UPPSALA
UNIVERSITET

Comparative genomics of Central Arctic Ocean microbiota for observation of Alternative Carbon Fixation Pathways

Kaavya Venkateswaran

Degree project in bioinformatics, 2023
Examensarbete i bioinformatik 30 hp till masterexamen, 2023
Biology Education Centre, Uppsala University, and Dept. of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences (SLU)
Supervisor: Stefan Bertilsson

Abstract

The Central Arctic Ocean is a repository of rich and diverse biota, whose major portion is one of the most important drivers of global biogeochemical cycles, including carbon cycling. In this study, the functional potential of the microbiota to fix carbon with alternative carbon fixation pathways were investigated along with their chemolithotrophic characteristics. Samples from two expeditions MOSAiC & SAS-Oden (2019-2021) resulted in metagenomic data consisting of about 1200 mOTUs (metagenomic Operational Taxonomic Units). Kofamscan based annotation explained by KEGG pathways database was analysed to explore prevalence of the alternative Carbon Fixation pathways across different taxa. From the six carbon fixation pathways, three were consolidated for their presence (rTCA, DC-HP, HP-HP). In order to explain the other metabolic processes that these organisms employ to survive, a functional annotation tool for metabolic pathways was used. that Reductive Tricarboxylic acid cycle pathway was found to be most present and observed in 5 out of 6 mOTUs selected from filtering the dataset. The taxa include bacterial phyla Proteobacteria, Actinobacteria, Chloroflexota and Marinisomatota and archaeal phyla Thermoplasmota. However, for the other pathways and less studied organisms less resolution were observed across the dataset for the presence of other pathways. These CFPs found were also supported by oxidation of inorganic compounds with high redox potentials. This study provides a glimpse of the metabolic potential of the Central Arctic Ocean microbiota, shines light on the importance of understanding and unravelling the intricacies of this rich and diverse environment.

What we do in the shadows? An Arctic Ocean story.

Popular science summary

Kaavya Venkateswaran

The Central Arctic Ocean is the home for huge icebergs, polar bears as well as a very interesting array of microorganisms. The major microbial components, bacteria, and archaea, strive to live in conditions of prolonged darkness, freezing temperatures and low energy availability under layers of ice. There had been previous expeditions in this unique ecosystem to understand them better. Recent expedition to the Central Arctic Ocean resulted in a wealth of metagenomic data that was used in the analysis.

Ocean microbes excel at assimilating and fixing carbon from its surroundings. While most organisms do this in the presence of sunlight, it is interesting to see what the Arctic microorganisms do in the shadows. Another special characteristic of these organism is that they make use of unique mechanisms to acquire energy to perform other metabolic processes to live in the extreme conditions.

This study investigated the data through the lens of the six carbon fixation pathways that were previously described in other microorganisms. In addition, some energy sourcing reactions were also looked for to explain how these microbes thrive. The study then delved into some members of the microbial community, since they were represented better than the others, for the different pathways and logic of dark carbon fixation in these microorganisms was explained.

This exploration provided insights into the complex adaptations and interactions within the Central Arctic ecosystem. Understanding these ecosystems is crucial for addressing the effects of global warming on its biodiversity, as well as leveraging their mechanisms to fix atmospheric carbon and in implementing the unique mechanisms for biotechnological advancements.

Degree project in bioinformatics, 2023

Biology Education Centre (UU) and Department of Aquatic Sciences and Assessment (SLU)

Supervisor: Stefan Bertilsson

Table of Contents

Abbreviations

3-HP - 3- Hydroxypropionate cycle

CAO - Central Arctic Ocean

CO_2 - Carbon dioxide

CFP - Carbon Fixation Pathway

CPS - Cellular Polymeric Substances

DC-HB - Dicarboxylate-hydroxybutyrate cycle

ETC - Electron Transport Chain

HMM – Hidden Markov Model

HP-HB - Hydroxypropionate-hydroxybutylate cycle

JSON – JavaScript Object Notation

KEGG – Kyoto Encyclopedia of Genes and Genomes

KO - Kegg Orthology terms

MAG - Metagenomic Assembled Genomes

MOSAiC - Multidisciplinary drifting Observatory for the Study of Arctic Climate

mOTU - metagenomic Operational Taxonomic Units

NADH/NADPH - Nicotinamide adenine dinucleotide/Nicotinamide adenine dinucleotide phosphate

NCBI - National centre for BIotechnology Information

rTCA - reductive Tricarboxylic acid cycle

sq.km - Square Kilometer

1 Introduction

A microbiome is the assemblage of all microbial organisms that live in an ecosystem. It is a characteristic component of any ecosystem, including atmosphere, water (fresh and marine), soil, points of interaction with living organisms, that centrically drive the physicochemical and biogeochemical cycles of the earth. The components of the said microbiota include bacteria, archaea, fungi and viruses which constitute the most diverse, abundant and metabolically rich consortia of organisms.(Azam *et al.* 1983) They tend to strive to live in extreme conditions found on the earth. The Ocean ecosystem is one of the most important drivers of critical biogeochemical cycles and is abundant in organisms performing a variety of metabolic processes. A major driving factor of the ocean biome, crucial for the functioning of the food web, hence sustaining its entirety, is its microbiome. The ocean microbiome constitutes approximately 70% of the marine biomass and is a major component of earth's primary contribution, pertaining to approximately 50% by performing nutrient cycling across the system. (Krabberød *et al.* 2022)

1.1 The Central Arctic Ocean

The polar regions covering 3.9 - 4.3 % of earth's surface, including the Arctic and Antarctic, have significance in their contribution to the biogeochemical cycles. The region in focus in this study, the Central Arctic Ocean (CAO) (refer to Fig1) that has an area of 3.3 million sq.km and is enclosed by continents on all sides. The Arctic marine region is characterised by four vertical layers of water or water masses. The first is the surface layer that has low salinity since it is in proximity to the polar ice influenced by the melt off. The second layer is the gradient layer, which has higher salinity than the surface layer. This layer has an influence of water from the Atlantic. The intermediate layer is warmer than the above said layers with increased salinity along with Atlantic water. The last layer is the deep water, which has conditions of high salinity and oxia (Jang *et al.* 2023, p. 2)

The important characteristic of this cold polar region is seasonal shifts in light conditions, cold temperatures with extreme winters and sea shelves around a deep central ocean basin that results in its own unique ecosystem. Within this icy environment, microbial communities flourish, forming diverse assemblages with specific adaptations and ecological dynamics and have evolutionarily adapted to thrive in conditions of low temperatures. These cold adapted organisms are called psychrophilic or psychrotolerant organisms. The CAO microbiota is one of the most important components of the nutritional cycles of the world. Its organisation and composition is influenced by factors such as oceanic currents, salinity, temperature, nutrient availability and seasonal light changes that influence the community composition. This ecosystem is also characterised by the huge influx of organic carbon from events of fresh water that drain the continents that surround it. This results in parallel existence of autotrophy, heterotrophy as well as chemolithotrophy to utilize this influx of nutrients and other sources. (Dang & Chen 2017)

Unlike the mammals and other multicellular organisms in this ecosystem, the planktons and microbes are not well studied for the impact that they experience due to anthropogenic activities and global warming. The vulnerability of the Arctic Ocean to rising atmospheric CO_2 levels is amplified by the strong freshwater influence, primarily because of its significant riverine inputs. This freshwater has limited buffering capacity, which along with its frigid Arctic temperatures, facilitates greater solubility of $CO_2$2 in the water. This alters the pH balance of these ecosystems with major effects on the structure of the biome such as alterations in food web and change in biodiversity. (Le Moullec & Bender 2021)

Projections indicate that the Arctic Ocean's surface waters will witness the most substantial decrease in pH globally throughout this century, with an estimated decline of approximately 0.45 units (Qi *et al.* 2022). This calls for the need of better and deeper understanding of how these organisms exist and sustain as well as how the rapidly changing environmental conditions affect them.

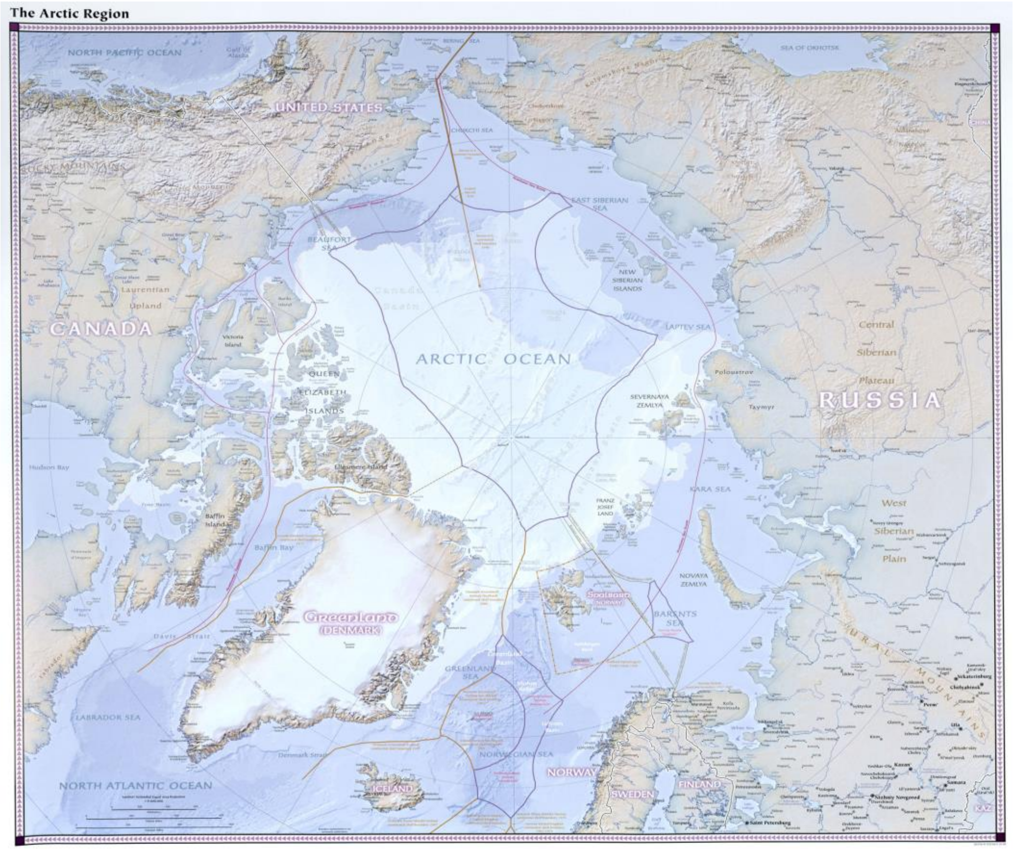


Fig 1. Map of the CAO from https://images.nationalgeographic.org/image/upload/v1638890479/EducationHub/photos/arctic-circle.jpg

1.2 The biogeochemical cycles

The microbes are integral components of the biogeochemical cycles, influencing the energy flow across the biosphere starting from heat and light energy from the sun and heat from the earth's core. A major characteristic is the ability to fix inorganic compounds in the atmosphere and the ecosystem around it, including carbon, nitrogen, sulphur and so on as constitutional building blocks of life. Among the elemental cycles, the carbon cycle is the focus of the study. Oceans currently contribute to the absorption of 30 - 50% of the world's atmospheric carbon and its net fixation by deploying different metabolic pathways by the microbial ecosystem. Carbon fixation is the process of assimilation of free inorganic carbon as building blocks into biomass, as well as being involved in energy metabolism. This process is mainly done by plants on land in the presence of sunlight. It is analogous to the processes carried out by aerobic and anaerobic organisms in many microbiomes. These microorganisms obtain energy from different sources such as light or reduced inorganic compounds. Even with much less quantity by mass compared to the autotrophs on land, the ocean microbiota has the ability to have much higher biomass turnover than their terrestrial counterparts which makes them important drivers of the nutrient cycles. (Hügler & Sievert 2011) (for summary of Ocean biogeochemical cycles, refer to Fig 2)

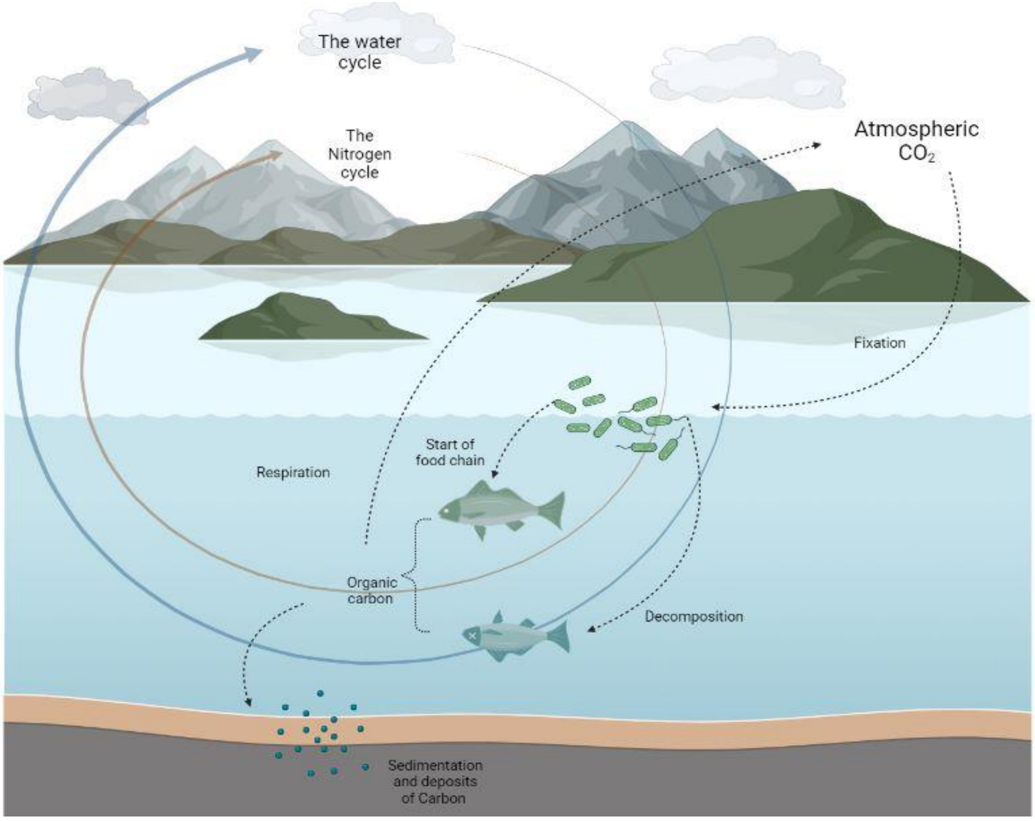


Fig 2. Carbon cycle in land and ocean environments. Microbes act as the primary producers, perform nutrient cycling, carbon sequestration, maintain symbiotic relationships with other members of the ecosystem, disease regulation and decomposition and recycling of dead organic matter.

1.3 Different nutrition modes

The organisms that are the key drivers of these nutrient cycles accomplish them through the processes of acquiring, processing, and releasing the different nutrients into the ecosystem around them. The major types of such nutrient modes are autotrophy and heterotrophy. Autotrophy is the process by which an organism makes complex organic compounds from simpler compounds while a heterotroph depends on the autotrophs to provide organic substances for their nutritional needs. Autotrophs harvest energy from several sources. Apart from phototrophy, which is harvesting energy from light, there are other mechanisms, wherein the organisms derive energy from inorganic molecules(lithotrophs) that are electron donors such as hydrogen, sulphur, nitrogen and iron and organic molecules (organotrophs). There is another category of organisms that utilise several such metabolic processes to harvest energy, called the mixotrophs.

Chemolithotrophs, specifically Chemolithoautotrophs, are predominantly autotrophic microorganisms that utilize atmospheric carbon dioxide as the primary source to synthesize essential organic compounds. These organisms rely on ATP and reducing power (NADH/NADPH) to convert oxidized carbon dioxide into highly reduced organic compounds, such as glucose. However, when a chemolithoautotroph employs an electron donor with a higher redox potential than NAD+/NADP, it necessitates the utilization of reverse electron flow to redirect electrons up the electron tower. This process is beneficial, yet it consumes energy from the proton motive force to drive electrons in a reverse direction through the electron transport chain (ETC). These organisms hence rely on other metabolic processes of utilising inorganic molecules as electron donors, especially molecules with high redox potential. Some of these compounds include hydrogen sulfide, hydrogen gas, iron, ammonia, sulfur compounds and so on.

Alternatively, some microbes are chemolithoheterotrophs, satisfy their energy and electron requirements from inorganic chemicals rely on organic compounds in their environment as a carbon source. These organisms are also known as mixotrophs since they necessitate both inorganic and chemical compounds for growth and reproduction. Chemolithoautotrophs are also mixotrophs, assimilating carbon and producing organic compounds for the other heterotrophs to feed on, which makes up the important characteristic of most of the population of ocean microbiota.

1.4 The Carbon Fixation Pathways

A major phenomenon in the carbon cycle is the process of assimilating and fixing carbon for sourcing energy, to make the basic functioning components of the organism as well as special adaptation, especially for the psychrophilic & psychrotolerant organisms to produce cellular polymeric substances (CPS). There are six different carbon fixation pathways that have been identified in living organisms across the tree of life. The photoautotrophic organisms majorly harness the sun for energy and fix carbon using the Calvin-Benson-Bassham (CBB) cycle (Bassham *et al.* 1950) among other special pathways. The other pathways include the reductive Tricarboxylic Acid (Arnon-Buchanon) pathway (Wächtershäuser 1990, Berg 2011), reductive acetyl-CoA (also called Wood - Ljungdhal) pathway

(Ljungdhal 1986), 3-Hydroxypropionate Bicycle(Fuchs–Holo Pathway)(Holo 1989), 3-Hydroxypropionate /4-Hydroxybutyrate Cycle(Berg 2011), Dicarboxylate/4-Hydroxybutyrate Cycle (Huber *et al.* 2008) and the reductive glycine pathway (Sánchez-Andrea *et al.* 2020). These carbon fixation pathways exist in many diverse taxa of bacteria and archaea found in the ocean wherein they contribute accordingly in areas of abundance and energy availability. (for details, see Section 2)

1.5 Aim and scopes

The project aims to investigate alternative modes of carbon fixation (CFP) in the Central Arctic Ocean. The goal is shedding light on the special and varied mechanisms employed by microbes to fix inorganic carbon in dark and low-energy conditions. Additionally, in order to explain use energy from the surroundings to perform energetically costly autotrophy, other pathways indicative of chemolithotrophy or mixotrophy will also be explored.

In addition, this study will compare the distribution of the microbes in relation to the surfaces or regions of varying water depths in which the different CFPs occurIn addition, this study will compare the distribution of the microbes in relation to the surfaces or regions of varying water depths in which the different CFPs occur. Pangenomic tools (Anvi'o) will be utilized to visualize such relationships, providing a comprehensive overview of microbial ecology in the CAO.

Through these comprehensive analyses, our project aims to unravel carbon fixation modes, microbial diversity, and ecological dynamics in the Central Arctic Ocean. This knowledge will contribute to advancing our understanding of microbial ecology, adaptation, and biogeochemical cycling in extreme environments. Insights from this research can improve understanding how Arctic microbial communities are affected by anthropogenic activities that distort their natural ecosystem, and application in the fields of biotechnology and bioengineering, such as carbon capture and storage strategies or the development of efficient microbial-based bioprocesses.

2 Literature survey

At present, there are six known CO_2 fixation pathways that operate in microorganisms. These organisms have been systematically studied in several taxa living in different conditions across the globe. The commonly found pathway in autotrophs is the Reductive Pentose Phosphate (Calvin-Benson-Bassham) Cycle (CBB) that is performed by all autotrophic plants on the terrestrial ecosystem as well as many aquatic photosynthetic organisms. There is also a seventh pathway that has been found recently that requires further research which as it seems to be prevalent in organisms that perform dark carbon fixation. The major carbon fixation pathways in prokaryotes are as follows:

2.1 Reductive Tricarboxylic Acid Cycle (rTCA) or Arnon-Buchanan Cycle

The rTCA cycle is a carbon fixation pathway utilized by certain bacteria and archaea. It can use different carbon sources, including carbon dioxide (CO_2), acetate, or pyruvate. The specific carbon sources vary among different groups of microorganisms. The enzymes involved in this pathway include malate dehydrogenase, fumarate hydratase, succinyl-CoA synthetase, isocitrate dehydrogenase, aconitate hydratase, fumarate reductase, 2-oxoglutarate synthase, and citrate cleaving enzymes which are the key identifying enzyme for this pathway. The rTCA cycle is found in anaerobic or microaerophilic bacteria and archaea such as Chlorobiales, Aquificales, Epsilonproteobacteria, Nitrospirae, and some proteobacterial groups.(Wächtershäuser 1990, Berg 2011, Garritano *et al.* 2022)

2.2 Reductive Acetyl-CoA (Wood-Ljungdahl) Pathway

The Wood-Ljungdahl pathway is another carbon fixation pathway primarily used by bacteria and archaea. It mainly utilizes carbon dioxide (CO_2) as the carbon source, but some microorganisms can also assimilate other C1 compounds such as carbon monoxide (CO), formaldehyde, methanol, methylamine, and methylmercaptane. The pathway involves the conversion of CO_2 into acetyl-CoA, which serves as a building block for organic compounds. Key enzymes in this pathway include CO dehydrogenase, acetyl-CoA synthase, and enzymes involved in the conversion of acetate to methane. Bacteria such as Firmicutes, Planctomycetes, Deltaproteobacteria, Spirochaeta, and archaea from the Euryarchaeota phylum utilize this pathway.(Ljungdhal 1986, Wood 1991, Garritano *et al.* 2022)

2.3 3-Hydroxypropionate Bicycle (Fuchs-Holo Pathway)

The 3-Hydroxypropionate Bicycle pathway is a carbon fixation pathway found in certain microorganisms, including green nonsulfur bacteria Chloroflexus, Metallosphaera, Sulfolobus, Archaeoglobus, and Cenarchaeum (a symbiont of marine animals). It utilizes bicarbonate (HCO_3-) or carbon dioxide (CO_2) as the carbon source. The pathway involves the fixation of bicarbonate to form glyoxylate and the disproportionation of glyoxylate and propionyl-CoA to pyruvate and acetyl-CoA.

Key enzymes in this pathway include 4-hydroxybutyryl-CoA dehydratase, malonyl-CoA reductase, and propionyl-CoA synthase. (Holo 1989, Garritano *et al.* 2022)

2.4 3-Hydroxypropionate/4-Hydroxybutyrate Cycle

The 3-Hydroxypropionate/4-Hydroxybutyrate cycle is a carbon fixation pathway utilized by autotrophic thermoacidophilic microorganisms from the Sulfolobales order. It utilizes carbon dioxide (CO_2) as the carbon source and involves a series of enzymatic reactions to convert CO_2 into acetyl-CoA. The pathway consists of two parts (P1 and P2) and involves enzymes such as acetyl-CoA to succinyl-CoA and 4-hydroxybutyryl-CoA dehydratase. Autotrophic thermoacidophilic microorganisms from the Sulfolobales order utilize this pathway. (Huber *et al.* 2008, Garritano *et al.* 2022)

2.5 Dicarboxylate/4-Hydroxybutyrate Cycle

The Dicarboxylate/4-Hydroxybutyrate cycle is a carbon fixation pathway observed in the thermophilic crenarchaeon Ignicoccus and some Crenarchaeota. It utilizes organic dicarboxylates (e.g., malate or succinate) as carbon sources. The pathway incorporates enzymes from the reductive tricarboxylic acid cycle (rTCA) and the 4-hydroxybutyrate portion of the 3-hydroxypropionate/4-hydroxybutyrate cycle. Along with the utilization of carbon dioxide, the dicarboxylates are metabolized within the pathway. (Huber *et al.* 2008, Garritano *et al.* 2022)

The important alternative carbon fixation pathways as summarized by the KEGG pathway database, that includes all reaction steps and substrates that are used in the different pathways is depicted in Fig 3. The pathway maps by the KEGG database is a curated, standard source of reference.

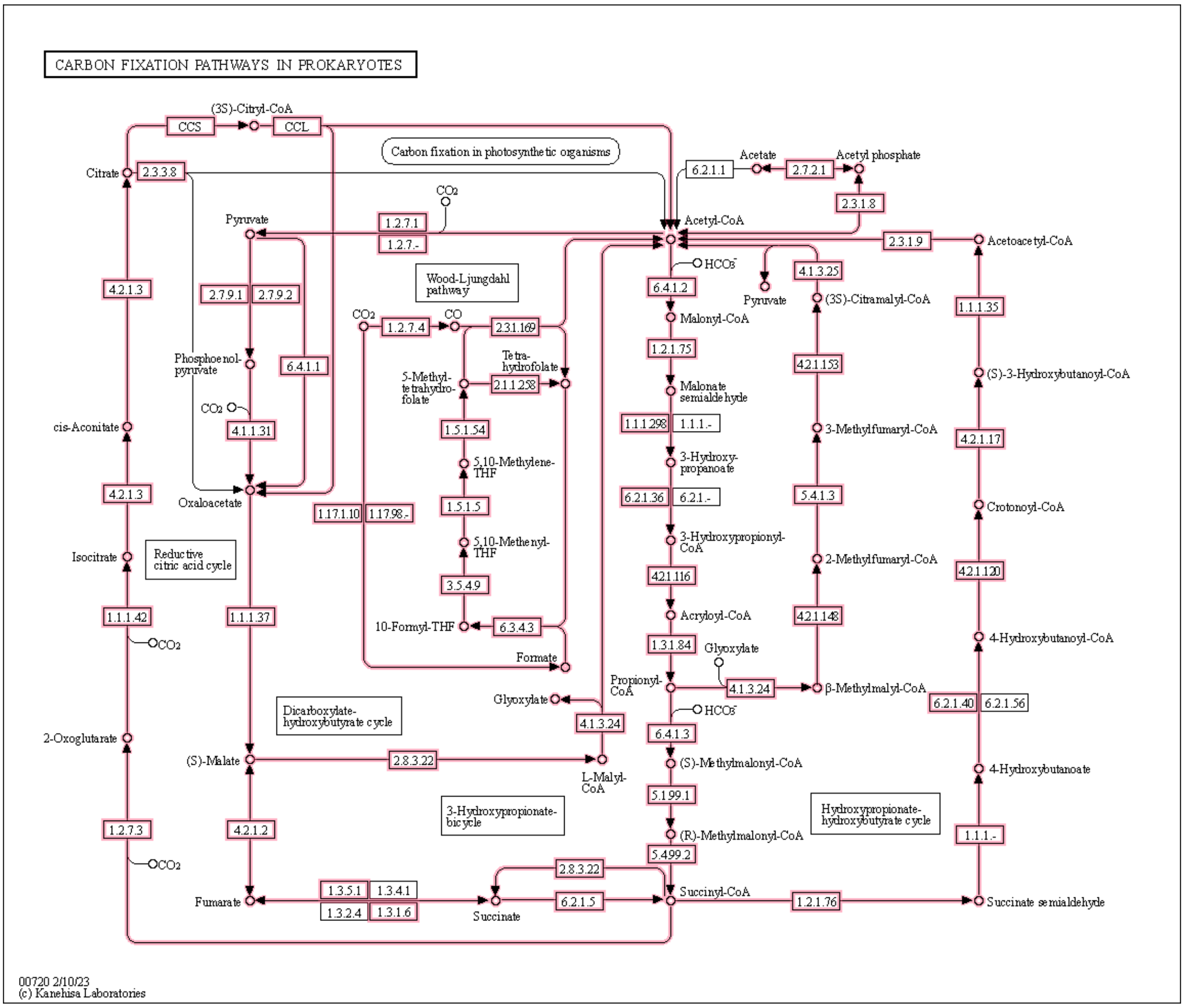


Fig 3. An overview of all alternative carbon fixation pathways used by prokaryotes. The boxes indicate the different enzyme components that perform the pathway, while the substrates are signified as circles. The arrow point the direction the map signifies the interconnections (on a molecular context) of these pathways. The five pathways under focus are highlighted in pink. Map map00720 from https://www.genome.jp/pathway/map00720.

3 Methods

3.1 Data collection, Sequencing and Assembly

Water samples were collected from the MOSAiC (Multidisciplinary drifting Observatory for the Study of Arctic Climate, Mock et al. 2022) and DAS-Oden expeditions These expeditions traversed through the CAO from the coast of northern Greenland to the polar north that encompasses the epipelagic, mesopelagic and bathypelagic regions over a period of 12 months that included collection of samples with temporal resolution of one week. This spatio-temporal recording was done, the samples were processed for metabarcoding, metagenomic and metatranscriptomic sequencing as well as for several bioassays. The metagenomic and metatranscriptomic data are in focus for this study. Samples were collected from different surfaces such as Subsea ice, maximum temperature and maximum chlorophyll. (For purposes of other studies) samples were subjected to different treatments in conditions of Light or Darkness, temperature and chlorophyll content. The samples were subjected to Illumina shotgun sequencing. The sequences were assembled with the Metasssnake pipeline (Metasssnake) which uses the Snakemake streamlining tool (Snakemake). The pipeline encompasses tools all the way from the first step of checking the quality of the sequences, to providing functional annotations for the genomes assembled (Metagenome Assembled Genomes or MAGs). The MAGs formed were then assorted into bins based on sequence identity that is assumed to be identical with phylogeny. For these bins mOTUs were created using the mOTUliser tool (Motulizer). The assembly, annotation, and creation of mOTUs is a standard analysis pipeline were previously done by the lab. The metadata table included the information about these mOTUs such as completeness and taxonomic classification and many others.

3.2 Exploring the Annotation of Metagenomes

3.2.1 Preliminary exploration with EggNOG annotation

Metagenomes previously assembled by the metasssnake2 metagenomic assembly pipeline mentioned above were annotated using EggNOG mapper (EggNOG Mapper). Gene clusters made for the bins by MMseqs2 (MMseq2), the annotation was explored for the presence of the 6 different metabolic pathways that are described by the KEGG pathways (https://www.genome.jp/pathway/map00720) database. A summary of the different carbon fixation pathways in focus, the important enzyme components and their respective identifiers are summarised below. KEGG pathways database explains 7 pathways in total. But for this study, upon preliminary data analysis to contemplate the overall distribution of different metabolic pathways, the incomplete TCA cycle and the Phosphate acetyltransferase-acetate kinase pathway are not included since their prevalence was not strongly supported.

Table 1. Summary of Carbon fixation pathways by prokaryotes as described by KEGG Pathways database. The first column is the Broad classification, Energy metabolism in prokaryotes, followed by the module identifiers, the respective pathway names and key enzymes found in literature for the pathways

	Module Identifiers	**Pathway names**	**Key enzymes**
Energy metabolism **00720** (KEGG)	M00173	Reductive citrate cycle (Arnon-Buchanan cycle)	Citrate Lyase, Citryl CoA lyase
	M00374	Dicarboxylate-hydroxybutyrate cycle	4-Hydroxybutyrl CoA Dehydrogenase, Pyruvate : ferredoxin oxidoreductase, Phosphoenol pyruvate carboxylase
	M00375	Hydroxypropionate-hydroxybutylate cycle	4-Hydroxybutyrl CoA Dehydrogenase, Acetyl CoA synthase
	M00376	3-Hydroxypropionate bi-cycle	4-Hydroxybutyrl CoA Dehydrogenase, Malonyl CoA Reductase, Acetyl CoA Carboxylase
	M00377	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)	CO dehydrogenase/acetyl CoA synthase, Acetate kinase, Methyl transferase, Phosphotransacetylase
	M00620	Incomplete reductive citrate cycle	2-oxoglutarate decarboxylase, Succinic semialdehyde dehydrogenase
	M00579	Phosphate acetyltransferase-acetate kinase pathway	-

The result of the EggNOG mapper is a JSON file that consists of several JSON objects for each of the gene clusters formed with annotated representative protein sequences. The tool uses precomputed orthologous groups and phylogenies from the EggNOG database (EggNOG) to transfer functional information from fine-grained orthologs only. These annotations were further parsed to extract the desirable candidates, namely the bins containing KEGG Orthology terms corresponding to the drivers of the pathways of interest. The important information extracted were the KOs. Member bins JSON files were parsed using the python package json. The KOs corresponding to the enzymatic components of the different carbon fixation pathways were also extracted. The KOs selected for the different bins were grouped and sorted for each pathway. For the pathways, the completeness of the bins was calculated using an available in-house function. This function takes into account the definition of the pathway, that is the chronological order of the enzymes involved, most importantly considering the alternatives to certain enzymes. Finally, the function calculates the percent completeness of the pathway for a specific member bin in relation to the described complete pathway. Heatmaps were produced using the Seaborn package and pheatmap package in R in order to cluster and visualise the presence of different KOs for the pathways in the bins. Hierarchical clustering was applied on the bins to visualise the relatedness of the bins based on their completeness. All of the analyses were carried out using a local high performance computing station (parsing big data files) and a personal computer (R studio).

3.2.2 Kofamscan Annotation

For the dataset that contains protein sequences of all the bins, the Kofamscan (Kofamscan) tool was used to annotate the proteins directly. The curated Kofam database uses HMM profiles previously described for different KEGG Orthology terms to annotate the protein sequences. This annotation was done with a threshold of 50% strictness for the dataset annotation and the number of threads to be utilised was set to 12 to parallelise the processing. Firstly, the hits marked for significance by the tool (denoted by a *) were selected for parsing and extraction of information. This resulted in a significant loss of desired KOs, hence, the gene clusters were selected for which the ratio of score to threshold was more than 50%, to ensure reliable annotation as well as conservation of the KOs that are required to spot the presence of the pathway. The Kofamscan output only has annotations for gene clusters, for which the member bin details were extracted later from the metadata table, as each member bin will constitute several gene clusters. This was then followed by calculating the percent completion and producing heatmaps to visualise the presence of the different pathways as followed for the EggNOG annotation dataset.

3.2.3 Filtering and extraction of taxa for the pathways

Further steps of filtering and consolidation were done for selected member bins that have at least 25% of the components of the pathway. Apart from pathway completeness, bins with a completeness threshold of at least 60% were applied. Further bin information, such as the mOTUs that they're present in, the mOTU completeness, and the GTDBTK classification were extracted from the metadata file. In the next steps mOTU bins with completeness zero and unbinned members were removed and then the overall dataset was filtered for percent completion that is more than 60% consecutively. The bins were then aggregated according to the mOTU affiliation, wherein the presence-absence of the KOs were also combined. The mean completeness for all the bins belonging to the mOTU along with the number of bins were noted. The KOs were calculated for their frequency of occurrence in the bins. After the first round of filtering, we focused on the KOs for the key enzymes, combining and plotting the mOTUs for these KOs. The mOTUs were filtered again, this time for their overall completeness and for the presence of the key enzymes. For the mOTUs selected, a representative member bin was selected in order to reduce the dimensionality of the dataset and to make representations easier for the downstream processes. External sources for the selected taxa, such as literature about their ecology and distribution, and other information were collected. The representative bins from the select mOTUs were also checked for their completeness.

3.3 Exploration for other metabolic pathways - Gapseq

In order to explore the other metabolic pathways of carbon fixation and explaining the functioning of the organism, the Gapseq tool(Zimmermann *et al.* 2021) was used. This tool enables the prediction of metabolic pathways and automatic reconstruction of microbial metabolic models. It predicts enzyme activity, carbon source utilization, fermentation products, and metabolic interactions of the microbes based on scientific literature and empirical data encompassing 14,931 bacterial phenotypes. Gapseq facilitates the construction of comprehensive metabolic models. This tool has two modes: 1. pathway and transporter prediction, 2. gap filling and network reconstruction. The pathway and transporter prediction explores the genomic sequence for different metabolic pathways that the microbe might harbour along with parameters such as percent completion and based on that presence/absence prediction. Then the tool reconstructs the pathway networks predicted to be present in the microbe by employing growth media parameters in order to fill gaps as well as predict genome scale metabolic models for cultured microbes. The tool was run in the recommended conda environment set up on the local workstation, using 20 threads in parallel, on genomic sequences in. fna format as input files. . All select member bins, 18 in total for the 4 metabolic pathways M00173 (rTCA), M00374 (DC-HP), M00375 (HB-HP) & M00376 (3-HP) of the representative mOTUs were processed through this tool in order to explore the other metabolic pathways. , . The reference database was the GTDBtk database, and the default reference organisms were Bacteria and Archaea. In order to perform all the predictions, the tool takes up to 4 hours (as specified in the official documentation) for each sample. Hence the samples were first processed for only predicting all metabolic pathways with the find -p all command, which takes about 2 hours per sample. Furthermore, the select bins were explored for carbon fixation and related chemolithotrophic pathways such as nitrogen fixation, sulphur oxidation, etc to confirm its presence and correlation.

3.4 Pangenome analysis with Anvi’o tool consortium

The Anvi’o tool (Eren *et al.* 2021) was used for pangenome analysis, in order to look at the overall distribution for the different CFPs in the different bins for the mOTUs selected. The mOTUs picked for this analysis are the ones that had several bins as components for the specific pathway and its component bins checked for its completeness, redundancy in order to eliminate contaminated bins. The first step is to make contigs.db for the bins of each mOTU using the anvi-gen-contigs-database. The contig databases were then processed to retrieve the HMM profiles for the gene clusters found in the contigs. The annotation using the KEGG Kofam database with the permissible threshold to have heuristic values for the threshold set to 0.5(50% threshold) was done with the programs anvi-run-hmms and anvi-run-kegg-kofams respectively. The several metabolic pathways that the bins possess were then determined by the program anvi-estimate-metabolism. This is stored as text files for each bin for the mOTU. Following the creation and annotation of the contig databases, another database for all the contigs, a genome database, was made with the anvi-gen-genomes-storage program. The next step was to generate a pangenome database. This was done by the program anvi-pan-genome. Then the probability of occurrence of a gene in the core or accessory genome was calculated by the tool anvi-script-compute-bayesian-pan-core that utilises the pangenome and the genome database.

4 Results

4.1 The Dataset

Metagenomic data from the Central Arctic Ocean was scarce until recently but the MOSAiC (Multidisciplinary drifting Observatory for the Study of Arctic Climate) expedition that took place from

September 2019 until October 2020 enabled collection of such data during a full yearly cycle. The result of this expedition is a large collection of meta barcoding, metagenomic, and meta transcriptomic data over both seasonal and spatial scales. An existing large metagenomic dataset from the MOSAiC cruise containing ~10000 Metagenomic Assembled Genomes (MAGs – referred to as bins here) that represents ~1200 mOTUs (which are MAGs that were clustered together – equivalent to species) combined with binning and assembly of MAGs from 500+ more recently sequenced metagenomics samples collected from the Central Arctic Ocean during the MOSAiC expedition and the shorter SAS- Oden expedition that took place in 2021. (Mock *et al.* 2022).

4.2 Parsing the EggNOG annotation

The output file from the EggNOG mapper contained annotation for the gene clusters of the complete set of bins. Upon parsing this file, extracting the desired KOs that represent for the pathways of interest, 9610 bins were extracted in total. These bins were assorted for each of the pathways, according to their KO composition. The percentage completion of the pathway was calculated for these bins. This resulted in a significant reduction in dimensionality wherein the bins were selected to have the desired pathways. A heatmap was produced for these bins that are at least 25 % complete for the pathways to select for strong representation of presence of different pathways (see Appendix 1). In total, 143 KOs were found to be involved in all the CFPs of interest. Out of the 99 unique KO terms (of the total 143), only 53 were detected when they were selected in such a way that these KOs occur 5% more than the next best (most occurring) KO.

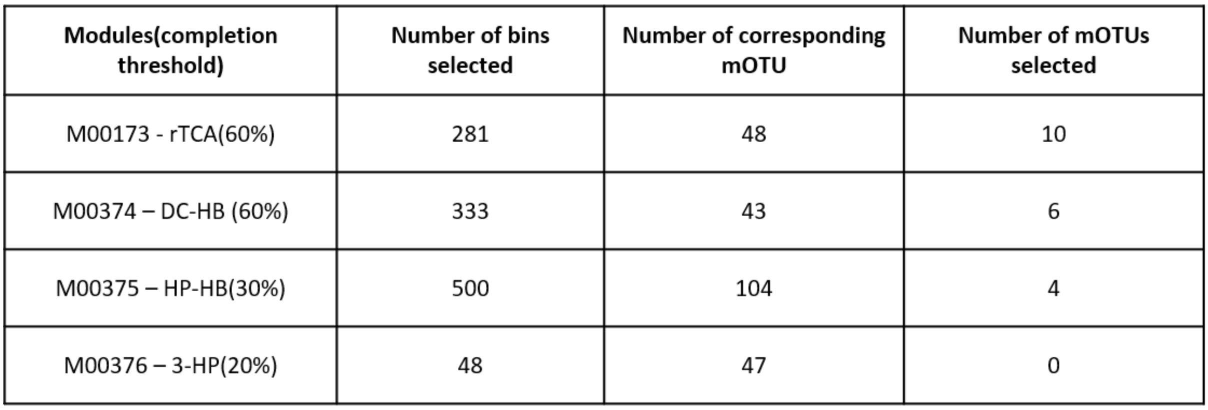
4.3 Parsing the Kofamscan annotation

The Kofamscan tool run resulted in annotation of gene clusters for the assembled bins. There were annotations for the complete gene clusters for the entire dataset of 6 542 484 gene clusters. Each gene cluster is the representation of genes present in different member bins. Since the downstream processes require processing of the member bins, the results from parsing the Kofamscan output were mapped to the member bin list available from metadata. This process of mapping was done after the filtering of the gene clusters which had the ratio between the score for annotation given by Kofamscan to the threshold with which the annotation was done (i.e >30%). It was found that 11 KOs did not pass the filter for the threshold values in the annotation results. The reason for this observation might be that these KOs were underrepresented while the orthologs were formed or in comparison to KOs containing well annotated genes. As a result, 8452 bins were selected that passed the above threshold. These bins were then segregated based on the KOs present, into different CFPs of interest.

4.4 Filtration and extraction of representative mOTUs

Heatmap visualizing the presence and absence of the pathways were plot for the 8452 bins. The following completion threshold were used: at least 60% complete for the module M00173(rTCA) and M00374(DC - HB), 30% complete for the module M00375(HP-HB) and 20 % complete for M00376. The bins were then grouped by the mOTUs, and presence indicates that the KO is present in at least one of the bins. This aggregation was also used to count the frequency of occurrence of the KOs in their respective mOTUs.

Table 2. Summary table for selection of mOTUs from the complete dataset after parsing the annotations from Kofamscan tool. The table contains the modules and the minimum percentage of completion for which the bins were selected, the corresponding number of bins, their respective mOTUs and the number of representative mOTUs selected for further analysis.

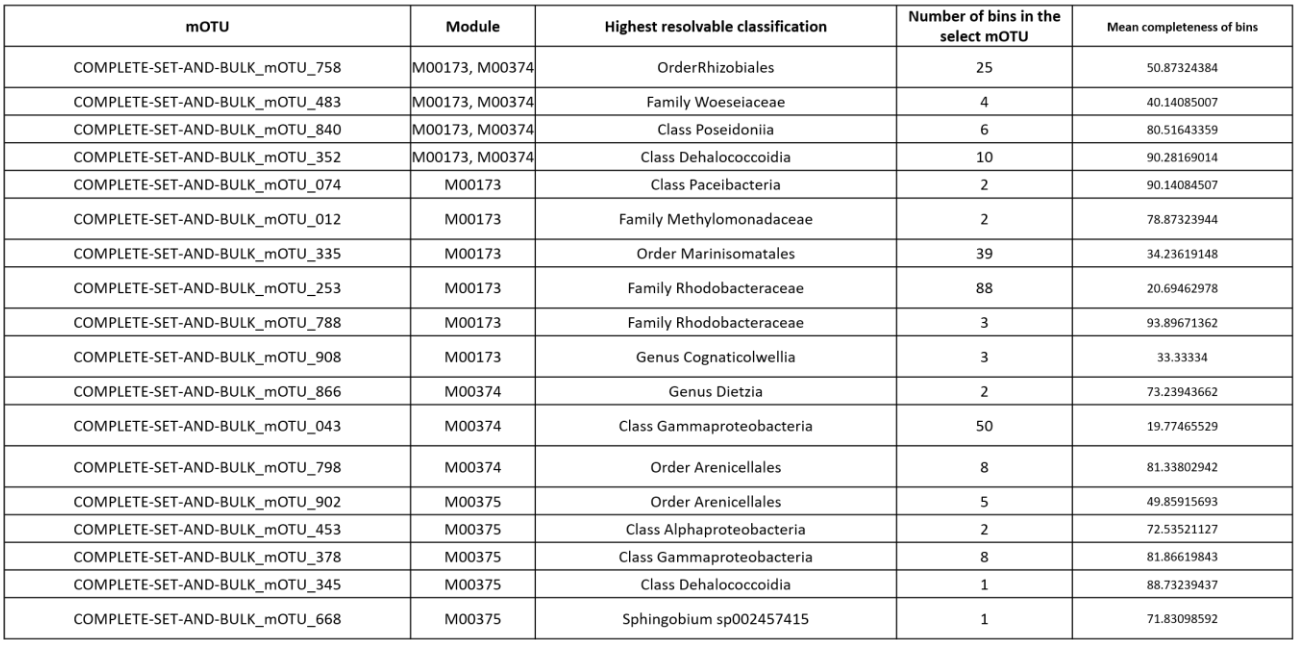


Modules(completion threshold)	Number of bins selected	Number of corresponding mOTU	Number of mOTUs selected
M00173 - rTCA(60%)	281	48	10
M00374 – DC-HB (60%)	333	43	6
M00375 – HP-HB(30%)	500	104	4
M00376 – 3-HP(20%)	48	47	0

In order to consolidate these bins on the basis of not only how complete their genomes are, but it is also presumed that even incomplete bins will contain a significant portion of the pathway that is worth recording. Hence mOTUs that are at least 60% complete were considered for the analysis. These mOTUs were then plotted to visualise the presence and absence of the different pathways. Similar heatmaps made for the key enzymes were also analysed. On the basis of overall completeness and presence of key enzymes (or its alternates as described in KEGG pathway database), 18 mOTUs were selected for the pathways M00173(rTCA), M00374(DC-HP), M00375(HB-HP) & M00376(3-HP) (summarised in Table 2). The representative bins for the mOTUs were recorded for the select mOTUs as well as their respective GTDBtk classification.

Then for the respective taxonomic classification, a background check was done to record the most accessible category for its description of the desired CPFs. Different thresholds of completion for the selection of the mOTUs were used since a lot of bins that contained the required combination of KO terms of the key enzymes were not found with higher minimum completion values. A summary of the mOTUs selected is given in the table 3.

Table 3. Summary of all the select mOTUs for the pathways. This contains the name of the select mOTUs, the modules they were selected for, highest resolvable GTDBtk classification found, the number of bins selected for the mOTU chosen and the mean completeness of the bins in the mOTU.



mOTU	Module	Highest resolvable classification	Number of bins in the select mOTU	Mean completeness of bins
COMPLETE-SET-AND-BULK_mOTU_758	M00173, M00374	OrderRhizobiales	25	50.87324384
COMPLETE-SET-AND-BULK_mOTU_483	M00173, M00374	Family Woeseiaceae	4	40.14085007
COMPLETE-SET-AND-BULK_mOTU_840	M00173, M00374	Class Poseidoniia	6	80.51643359
COMPLETE-SET-AND-BULK_mOTU_352	M00173, M00374	Class Dehalococcoidia	10	90.28169014
COMPLETE-SET-AND-BULK_mOTU_074	M00173	Class Paceibacteria	2	90.14084507
COMPLETE-SET-AND-BULK_mOTU_012	M00173	Family Methylomonadaceae	2	78.87323944
COMPLETE-SET-AND-BULK_mOTU_335	M00173	Order Marinisomatales	39	34.23619148
COMPLETE-SET-AND-BULK_mOTU_253	M00173	Family Rhodobacteraceae	88	20.69462978
COMPLETE-SET-AND-BULK_mOTU_788	M00173	Family Rhodobacteraceae	3	93.89671362
COMPLETE-SET-AND-BULK_mOTU_908	M00173	Genus Cognaticolwellia	3	33.33334
COMPLETE-SET-AND-BULK_mOTU_866	M00374	Genus Dietzia	2	73.23943662
COMPLETE-SET-AND-BULK_mOTU_043	M00374	Class Gammaproteobacteria	50	19.77465529
COMPLETE-SET-AND-BULK_mOTU_798	M00374	Order Arenicellales	8	81.33802942
COMPLETE-SET-AND-BULK_mOTU_902	M00375	Order Arenicellales	5	49.85915693
COMPLETE-SET-AND-BULK_mOTU_453	M00375	Class Alphaproteobacteria	2	72.53521127
COMPLETE-SET-AND-BULK_mOTU_378	M00375	Class Gammaproteobacteria	8	81.86619843
COMPLETE-SET-AND-BULK_mOTU_345	M00375	Class Dehalococcoidia	1	88.73239437
COMPLETE-SET-AND-BULK_mOTU_668	M00375	Sphingobium sp002457415	1	71.83098592

4.5 Prevalence of Different CPFs in the select mOTUs

After manual data analysis and filtering a total of 18 mOTUs were selected for the presence of different CPFs in the CAO microbiota. The basis of selection is as follows. 1. For the select mOTUs, the ones that corresponded to taxa known from literature to have the pathways present (or other probability of presence thereof) were considered. Along with this, the number of KOs that the mOTU has for the pathways that it was filtered for was also critical. 2. In order to validate the CPFs and other pathways presence the support for chemolithotrophy was looked for from the Gapseq tool results for the representative bins for each of these mOTUs.

The results of the Gapseq analysis were used as evidence for the presence of the CFPs that were detected in the selected mOTUs. Out of the 10 mOTUs selected for the presence of the reductive Tricarboxylic acid cycle of fixing carbon, one mOTU (COMPLETE-SET-AND-BULK_MOTU_908) was confirmed. This mOTU was also found to have pathways for Hydrogen oxidation. The mOTU COMPLETE-SET-AND-BULK_MOTU_352 however was annotated for the presence of the incomplete rTCA cycle (M00620) on contrary to rTCA. This mOTU however did not show signs of pathways that would suggest its ability of chemolithotrophy. The mOTU COMPLETE-SET-AND-BULK_MOTU_453 showed the presence of both HP-HB cycle & rTCA cycle. The rest of the pathways, in addition to other special pathways, showed the presence of components of the rTCA cycle. A summary of the results can be found in the table 4 below.

Table 4. Summary of results from the Gapseq tool run for the representative bins for the select mOTUs. This table summarises the mOTU names, the respective representative bins, the CPFs found, and the percentage of the respective pathways found in the representative bins and other "complimentary" pathways of energy sourcing.

mOTUs	Representative bins	Alternative Carbon Fixation Pathways found	Other pathways for energy source
COMPLETE-SET-AND-BULK_mOTU_758	bulkbins-MOSAIC-MIME-BINNING-603_bin-052	rTCA(77%)	Aerobic Sulphur Oxidation(100%)
COMPLETE-SET-AND-BULK_mOTU_352	bulkbins-MOSAIC-MIME-BINNING-ClusterAss-1_bin-070	Incomplete rTCA(71%)	Ammonia assimilation(100%)
COMPLETE-SET-AND-BULK_mOTU_253	bulkbins-MOSAIC-MIME-BINNING-2003_bin-095	rTCA(51%) 3-HP, HP-HB(61%)	Aerobic Hydrogen Oxidation(100%)
COMPLETE-SET-AND-BULK_mOTU_335	bulkbins-MOSAIC-MIME-BINNING-1503_bin-091	rTCA(40%)	
COMPLETE-SET-AND-BULK_mOTU_908	bulkbins-MOSAIC-MIME-BINNING-ClusterAss-2_bin-258	rTCA(75%)	Hydrogen oxidation(aerobic&anaerobic)(50%)
COMPLETE-SET-AND-BULK_mOTU_043	bulkbins-MOSAIC-MIME-BINNING-1703_bin-049	rTCA(66%)	

COMPLETE-SET-AND-BULK_mOTU_453	SRR5819383.103	rTCA(75%)
		Hp-HB(75%)
			Anaerobic Hydrogen oxidation(100%)
			Aerobic sulfur oxidation(100%)

COMPLETE-SET-AND-BULK_mOTU_483	bulkbins-MOSAIC-MIME-BINNING-ClusterAss-2_bin-691	rTCA(66%)	
COMPLETE-SET-AND-BULK_mOTU_840	bulkbins-MOSAIC-MIME-BINNING-2903_bin-044	rTCA(66%)	
COMPLETE-SET-AND-BULK_mOTU_074	bulkbins-MOSAIC-MIME-BINNING-ClusterAss-3a_bin-559	rTCA(50%)	Aerobic Hydrogen oxidation (100%)
COMPLETE-SET-AND-BULK_mOTU_012	MOSAIC-MIME-BINNING-ClusterAss-2_bin-193	rTCA(50%)	

4.6 Pangenomes from Anvi’o

For the pangenome analysis, the mOTUs were filtered for the number of component bins that make up for the evidence of presence of the pathway, as well as the number of KO components pertaining to enzymes of the pathway carefully examined for the presence of key genes as summarised in table 5.

Table 5. The mOTUs selected for the different modules and the number of KO components that were found to be present in those mOTUs.

Module	mOTUs selected	Number of KOs present
M00173	COMPLETE-SET-AND-BULK_mOTU_253	24(of 43)
M00173	COMPLETE-SET-AND-BULK_mOTU_335	25(of 43)
M00173	COMPLETE-SET-AND-BULK_mOTU_352	31(of 43)
M00173	COMPLETE-SET-AND-BULK_mOTU_758	32(of 43)
M00173	COMPLETE-SET-AND-BULK_mOTU_908	23(of 43)
M0374	COMPLETE-SET-AND-BULK_mOTU_043	15(of 19)

The results from the pangenome analysis of the above mOTUs are interactive graphs. These graphs consist of a hierarchical clustering of gene clusters found among different member bins of an mOTU. It also specifies the presence of components of KEGG modules, quality measures of the member bins such as the GC content, redundancy, completion and bin length as assessed by the Anvi’o program. Pangenome plots for the six select mOTUs (in table 4) were done along with distinction of core and accessory genome as well as highlighting the presence of the enzymes that perform the different CFPs. The bins are clustered (their contigs are the elements of clustering) by the probability of their occurrence in either the core or the accessory genomes. This was done in order to distinguish if the enzyme components of the pathway are present in the core or the accessory genome of the organism. The mOTUs were highlighted primarily for the CFPs that they were selected for and eventually, other pathways were also checked for their presence (separately & in combination). The evidence of presence of the combination of pathways can be seen in the search results sections of the search tab. The pangenome plot for one of the mOTUs is explained in detail here and is followed by discussion of the same featured for the rest of the mOTUs.

For the mOTU COMPLETE-SET-AND-BULK_MOTU_253, which belonged to the family Rhodobacteriaceae, the pangenome plot is explained in details here. As can be seen in the plot (highlighted in red as the outermost layer) the presence of M00173(rTCA cycle) can be observed. The presence of other CFPs were also annotated but strong evidenced were found only for M00173 (rTCA) and M00374 (DC-HB). The mOTU is seen to have well annotated component bins(as indicated by the Kofam & KEGG modules layers). The clustering of the bins is done in such a way that the gene clusters are assorted according to their presence in the core or accessory genome (green for core and red for accessory) which can be seen right above the bins in purple. The clustering shows the presence of a significant portion of the genome as accessory that harbours genes for the pathway, wherein the majority of the genes are marked to be in the core of the genomes. It is also observed that there are specific clusters of gene in the accessory genome far from the core (that is not clustered together) that might indicate that the organisms with these parts harbour special genomic segments that are conserved for that particular taxon. The genomic regions that are clustered together on the accessory genome at the far end from the core genome is suspected to occur due to events of horizontal gene transfer, for which the presence of the pathway components was also noted. There happens to be an anomaly bin wherein it is much larger than the other bins as well as having some unique hits. This bin is regarded as contamination and is ignored, but visible in the figure.

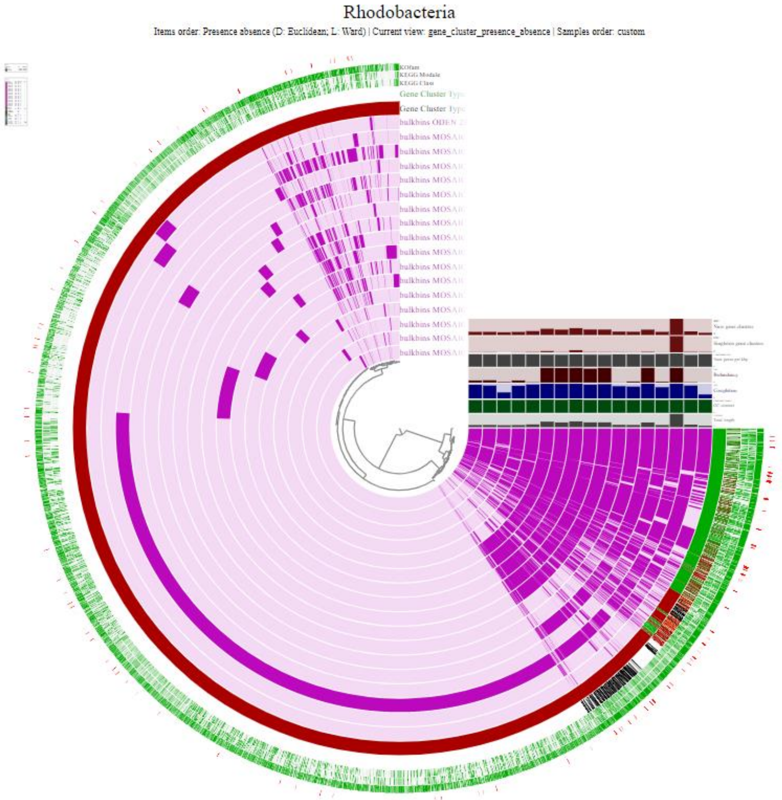


Fig 4. Pangenome plot of COMPLETE-SET-AND-BULK_MOTU_253, Order Rhodobacteria. The purple segments indicate the member bins of the mOTU, followed by the summary statistics at the end of the circle. The segments in green/red around the bins is the presence of the gene clusters in the core(green) or accessory(red) genomes, the outer green layers indicate the presence of the Kofam and KEGG modules(all)

Then the pangenome plots were analysed in similar fashion for the rest of the mOTUs (plots included in Appendix 3). For the mOTU COMPLETE-SET-AND-BULK_MOTU_335, which belongs to the Order Marinisomatales, it was observed that the majority of the gene clusters belonged to the core genome. It can be observed that there is a significant clustering of the pathway components around the core genome as well as containing parts in the accessory genome. There seems to be a difference in the genome sizes of the bins which pertains to having incomplete member bins with different completion percentages or having bins with different genome sizes. (refer to Appendix 3(a))

The mOTU COMPLETE-SET-AND-BULK_MOTU_352(see Appendix 3(b)), that belongs to the order Dehalococcoidia, the genes for the rTCA pathway is spread across the genome (both core and accessory) whereas for the mOTU COMPLETE-SET-AND-BULK_MOTU_758 (see Appendix 3(c)), the genes are present mainly in the accessory genome. For the mOTU COMPLETE-SET-AND-BULK_MOTU_908 (see Appendix 3(d)), which was found to harbour the rTCA cycle from the literature review, the genes of the module were spread across the genome. The core/accessory genome calculation was not performed for this mOTU since it did not have bins with completions that was permissible for the tool that computes the probability.

For the mOTU COMPLETE-SET-AND-BULK_MOTU_043, which is a Gammaproteobacteria, was observed to have the HB-HP pathway (M00374). The distribution of bins shows that there is a bin with larger size which might pertain to redundancy that is considered contamination. The pangenome plot was found to have the components of both the rTCA and HB-HP pathways and both of these pathway components were found to be present in the core genome of the organism to a great extent. (see Fig 5) A lot of these hits were also found to be false positives since it occurs in the bin with high redundancy and hence these hits were ignored for the analysis. This mOTU was also observed to have the components of the rTCA pathway (M00173).

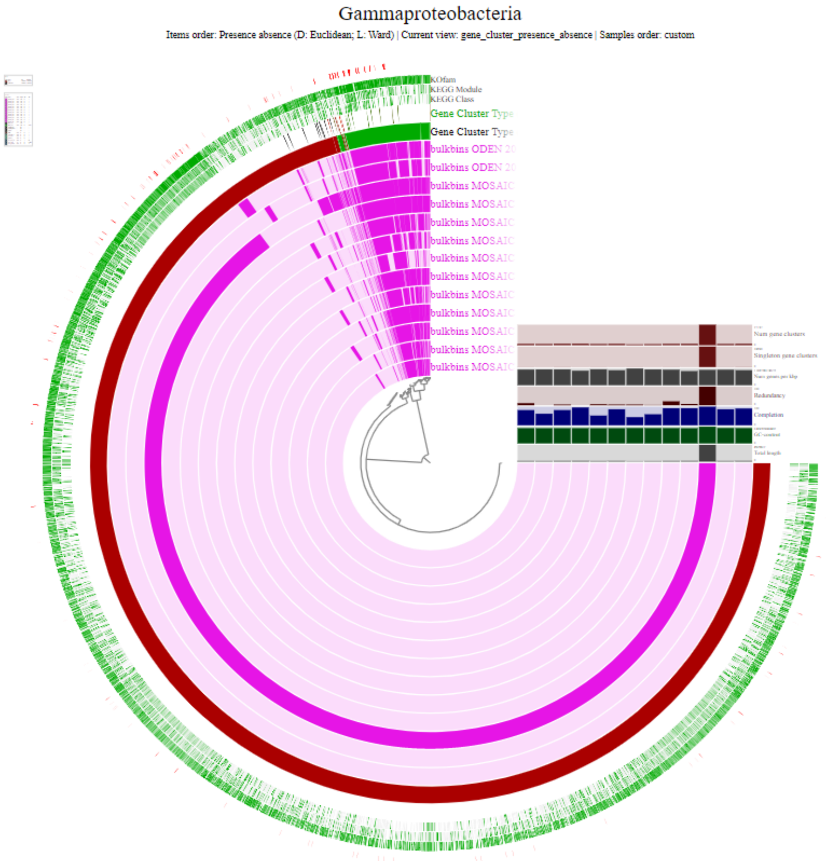


Fig 4. Pangenome plot of COMPLETE-SET-AND-BULK_MOTU_043, class Gammaproteobacteria. The purple segments indicate the member bins of the mOTU, followed by the summary statistics at the end of the circle. The segments in green/red around the bins is the presence of the gene clusters in the core(green) or accessory(red) genomes, the outer green layers indicate the presence of the Kofam and KEGG modules.

5 Discussion

The significance of Carbon fixation in the cycling of elements crucial for the existence and maintenance of life in various ecosystems including marine, atmosphere and continental biosphere. Hence understanding how the different components of the ecosystem that are major contributors in sustaining this cycle is deemed crucial. The CAO is a major unexplored biome that can act as a crucial carbon sink(Poff *et al.* 2021). This study hence focuses on how the CAO assimilates inorganic carbon and what special mechanisms enables this process to be maintained in conditions of low light and energy. The course of the Discussion section has defence and reasoning of adaptation made for analysis tools and methods followed by biological interpretation and inferences.

5.1 Choosing Kofamscan over EggNOG for reliable annotation

The study started with annotations for all the complete set of MAGs in the form of EggNOG annotations. Parsing of these files resulted in the loss of detection for some integral component KOs that validates the occurrence of the CPFs of interest. And hence a closer look into the annotation system and the construction of the reference database that EggNOG uses was done. EggNOG mapper utilizes blastx-like searches to map the protein sequences to putative ab-initio orthologous gene clusters (COGs). (Cantalapiedra *et al.* 2021) These COGs were formed by comparing both known and predicted proteins from all completely available microbial genomes. The proteins in each COG category will be orthologous in at least three lineages and correspond to an ancient, conserved domain. (http://clovr.org/docs/clusters-of-orthologous-groups-cogs/) Upon examining the relative assignment of KO terms to these COGs, it was found that some COGs contain genes annotated with several different KOs and many KO terms were found to be in several different COGs. Hence trying to annotate EggNOGs unambiguously with KOs led to loss of KOs we were interested in finding as well as made the annotation unreliable. This annotation was therefore disregarded from doing further analysis. In order to overcome the problem of ambiguity that EggNOG annotation restricted the workflow with, a rather direct annotation of the genomes with the Kofam Database was done with the tool Kofamscan which is equivalent to the online tool Kofamkoala (https://www.genome.jp/tools/kofamkoala/) by KEGG. This uses KEGG Orthologies and Hidden Markov Models, which are much more reliable and accurate than blast-like procedures such as EggNOG to annotate the data. This tool is also based on curated HMMs which makes the annotation more trustworthy. The problem of missing KOs for the pathways existed in the Kofamscan annotation as well but the count was relatively low, and it was also that the missing KOs were not key enzymes whose absence would signify that the pathway is not present.

5.2 Selection and Consolidation of the dataset

Upon parsing of the Kofamscan annotations, in order to reduce the dimensionality, the dataset was filtered in various steps (for details see Section 3.2.3). For the following steps of Gapseq analysis and Pangenomics, two independent consolidations were used sequentially, and they were performed atomically. This is because the two analyses were done in parallel and hence it wasn't streamlined. For the Gapseq tool, all the representative bins for the selected mOTUs were used. In order to perform pangenome analysis, three conditions were taken into account. The literature evidence that hint for the presence of the pathways that were found for the mOTU's taxa from the previous analyses, the number of pathway components that the mOTU contains and the number of component bins, which are the components of the mOTU that is suspected to employ the pathway. From the literature survey, the overall metabolic profile was noted for the mOTUs. The GTDBtk classification was employed for this step. The completeness, meaning the number of components that the mOTU possess was selected in such a way that at least 50% of the components are present. This threshold is low since the bins that make up the mOTUs might be of varied completeness and the genomic regions that had not been captured might contain components of the pathway. Next constrain is the number of bins that make up the mOTUs. False positives are less likely in the mOTUs with multiple bins. By the above criteria, 6 mOTUs were selected from the 18 mOTUs selected for the presence/absence of KO terms (key enzymes and other components) in the previous step. Among the six selected mOTUs, one mOTU, COMPLETE-SET-AND-BULK_MOTU_908 is an outlier in the number of component bins criteria since it has only 3 component bins compared, while the rest has >10 each. This mOTU was still chosen because literature survey supported the presence of the detected pathway.

5.3 The trend observed in occurrence of CFPs across the bacteria and archaeal tree of life

The CAO is characterised with thick layer of ice around the polar north that sits on deep oceanic trenches that harbour one of the most diverse ecosystems. The major driver of this ecosystem, primarily involved in nutrient cycling, being primary producers in the food web as well as sustaining the ecosystem is its microbiota. As primary producers many of these microbes, predominantly bacteria and archaea involve in the essential process of carbon fixation. These microbes, living in such ecosystems, will not be exposed to photic sources of energy and hence must rely on other processes in order to aid the organism in efficiently fixing carbon. Unlike a photosynthetic bacteria and archaea, that employ photoautotrophy to aid in the fixation of carbon, these prokaryotes employ specialised metabolic pathways to fix the carbon around them. They do it by fixing carbon in air or water, in organic or inorganic forms. These are defined as alternative Carbon fixation pathways in prokaryotes. The different carbon fixation pathways are adapted by different members of the bacterial and archaeal tree of life. The most prevalent among these pathways is known to be the rTCA cycle apart from the photosynthetic CBB cycle(refer Fig 4).(Garritano *et al.* 2022)

There are other defined pathways that the prokaryotes adapt, at times even employ two. These alternative pathways are found to be less expensive energetically(Hügler & Sievert 2011) since these organisms thrive in conditions of availability of less sources of energy. In order to adapt to the environment, they live in, these microbes derive energy in the form of electron donors by mechanisms of oxidation (sulfur, nitrogen, Hydrogen). Hence, these mechanisms can also be found in combination in the microbes and therefore they are termed Chemolithoautotrophs. This study attempts to look around for such Chemolithoautotrophs in the Central Arctic Ocean microbiota, which will explain how these microbes thrive in such environments. The aim was to identify the CFPs and to verify of the taxa in which they were previously regarded to occur (see Section 2) in was what is observed.

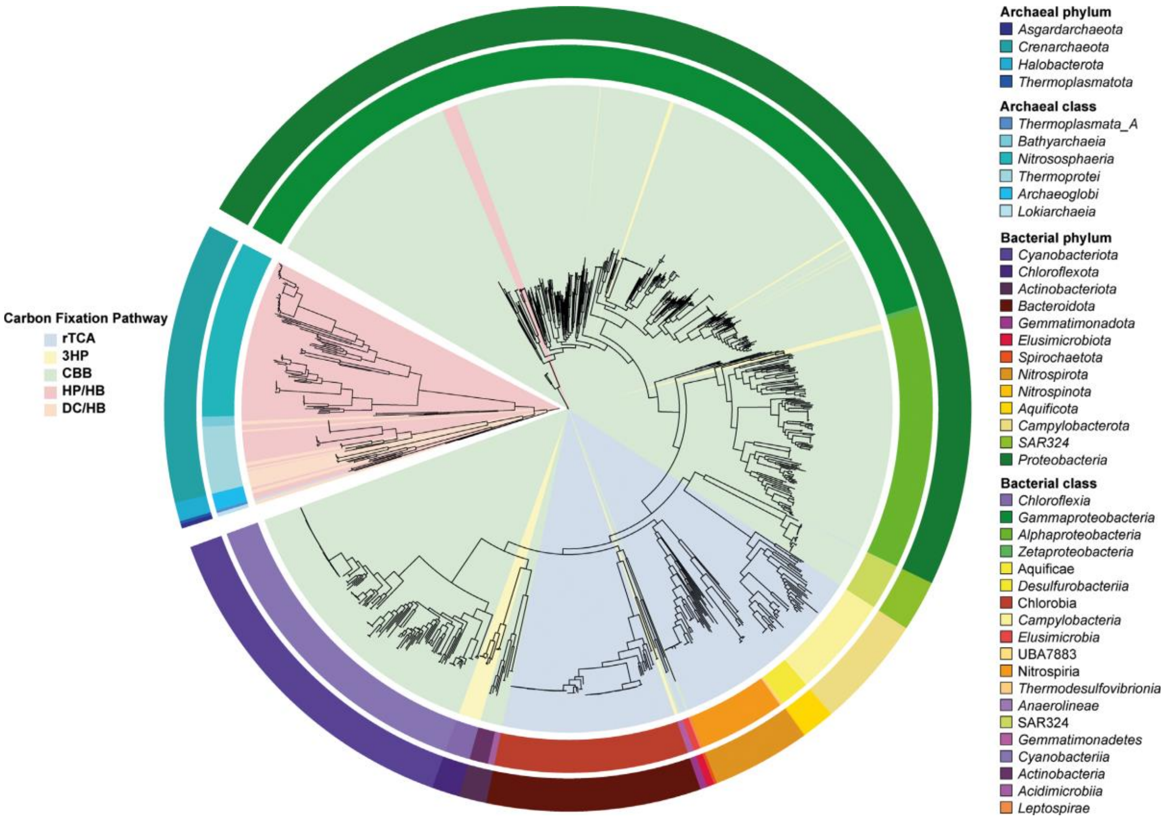


Fig. 5 Summary of distribution of different CPFs in bacteria & archaeal tree of life from (Garritano *et al.* 2022)

5.4 Observations from the Central Arctic Ocean Microbiota

Out of the observed 10 mOTUs that were proposed to undergo the process of rTCA, only the Gammaproteobateria, genus Cognaticowellia (mOTU 908) was observed to have strong evidence of a functional rTCA carbon fixation pathway. (Vineis et al. 2021) However the mOTU that was observed in the dataset had 17 out of 43 KO terms that correspond to the enzyme components of the pathway. Upon further scrutiny of this annotation upon the functional prediction of the Gapseq too, it was found that the genus Cognaticowellia indicated the presence of rTCA (75%), which also passed for the threshold of the tool. This tool also detected the presence of Hydrogen oxidation, with the highest confidence indicating that this organism oxidises hydrogen to derive energy and support the fixation of carbon by rTCA cycle. The pangenome (see Appendix 3(d)) of this mOTU was based on of three component bins. One of the three bins showed a high level of completeness as well as well-defined gene cluster (annotated by Kofam). This suggests that this bin(bulkbins-MOSAIC-MIME-BINNING-ClusterAss-2_bin-258) might be a defined species. The presence of the pathway, according to the pangenome analysis, indicates that the pathway components are present across the entire genome and along with the strength of the presence of the KEGG modules indicated, supports the presence of the rTCA cycle in this organism.

The mOTU (COMPLETE-SET-AND-BULK_mOTU_758) with the highest number of component KO terms (32 of 43) pertaining to the mOTU that was classified as a member of the order Rhizobiales, was previously recorded to have parts of both rTCA as well as DC-HB cycle. This suggests either the presence of both of the pathways or the fact that these pathways have the tendency to use the same enzymes from other metabolic pathways, which was also supported by the pangenome plot for the presence of both pathways (see Appendix 3(c)). The results of Gapseq confirmed the presence of rTCA in this organism but could not validate the presence of DC-HB cycle which might be due to the absence of components in the incomplete representative bin. This organism employs Aerobic Sulphur oxidation which is hypothesised as an energy source wherein it utilises sulfur compounds from its surroundings. While this order is studied well for its occurrence in ecosystems other than marine, no model organism or benchmark study was found for its behaviour in the marine ecosystem.

The other taxa with predominant features of this pathway present is the family Woeaseiale (COMPLETE-SET-AND-BULK_mOTU_483). This constitutes 25% of sediment microbiome diversity (Perner *et al.* 2022) that predominantly employs CBB cycle to fix carbon in contrary to the observation(31 of 43 KOs present). This suggests that the particular genera had not been well defined or studied might employ the rTCA cycle wich complies with the Gapseq results. It is also suggested to have a supporting chemolithoautotrophy wherein it oxidises sulphur to meet its energy needs. However, no evidence of such adaptation was spotted in Gapseq results. This organism also lacks well studied representatives in marine ecosystems (Mußmann *et al.* 2017, Buongiorno *et al.* 2020).

Another interesting find that happened to have the rTCA cycle was the class Dehalococcoidia (COMPLETE-SET-AND-BULK_mOTU_352). This bacteria, in other resources, was observed to have the Wood - Ljundhal pathway (M00377) and 3HP pathway (M00376)(Yang *et al.* 2020). This bacterium was observed to utilize Hydrogen to harvest energy while also carrying out C1 fixation by oxidising methane using incomplete 3HP pathway. It was also observed that the model organism, *D. mccarttyi* also employed nitrogen fixation in pristine conditions (Löffler *et al.* 2013). On contrary, the Gapseq results rendered the presence of incomplete TCA cycle (M00620) and whereas the ability of the bacteria to oxidise nitrogen was found. The pangenome plots however shows the presence of the rTCA cycle as well as the CBB cycle but not the Wood-Ljungdahl pathway. This might suggest the presence of the combination of the pathways or the presence of one of these pathways.

The other bacteria that utilise methane as a carbon source that was detected to have the rTCA cycle was the Methylominadaceae (COMPLETE-SET-AND-BULK_mOTU_012) family bacteria. Not a lot of information was found about the other metabolic pathways that they employ from literature as well as from Gapseq. This was also previously reported to be found in sediments (Orata *et al.* 2018). On the other hand, yet another sediment dwelling bacteria was found to carry out the rTCA cycle, a bacteria belonging to the class Paceibacteria(COMPLETE-SET-AND-BULK_mOTU_074), which also lacked extensive study on its marine dwelling species (Zhao *et al.* 2022). Gapseq however predicted that this bacterium possessed a part of rTCA cycle as well as an aerobic hydrogen oxidative mechanism.

It was also found that the bacteria of the order Marinisomatales, which was found to be in the Atlantic marine ecosystem. This order was however not been studied for its CPFs but it was observed to utilise the process of reduction of nitrogenated carbon compounds that might shine new light onto the metabolic pathway of these organisms in the Arctic ecosystem (Coutinho *et al.* 2021). Further two mOTUs (253 & 788) were found to belong to the family Rhodobacteraceae. They were observed to possess characteristics of photoheterotrophy or photolithoautotrophy by employing CBB pathway to fix carbon as well as employ dark carbon fixation coupled with nitrogen fixation (Pujalte *et al.* 2014, Connelly *et al.* 2014). However, pangenomics of the mOTU COMPLETE-SET-AND-BULK_MOTU_253 shows the presence of rTCA cycle. (See Fig 4)

For the DC-HB cycle and HP-HB cycle (M00374 & M00375), 11 mOTUs were identified to have a significant portion of the KOs corresponding to the overall as well as key components of the pathway through the selection with heatmaps. However, none of these mOTUs were found to have a significant presence of these modules in the Gapseq functional annotation tool results. This is potentially due to poor assembly or annotation of the sample bins in hand so that they were not detected during multiple rounds of filtration of the dataset. There was also some data loss occurred due to missing bin information and their sequences, incomplete bins, strict annotation leading to loss of less annotated KOs, etc. The loss of KOs due to strict annotation is very effective in the case of Ocean microbiota, especially one that is underexplored that these genes or pathways and its components might not be well studied nor have enough hits in the reference databases. The pangenome analysis and visualisation however indicated the presence of the HB-HP pathway(M00374) in the Gammaproteobacteria (COMPLETE-SET-AND-BULK_MOTU_043) found and selected. This pathway was also found to be present in a few other mOTUs, but it requires deeper exploration and confirmation of these pathways in the mOTUs.

There also happen to exist some controversial results indicating the presence/absence of pathways in the said mOTUs that they were selected for, which could be explained by the fact that the Anvi'o tool utilises its inbuilt gene clustering mechanism primarily on the data from the genome sequence, followed by hmm profiling and annotation. For the dataset utilised in the study, clustering was done separately and then annotated. This might have resulted in arbitrary changes in annotation and hence would have resulted in different results. The dataset analysed through Gapseq is also expected to be subjected to stricter constraints than that of Anvi'o which was much more flexible. Nevertheless, further experimental study of genetic composition, enzyme activity and pathway reconstruction would collectively help in better understanding of this ecosystem and its interactions.

6 Conclusion

The Central Arctic Ocean microbiota is a goldmine of unexplored biological diversity with numerous special characterisations that its components take up to survive in harsh conditions. Apart from photoautotrophy, the microbes employ mechanisms of chemolithotrophy, which aids in providing energy for other crucial metabolic processes essential for the survival of the organism. The aim of the study was to describe the presence of different alternative carbon fixation pathways found in prokaryotes in this ecosystem and their "supporting" metabolisms that feed them with energy. Alternative CFPs were found to occur in 192 mOTUs out of 910 mOTUs assembled for the complete dataset. It was found that the reductive Tricarboxylic Acid cycle is the most prevalent, in concordance with previous study, along with utilisation of inorganic electron donors to sustain life in conditions of prolonged darkness. This mechanism was found to occur in 6 different clades of bacteria, with varied adaptations in sourcing electrons. It was also observed that the reductive Acetyl-CoA cycle, which was said to be prevalent across several marine microbial taxa, was not found in the dataset which might be due to poor representation in unique microbial communities. Traces of 3-HP, HP-HB & and incomplete rTCA cycle was found but no strong evidence was recorded which is hypothesised to occur because of low resolution annotation due to lack representation in public databases. However, findings of these special adaptations are believed to build curiosity for further exploring this ecosystem for its wonders.

7 References

https://www.pame.is/images/03_Projects/EA/LMEs/Factsheets/13_Central_Arctic_Ocean_LME_.pdf

Azam F, Fenchel T, Field J, Gray J, Meyer-Reil L, Thingstad F. 1983. The Ecological Role of Water-Column Microbes in the Sea. Marine Ecology Progress Series 10: 257–263.

Bassham JA, Benson AA, Calvin Melvin. 1950. THE PATH OF CARBON IN PHOTOSYNTHESIS. Journal of Biological Chemistry 185: 781–787.

Berg IA. 2011. Ecological Aspects of the Distribution of Different Autotrophic CO2 Fixation Pathways. Applied and Environmental Microbiology 77: 1925–1936.

Buongiorno J, Sipes K, Wasmund K, Loy A, Lloyd KG. 2020. Woeseiales transcriptional response to shallow burial in Arctic fjord surface sediment. PLOS ONE 15: e0234839.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution 38: 5825–5829.

Connelly TL, Baer SE, Cooper JT, Bronk DA, Wawrik B. 2014. Urea Uptake and Carbon Fixation by Marine Pelagic Bacteria and Archaea during the Arctic Summer and Winter Seasons. Applied and Environmental Microbiology 80: 6013–6022.

Coutinho FH, von Meijenfeldt FAB, Walter JM, Haro-Moreno JM, Lopéz-Pérez M, van Verk MC, Thompson CC, Cosenza CAN, Appolinario L, Paranhos R, Cabral A, Dutilh BE, Thompson FL. 2021. Ecogenomics and metabolic potential of the South Atlantic Ocean microbiome. Science of The Total Environment 765: 142758.

Dang H, Chen C-TA. 2017. Ecological Energetic Perspectives on Responses of Nitrogen-Transforming Chemolithoautotrophic Microbiota to Changes in the Marine Environment. Frontiers in Microbiology 8:

Garritano AN, Song W, Thomas T. 2022. Carbon fixation pathways across the bacterial and archaeal tree of life. PNAS Nexus 1: pgac226.

Holo H. 1989. Chloroflexus aurantiacus secretes 3-hydroxypropionate, a possible intermediate in the assimilation of CO 2 and acetate. Springer 252–256.

Huber H, Gallenberger M, Jahn U, Eylert E, Berg IA, Kockelkorn D, Eisenreich W, Fuchs G. 2008. A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation cycle in the hyperthermophilic Archaeum Ignicoccus hospitalis. Proceedings of the National Academy of Sciences 105: 7851–7856.

Hügler M, Sievert SM. 2011. Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. Annual Review of Marine Science 3: 261–289.

Jang K, Woo KS, Kim J-K, Nam S-I. 2023. Arctic deep-water anoxia and its potential role for ocean carbon sink during glacial periods. Communications Earth & Environment 4: 1–10.

Krabberød AK, Deutschmann IM, Bjorbækmo MFM, Balagué V, Giner CR, Ferrera I, Garcés E, Massana R, Gasol JM, Logares R. 2022. Long-term patterns of an interconnected core marine microbiota. Environmental Microbiome 17: 22.

Le Moullec M, Bender M. 2021. Impacts of Global Warming on Arctic Biota. doi 10.1007/978-3-030-81253-9_11.

Ljungdhal LG. 1986. The Autotrophic Pathway of Acetate Synthesis in Acetogenic Bacteria. Annual Review of Microbiology 40: 415–450.

Löffler FE, Yan J, Ritalahti KM, Adrian L, Edwards EA, Konstantinidis KT, Müller JA, Fullerton H, Zinder SH, Spormann AM. 2013. Dehalococcoides mccartyi gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, Dehalococcoidia classis nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi. International Journal of Systematic and Evolutionary Microbiology 63: 625–635.

Mock T, Boulton W, Balmonte J-P, Barry K, Bertilsson S, Bowman J, Buck M, Bratbak G, Chamberlain EJ, Cunliffe M, Creamean J, Ebenhöh O, Eggers SL, Fong AA, Gardner J, Gradinger R, Granskog MA, Havermans C, Hill T, Hoppe CJM, Korte K, Larsen A, Müller O, Nicolaus A, Oldenburg E, Popa O, Rogge S, Schäfer H, Shoemaker K, Snoeijs-Leijonmalm P, Torstensson A, Valentin K, Vader A, Barry K, Chen I-MA, Clum A, Copeland A, Daum C, Eloe-Fadrosh E, Foster B, Foster B, Grigoriev IV, Huntemann M, Ivanova N, Kuo A, Kyrpides NC, Mukherjee S, Palaniappan K, Reddy TBK, Salamov A, Roux S, Varghese N, Woyke T, Wu D, Leggett RM, Moulton V, Metfies K. 2022. Multiomics in the central Arctic Ocean for benchmarking biodiversity change. PLOS Biology 20: e3001835.

Mußmann M, Pjevac P, Krüger K, Dyksma S. 2017. Genomic repertoire of the Woeseiaceae/JTB255, cosmopolitan and abundant core members of microbial communities in marine sediments. The ISME Journal 11: 1276–1281.

Orata FD, Meier-Kolthoff JP, Sauvageau D, Stein LY. 2018. Phylogenomic Analysis of the Gammaproteobacterial Methanotrophs (Order Methylococcales) Calls for the Reclassification of Members at the Genus and Species Levels. Frontiers in Microbiology 9:

Perner M, Wallmann K, Adam-Beyer N, Hepach H, Laufer-Meiser K, Böhnke S, Diercks I, Bange HW, Indenbirken D, Nikeleit V, Bryce C, Kappler A, Engel A, Scholz F. 2022. Environmental changes affect the microbial release of hydrogen sulfide and methane from sediments at Boknis Eck (SW Baltic Sea). Frontiers in Microbiology 13:

Pujalte MJ, Lucena T, Ruvira MA, Arahal DR, Macián MC. 2014. The Family Rhodobacteraceae. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed.). The Prokaryotes: Alphaproteobacteria and Betaproteobacteria, pp. 439–512. Springer, Berlin, Heidelberg.

Sánchez-Andrea I, Guedes IA, Hornung B, Boeren S, Lawson CE, Sousa DZ, Bar-Even A, Claassens NJ, Stams AJM. 2020. The reductive glycine pathway allows autotrophic growth of Desulfovibrio desulfuricans. Nature Communications 11: 5090.

Vineis JH, Bulseco AN, Bowen JL. 2021. Microbial dark carbon fixation fueled by nitrate enrichment. 2021.08.24.457596.

Wächtershäuser G. 1990. Evolution of the First Metabolic Cycles. Proceedings of the National Academy of Sciences of the United States of America 87: 200–204.

Wood HG. 1991. Life with CO or CO2 and H2 as a source of carbon and energy. FASEB journal: official publication of the Federation of American Societies for Experimental Biology 5: 156–163.

Yang Y, Zhang Y, Cápiro NL, Yan J. 2020. Genomic Characteristics Distinguish Geographically Distributed Dehalococcoidia. Frontiers in Microbiology 11:

Zhao R, Farag IF, Jørgensen SL, Biddle JF. 2022. Occurrence, Diversity, and Genomes of “Candidatus Patescibacteria” along the Early Diagenesis of Marine Sediments. Applied and Environmental Microbiology 88: e0140922.

Zimmermann J, Kaleta C, Waschina S. 2021. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biology 22: 81.

13 Central Arctic Ocean LME.

8 Appendices

Appendix 1

Heatmap from parsing annotations of EggNOG mapper

https://drive.google.com/file/d/1DOggN_ivPTv2nvnx-UO5T4rIEAQ6xPAu/view?usp=sharing

Appendix 2

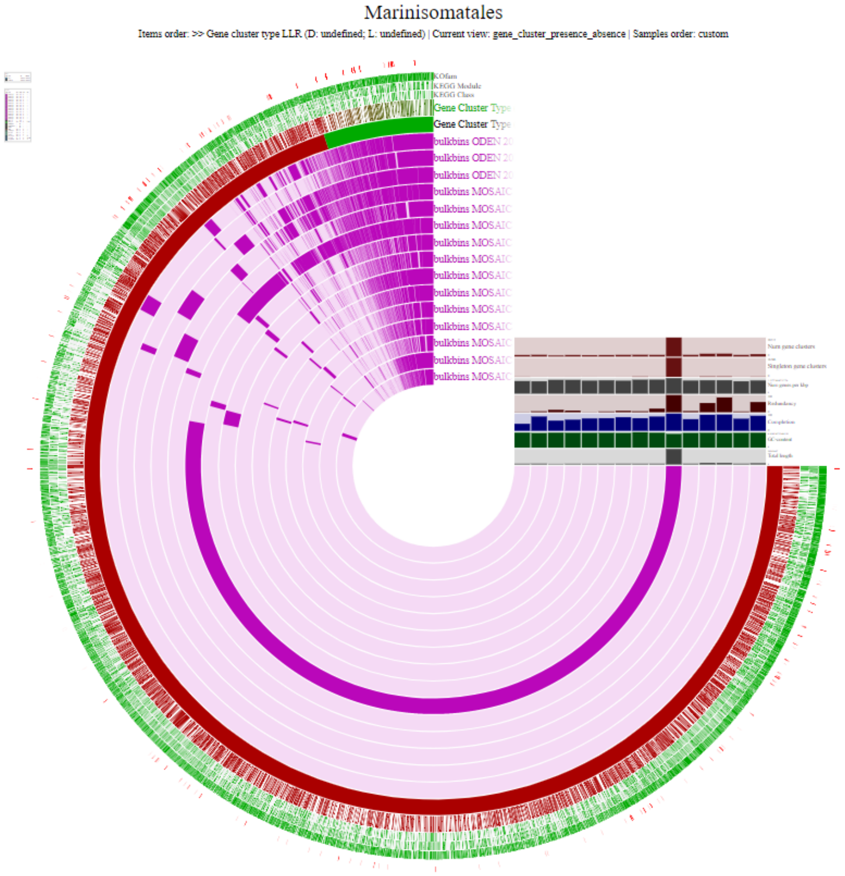
Heatmaps for Kofamscan annotation

https://drive.google.com/file/d/10IHhr4IMZfY3qqCfdH_IbhJP-LEfUCP7/view?usp=sharing

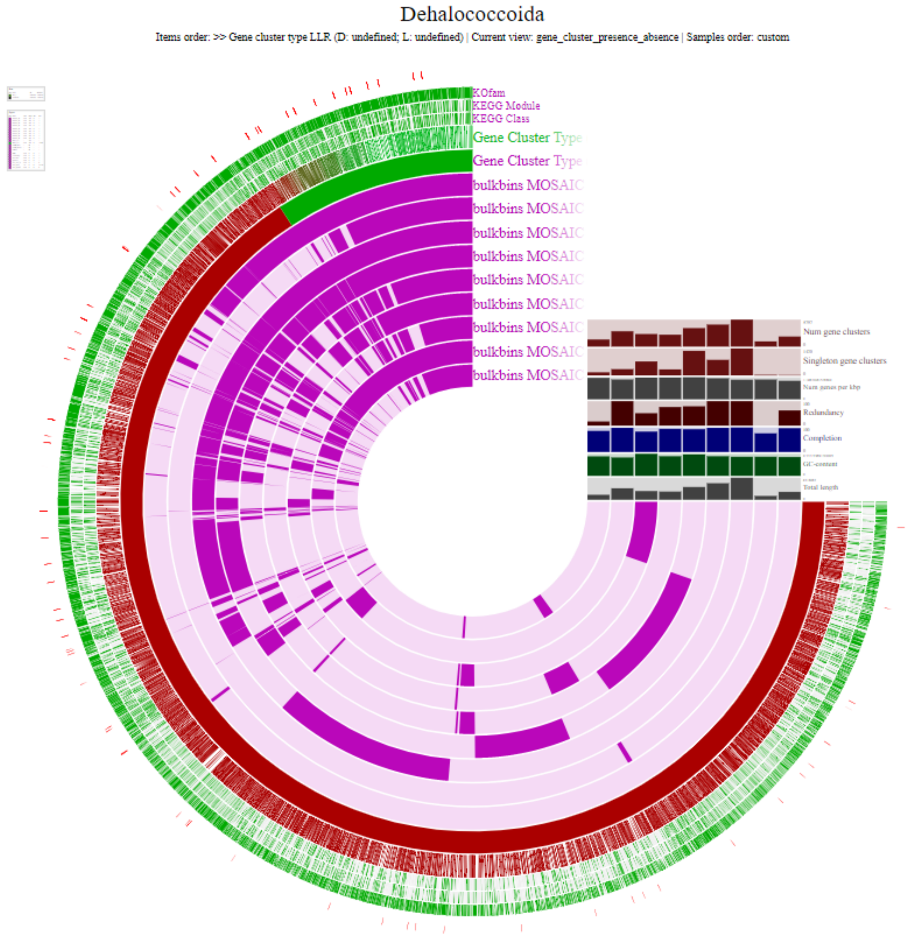
Appendix 3

Pangenome plots for the mOTUs

a) COMPLETE-SET-AND-BULK_MOTU_335, Marinisomatales



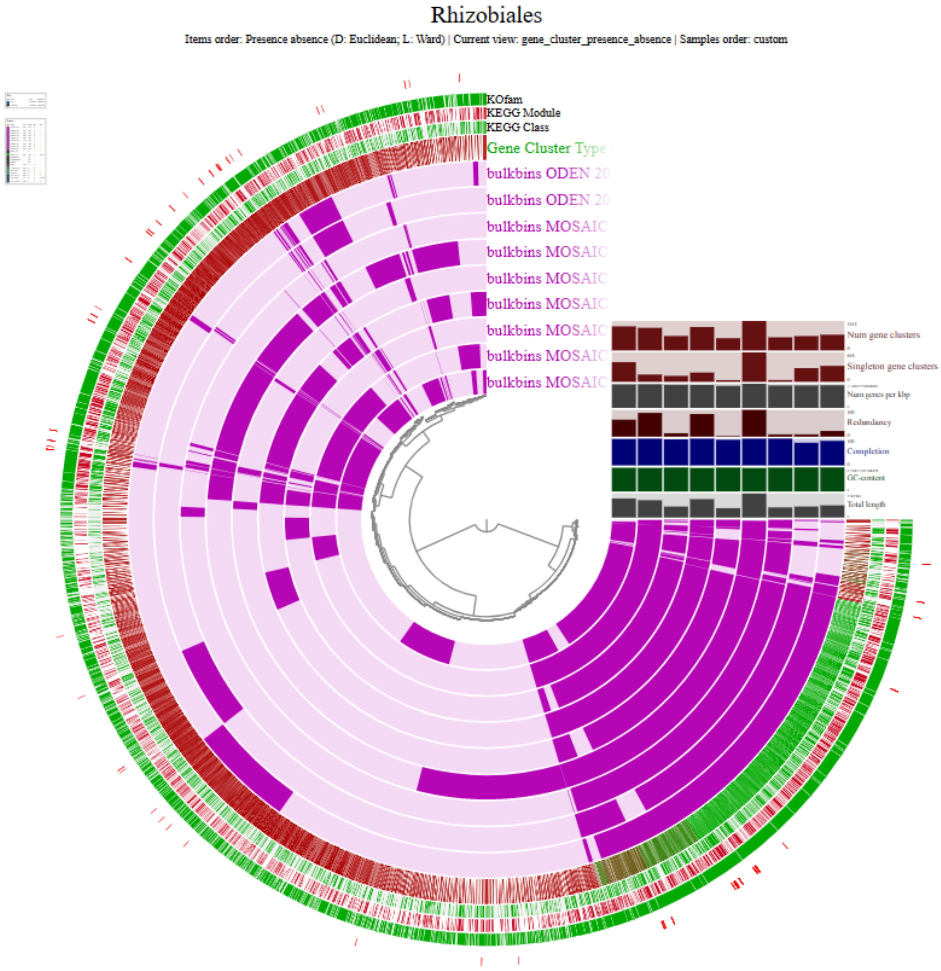
b) Pangenome plot for mOTU COMPLETE-SET-AND-BULK_MOTU_352, Dehalococcoidia



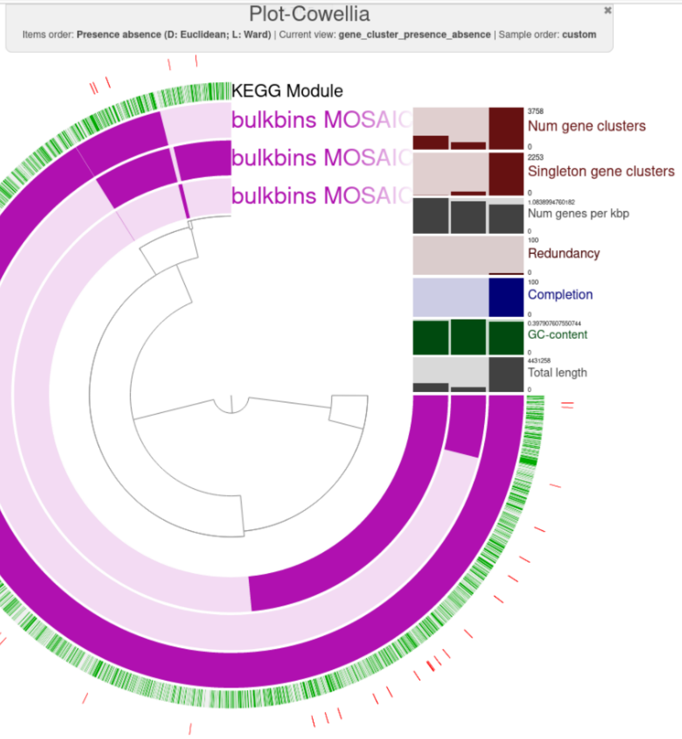
Dehalococcoida

Items order: >> Gene cluster type LLR (D: undefined; L: undefined) | Current view: gene_cluster_presence_absence | Samples order: custom

c) Pangenome plot for mOTU COMPLETE-SET-AND-BULK_MOTU_758, Rhizobiales



d) Pangenome plot for mOTU COMPLETE-SET-AND-BULK_MOTU_908, Cowellia



Appendix 4

Data collection from experiments https://drive.google.com/file/d/14JaesZxrNQ3JqbbOjQcp6D-fv2Oz5mzi/view?usp=sharing