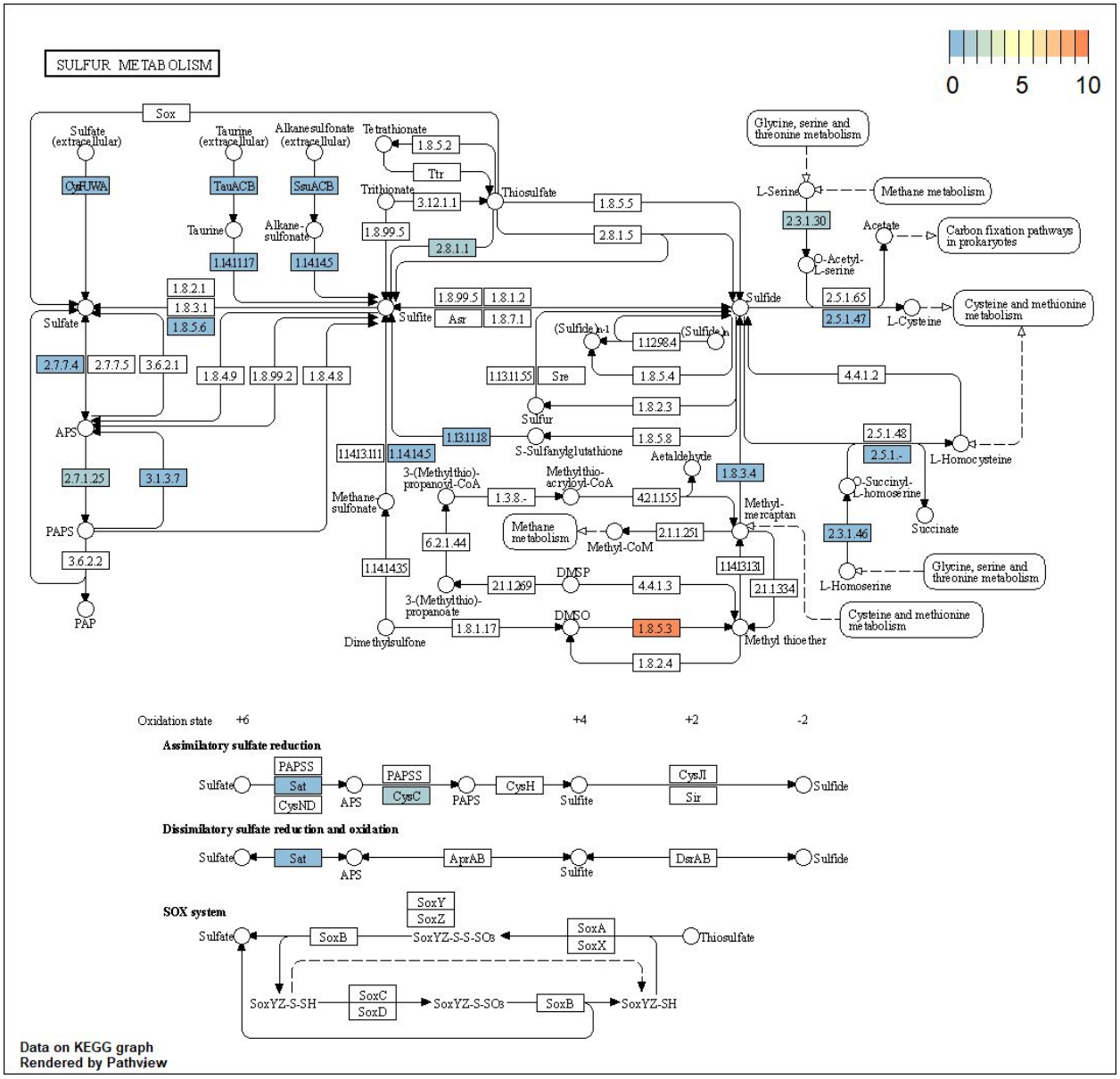

Figure 33 shows sulfur metabolism for all high and medium quality MAGs labelled as *Rhodobacteraceae* in the checkM results.

Figure 34: *Deltaproteobacteria*, Sulfur Metabolism

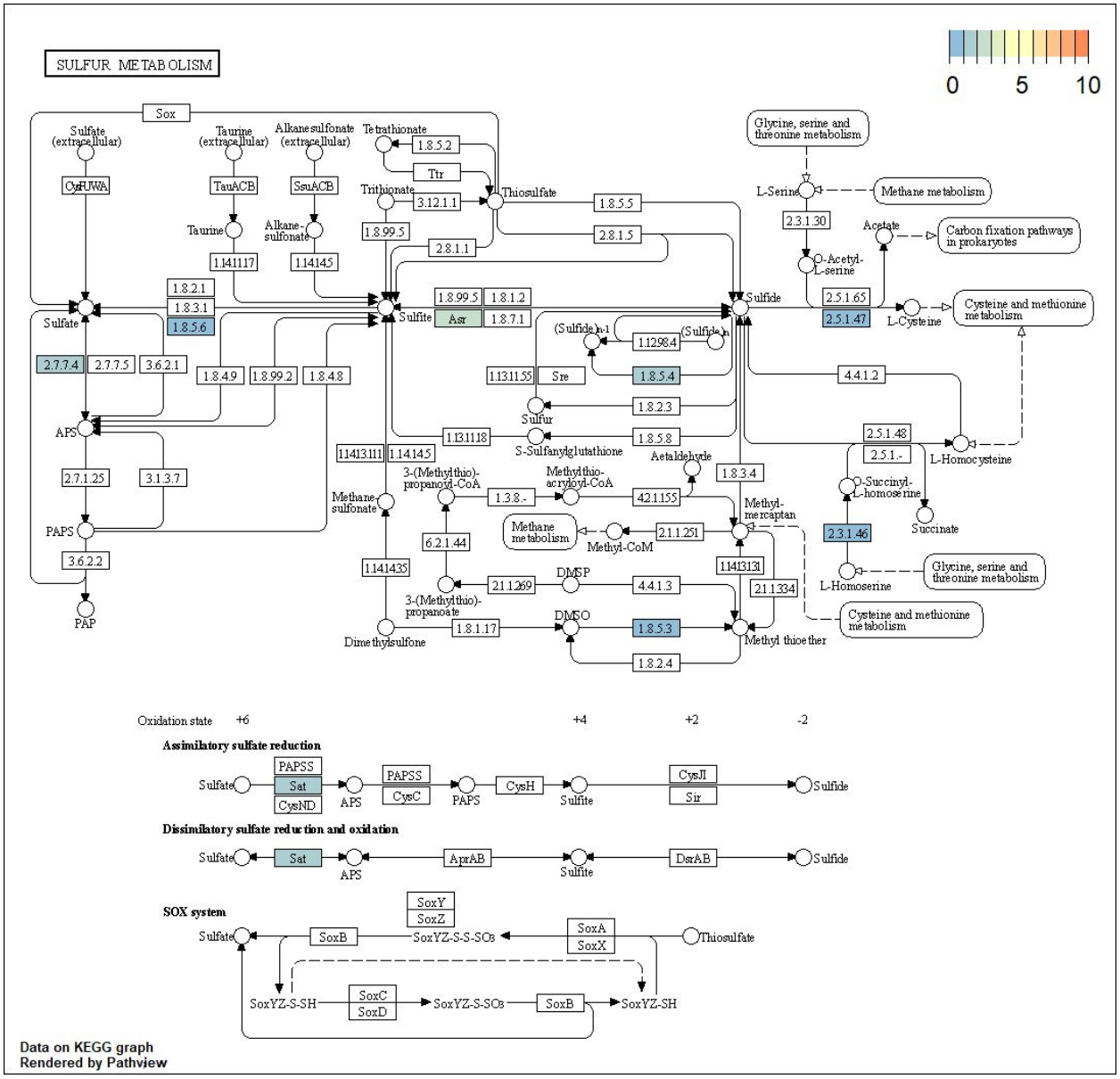

Figure 34 shows sulfur metabolism for all high and medium quality MAGs labelled as *Deltaproteobacteria* in the checkM results.

Figure 35: *Proteobacteria*, Sulfur Metabolism

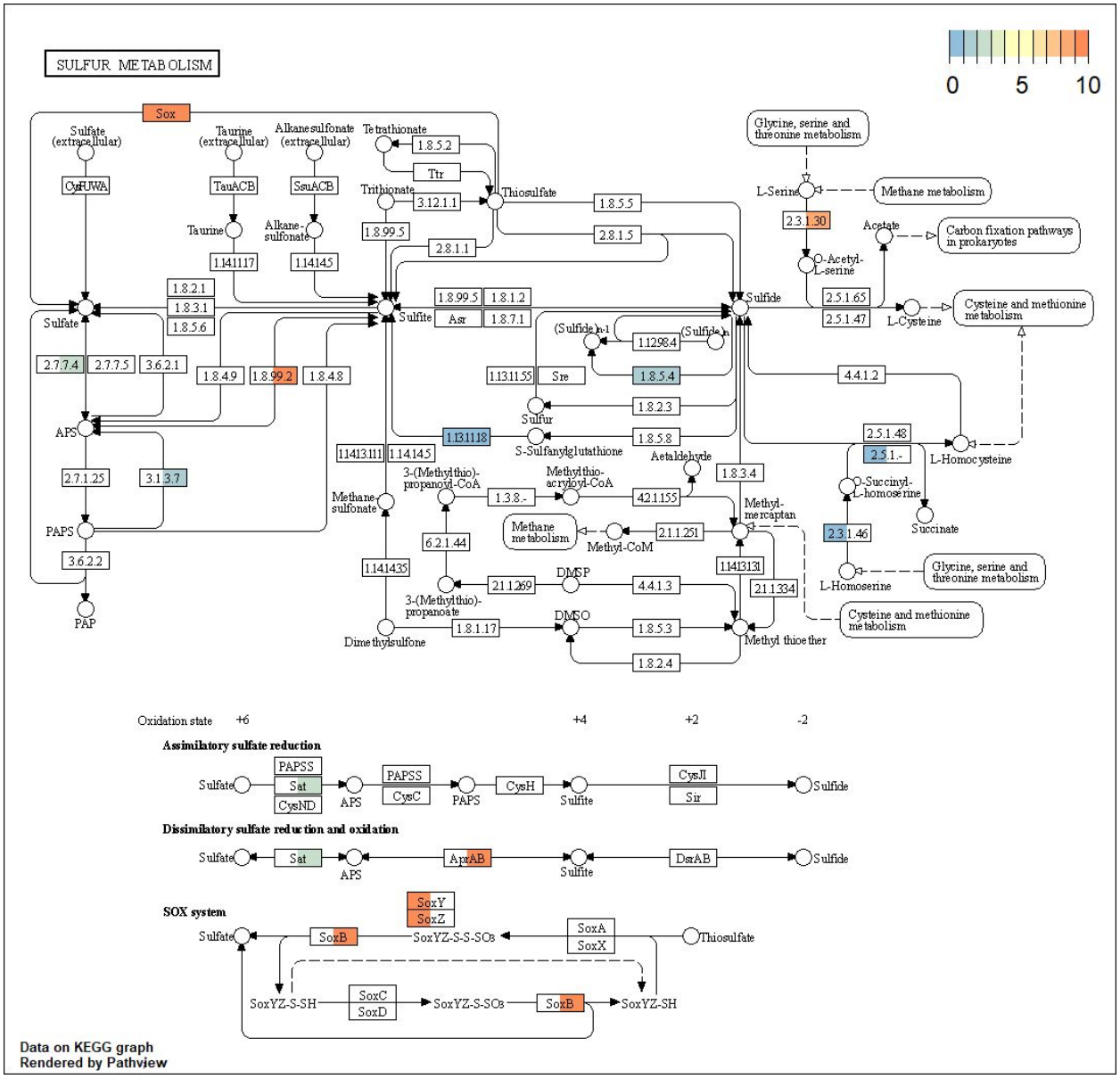

Figure 35 shows sulfur metabolism for all high and medium quality MAGs labelled as *Proteobacteria* in the checkM results.

Figure 36: *Rickettsiales*, Methane, Nitrogen, Sulfur Metabolism

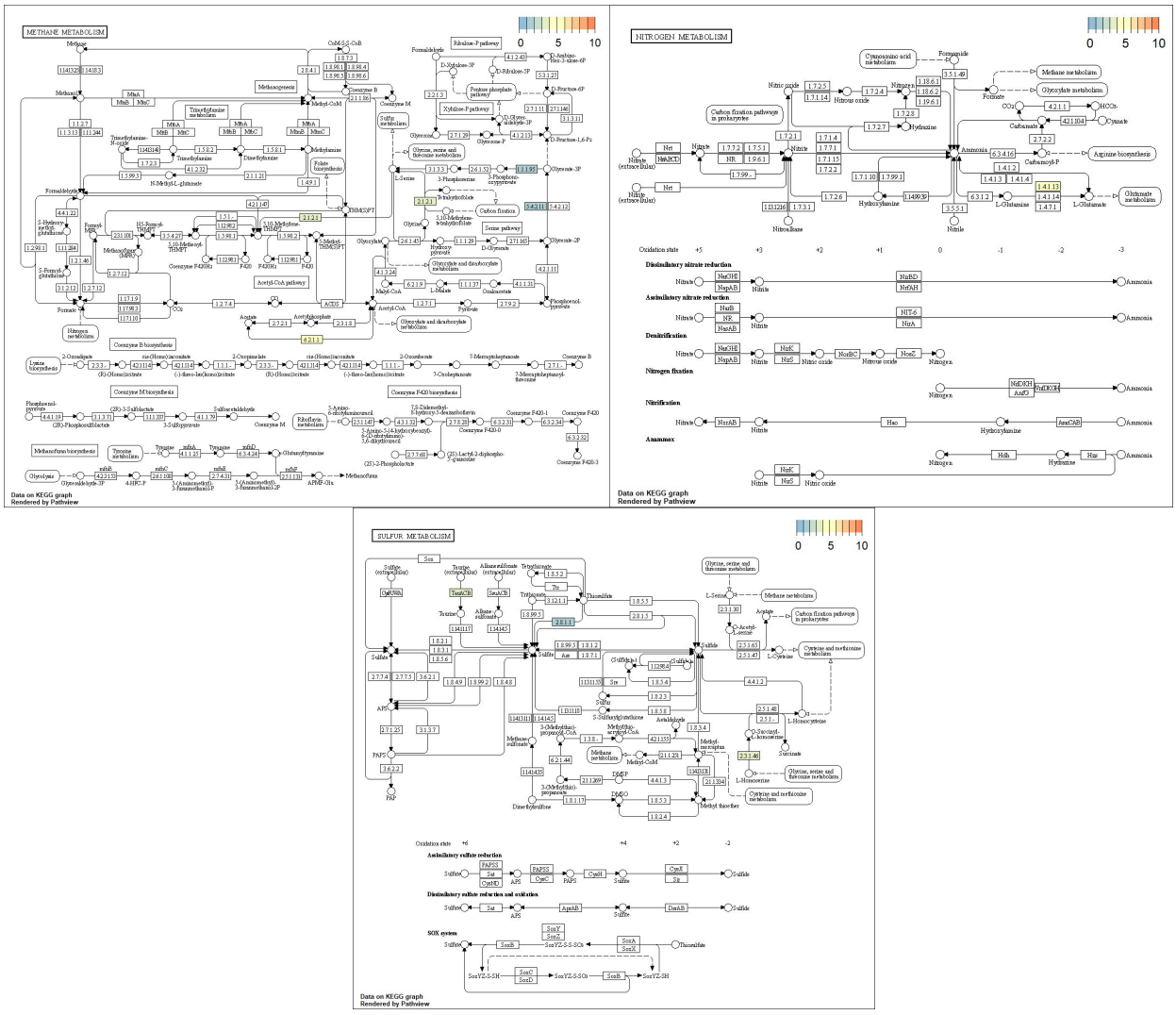

Figure 36 shows comparisons of methane, nitrogen, and sulfur metabolism for all high and medium quality MAGs labelled as *Rickettsiales* in the checkM results.

References

1. Hallam SJ, Torres-Beltrán M, Hawley AK. 2017. Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone. Scientific Data.

2. Torres-Beltrán M, Hawley AK, Capelle D, Zaikova E, Walsh DA, Mueller A, Scofield M, Payne C, Pakhomova L, Kheirandish S, Finke J, Bhatia M, Shevchuk O, Gies EA, Fairley D, Michiels C, Suttle CA, Whitney F, Crowe SA, Tortell PD, Hallam SJ. 2017. A compendium of geochemical information from the Saanich Inlet water column. Sci Data 4:170159.

3. Grantham BA, Chan F, Nielsen KJ, Fox DS, Barth JA, Huyer A, Lubchenco J, Menge BA. 2004. Upwelling-driven nearshore hypoxia signals ecosystem and oceanographic changes in the northeast Pacific. Nature 429:749–754.

4. Ellis RJ, Morgan P, Weightman AJ, Fry JC. 2003. Cultivation-dependent and -independent approaches for determining bacterial diversity in heavy-metal-contaminated soil. Appl Environ Microbiol 69:3223–3230.

5. Bowers RM, The Genome Standards Consortium, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Murat Eren A, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum

information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology.

6. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844.

7. Dubey RK, Tripathi V, Prabha R, Chaurasia R, Singh DP, Rao CS, El-Keblawy A, Abhilash PC. 2020. Single-Cell Genomics and Metagenomics for Microbial Diversity Analysis. Unravelling the Soil Microbiome.

8. Hawley AK, Torres-Beltrán M, Zaikova E, Walsh DA, Mueller A, Scofield M, Kheirandish S, Payne C, Pakhomova L, Bhatia M, Shevchuk O, Gies EA, Fairley D, Malfatti SA, Norbeck AD, Brewer HM, Pasa-Tolic L, Del Rio TG, Suttle CA, Tringe S, Hallam SJ. 2017. A compendium of multi-omic sequence information from the Saanich Inlet water column. Sci Data 4:170160.

9. Jackman SD, Mozgacheva T, Chen S, O'Huiginn B, Bailey L, Birol I, Jones SJM. 2019. ORCA: a comprehensive bioinformatics container environment for education and research. Bioinformatics 35:4448–4450.

10. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. 2011. The Sequence Read Archive. Nucleic Acids Research.

11. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2019. Database resources of the National Center for Biotechnology Information. Nucleic

Acids Res.

12. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics.

13. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics.

14. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359.

15. Thomas T, Gilbert J, Meyer F. 2012. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2:3.

16. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

18. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. Journal of Open Source Software.

19. Sommer DD, Delcher AL, Salzberg SL, Pop M. 2007. Minimus: a fast, lightweight genome

assembler. BMC Bioinformatics 8:64.

20. Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. 2011. Next generation sequence assembly with AMOS. Curr Protoc Bioinformatics Chapter 11:Unit 11.8.

21. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

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

23. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

24. Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics.

25. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research.

26. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361.

27. Luo W, Brouwer C. 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics.