Subantarctic Indian Ocean microbial gene duplications and their role in biogeochemical cycling

by

Caitlyn Fourie

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae: Genetics

In the Faculty of Natural and Agricultural Sciences Department of

Biochemistry, Genetics and Microbiology

University of Pretoria

Pretoria

24th April 2023

Supervisor: Prof. T. P. Makhalanyane

Co-supervisor: Dr. O. Mogase

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Declaration

I, Caitlyn Heather Fourie, declare that the thesis/dissertation ‘Subantarctic Indian Ocean microbial gene duplications and their role in biogeochemical cycling’, which I hereby submit for the degree Magister Scientiae (M.Sc.): Genetics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Caitlyn Heather Fourie

M.Sc. candidate

24th April 2023

Acknowledgements

First I'd like to thank my parents for their never ending support. Dad, thank you for encouraging me to never give up and inspiring me to work hard by your example. Thank you for giving me the courage and support to see this through. Mom, thank you for motivating me and checking in constantly. To the rest of my family, thank you for all the thoughts, prayers and kind words that kept me going.

To my 'ship-wife' Mancha, for teaching me 'how to science'. I wouldn't even know how to filter water if it was not for you. Thank you for all the laughs, cries, impromptu jam sessions and constant encouragement. Thank you Shads for helping me to keep going through all the hurdles. To the rest of my friends, thanks for listening to my constant complaints and sticking around, as well as all the encouragement and much needed breaks that kept me sane.

Thank you to my supervisor, Prof. Thulani Makhalanyane, for the support and opportunity. I have learnt and grown so much from this experience. Visiting Marion Island has been a dream of mine and the experience was priceless. Thank you Dr. Oliver Mogase for all your help and patience, I couldn't have done this without you and your many scripts. Thank you as well to the helpful crew of the S. A. Agulhas II and the EU Horizon 2020 project and AtlantECO for funding this project.

Table of contents

List of Figures viii

List of Tables xii

List of Supplementary Data xv

List of abbreviations xvii

Abstract 1

Chapter 1 3

1. Introduction 4

2. Gene duplication 7

2.1. The role of duplicated genes 8

2.2. The fate of gene duplicates 9

2.3. The rate of gene duplication 11

2.4. Examples of gene duplications 12

2.5. Gene duplicates in marine microbes 14

3. Adaptation 15

3.1. Mechanisms of adaptation in microbial communities 16

4. Evolution of microbial genomes 17

5. Microbial communities of the Subantarctic Indian Ocean 18

6. Oceanic microbial community functions 22

6.1. Nitrogen cycling 25

6.2. Carbon cycling 25

7. Techniques used to study microbial genomes and gene duplications 26

7.1 Metagenomics 27

7.2 Studying marine microbiomes with metagenomic approaches 29

8. Pangenomes 29

9. Summary 31
10. Problem statement 32
11. Hypothesis, aims and Objectives 33
11.1. Hypothesis: 33
11.2. Aim: 33
11.3. Objectives: 33
12. References 33
Chapter 2 47
1. Sampling 48
1.1. Sample collection from the Subantarctic Indian Ocean (SIO) 48
1.1.1. Processing of samples 50
1.2. Nutrient composition analysis 51
1.2.1. Nutrient measurements 51
1.3. Extracellular enzyme activity assays 52
1.3.1. Extracellular enzyme activity test assays 52
1.3.2. Extracellular enzyme activity measurements 53
1.4. DNA extraction and agarose gel electrophoresis 54
2. Metagenome-assembled genomes (MAGs) curation 54
2.1. Shotgun metagenomic sequencing 54
2.2. MAGs assembly 55
2.3. MAGs classification 55
3. Gene duplication analysis and assignment of putative paralogs 56
3.1. Gene duplication analysis 56
3.2. Assigning function to the putative paralogs 56
4. Statistical analyses 58

5. A Genus-level pangenomic analysis of high and medium quality MAGs 59
6. References 60
Chapter 3 63
1. Abstract 64
2. Introduction 66
3. Results 70
3.1. Physiochemistry of the PEI 70
3.1.1. Physical characteristics 70
3.1.2. Chemical characteristics 73
3.2. Extracellular enzymatic activity analysis 79
3.3. DNA extractions and shotgun metagenomic sequencing 83
3.4. Phylogenetic diversity of the MAGs 85
3.5. Gene duplications 87
3.5.1. Putative paralogs 87
3.5.2. Duplicated gene functions 90
3.6. The location of putative biogeochemical cycling genes within the SIO pangenome 93
4. Discussion 95
4.1. Chemical characteristics of PEI marine microbiomes 96
4.2. Extracellular enzymatic activities of PEI marine microbiomes 98
4.2.1. Carbon acquisition (BX and BG) 99
4.2.2. Phosphorus acquisition (AP) 100
4.2.3. Nitrogen acquisition (NAG) 101
4.3. Physical factors of PEI marine microbiomes 102
4.4. Taxonomy of the PEI marine microbiomes 104

4.5. Factors affecting gene duplications in the SIO 106

4.6. The functional capacity of gene duplicates in the SIO 108

4.6.1. Nitrogen cycling 109

4.6.2. Carbon cycling 110

4.6.3. Phosphorus cycling 113

4.7. Pangenome 114

5. References 116

Chapter 4 127

Synthesis 127

Supplementary data 132

List of Figures

Figure 1: The fate of duplicated genes based on an image by Voordeckers and Verstrepen (2015). Most duplicates will be lost through the accumulation of deleterious mutations. Duplicates can be retained in subfunctionalization, neofunctionalization, nonfunctionalization or if they play a role in gene dosage..11

Figure 2: Map of the Prince Edward Islands showing the sampling locations and study area. Water was collected from 14 stations around the PEI's in the 'T-transect' and the 'X-M-transect', as well as 11 stations between Cape Town, South Africa and the PEI's. The red points represent the stations that were sampled. Maps were created using Ocean Data View version 5.6.2 (Schlitzer 2023)……...50

Figure 3: Boxplots of the physiochemical variables according to depth including a) pressure, b) temperature, c) salinity, d) electrical conductivity, e) oxygen measured at all the inter-island stations on the 'T-transect' and the 'X-M-transect'. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……..……71

Figure 4: A temperature-salinity plot of stations showing the water masses sampled along the 'T-transect' (Plot by Gavin Tutt, Tesha Toolsee Marcel van den Berg, Oceans & Coasts Research at the Department of Forestry, Fisheries and the Environment). SASW = subantarctic Surface Water; AASW = Antarctic Surface Waters; AAIW = Antarctic Intermediate Water; CDW = Circumpolar Deep Water; AABW = Antarctic Bottom Water……..……72

Figure 5: Boxplots showing nutrient concentrations at each depth for all samples for a) nitrate, b) nitrite, c) nitrate + nitrite measured as combined nitrate and nitrite, d) ammonium, e) phosphate and f) silicate. Samples from the 'Crossroads line' were excluded since there were no nutrient concentration measurements for those samples. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……..74

Figure 6: Boxplots showing nutrient concentrations at each position (open ocean and inter-island) for all samples for a) nitrate, b) nitrite, c) nitrate + nitrite measured as combined nitrate and nitrite, d) ammonium, e) phosphate and f) silicate. Samples from the 'Crossroads line' were excluded since there were no nutrient concentration measurements for those samples. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……76

Figure 7: Bar graphs showing nutrient concentrations at each station for all samples for a) nitrate + nitrite measured as combined nitrate and nitrite, b) ammonium, c) phosphate and d) silicate. 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs. Samples from the 'Crossroads line' were excluded since there were no nutrient concentration measurements for those samples……..…78

Figure 8: Boxplots showing potential extracellular enzyme activity at each depth for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……...80

Figure 9: Boxplots showing potential extracellular enzyme activities at each position (open ocean and inter-island) for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……81

Figure 10: Bar graphs showing potential extracellular enzyme activities at each station for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs. Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase…..……82

Figure 11: The phylogenetic placement of 58 reconstructed metagenome-assembled Archaeal genomes. A maximum-likelihood phylogenomic tree was built based on concatenated amino acid sequences GToTree. The scale represents one amino acid substitution per sequence position…….....85

Figure 12: The phylogenetic placement of 170 reconstructed metagenome-assembled Bacterial genomes. A maximum-likelihood phylogenomic tree was built based on concatenated amino acid sequences GToTree. The scale represents one amino acid substitution per sequence position……..……..86

Figure 13: Boxplots of the number of paralogs according to a) depth and b) position (open ocean or inter-island). Two outliers from 5 m in the open ocean were removed. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively……..88

Figure 14: Boxplot showing the number of paralogs for all samples for each phylum. One Cyanobacterium and one Proteobacteria were removed because they were outliers……..……89

Figure 15: Bar graph showing the number of paralogs for all samples for each depth. 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs. One sample from S01 and S02 were removed because they were outliers……89

Figure 16: Bar graph showing the number of paralogs for all samples for each extracellular enzyme sampled for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs…….91

Figure 17: Heatmap showing putative paralog functions of all 228 MAGs for each phylum for each depth. Pathways are separated into carbon, nitrogen and phosphorus, as well as methanogenesis and methane oxidation. Paralogs responsible for β-glucosidase (BG) in Bacteroidota from deep samples were excluded since they are outliers. The scale represents the completeness of each pathway from 0.0 being incomplete to 1.0 being complete. 'MO' = methane oxidation; 'PT' = phosphorus transport……..…..92

Figure 18: Bar graph of the pangenome of the 228 metagenome-assembled genomes from the Subantarctic Indian Ocean. Core genes = genes found in 99% < = strains < = 100; soft core genes = genes found in 95% < = strains < 99%; accessory genes = genes found in 95% > strains…….……...94

List of Tables

Table 1: Examples of studies that were conducted on gene duplications in different organisms, as well as the authors and the references of each study……14

Table 2: The coordinates of the stations that were sampled, as well as the sample type (CTD or underway) and transect the station belongs to. S = S-transect; CR = Crossroads transect; T = T-transect; X-M = X-M-transect……49

Table 3: The specific substrates, concentrations of substrates and incubation times used for each of the extracellular enzyme activity assays……53

Table 4: Kruskal-Wallis non-parametric test of significance for each physical factor for each depth and station, as well as Wilcoxon test of significance for each physical factor for each position. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.....73

Table 5: Kruskal-Wallis non-parametric test of significance was applied to test the differences linked to each nutrient at each depth and station. In addition, the Wilcoxon test for significance was also applied to test differences in each nutrient at each position. The significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.……..……..……..……75

Table 6: Kruskal-Wallis non-parametric test of significance for each extracellular enzyme activity assay for each depth and station, as well as Wilcoxon test of significance for each extracellular enzyme activity assay for each position. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001…….....83

Table 7: Kruskal-Wallis non-parametric test of significance for the number of paralogs for each depth, phylum and station. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001……..……..90

Table 8: Kruskal-Wallis non-parametric test of significance for the number of paralogs for each extracellular enzyme assayed for each station. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001…..…..93

Table 9: The pangenome of the 228 metagenome-assembled genomes from the Subantarctic Indian Ocean…….……94

List of Supplementary Data

Figure S1: A scatterplot of the number of MAGs belonging to each of the 11 phyla against the total number of paralogs observed in each phyla, including the correlation statistics of the two non-parametric variables. *S* = the observed value of S statistics; *p* = significance; $\hat{\rho}_{Spearman}$ = Spearman's rank correlation coefficient; $CI_{95\%}$ = 95% confidence interval; n_{pairs} = number of observations……132

Figure S2: Scatterplots of the number paralogs assigned to a) BG, b) BX, c) NAG and d) AP against their respective activity (nmoles/hr/ml), including the correlation statistics of the two non-parametric variables. *S* = the observed value of S statistics; *p* = significance; $\hat{\rho}_{Spearman}$ = Spearman's rank correlation coefficient; $CI_{95\%}$ = 95% confidence interval; n_{pairs} = number of observations……..…133

Table S1: Quality and depths of the 228 high and medium quality MAGs from the SIO. Fmax = Fluorescence maximum; OMZ = Oxygen minimum zone; Deep = Deep; 5 m = Surface. 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs.

Table S2: Physical factors for all sample sites at each depth, as well as the water masses each sample belongs to. Fmax = Fluorescence maximum; OMZ = Oxygen minimum zone; Deep = Deep; 5 m = Surface; N/A: non-applicable (measurements were not taken for the site). 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents

samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs. SASW = subantarctic Surface Water; AASW = Antarctic Surface Waters; AAIW = Antarctic Intermediate Water; CDW = Circumpolar Deep Water; AABW = Antarctic Bottom Water.

Table S3: Concentrations of nutrients (µmol/L) for all sample sites at each depth. Fmax = Fluorescence maximum; OMZ = Oxygen minimum zone; Deep = Deep; 5 m = Surface; BD = below detection limit; N/A = non-applicable (measurements were not taken for the site). 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs.

Table S4: Potential extracellular enzyme activities (nmoles/hr/ml) for all sample sites at each depth. Fmax = Fluorescence maximum; OMZ = Oxygen minimum zone; Deep = Deep; 5 m = Surface. 'S' and 'CR' represent samples taken from the underway system between South Africa and Marion Island, while 'T' represents samples from the 'T-transect' and 'X' and 'MT' represent samples from the 'X-M transect' taken around the PEIs. Outliers were excluded for N-acetylglucosaminidase and alkaline phosphatase.

Table S5: Classification of each of the 228 metagenome-assembled genomes using GTDB-tk.

Table S6: Paralog pathway completeness of carbon, nitrogen, phosphorous and sulphur cycles, as well as methanogenesis and methane oxidation.

List of abbreviations

Abbreviation	Definition
3HP/4HB	4-hydroxybutyrate/ 3-hydroxypropionate
4-MUB	4-methylumbelliferone
5 m	Surface
AABW	Antarctic Bottom Water
AAIW	Antarctic Intermediate Water
AASW	Antarctic Surface Waters
ACC	Antarctic Circumpolar Current
AOA	Ammonia oxidizing archaea
AP	Alkaline phosphatase
APFZ	Antarctic Polar Frontal Zone
AZ	Antarctic zone
BG	β-glucosidase
BX	β-xylosidase
CAZ	Ceftazidime
CBB	Calvin–Benson–Bassham
CDW	Circumpolar Deep Water
CO_2	Carbon dioxide

CTD	Conductivity, temperature and depth
DOC	Dissolved organic carbon
DOM	Dissolved organic matter
DON	Dissolved organic nitrogen
Fmax	Fluorescence maximum layer
FL	Free-living
HGT	Horizontal gene transfer
IME	Island Mass Effect
IS	Insertion sequence
LGT	Lateral gene transfer
MAGs	Metagenome-assembled genomes
MESO	Mesopelagic zone
MGII	Marine Group II
ML	Epipelagic mixing layer
MO	Methane oxidation
MPA	Marine protected area
NAG	N-acetylglucosaminidase or chitinase
N_2	Dinitrogen gas

NO_2^-	Nitrite
NO_3^-	Nitrate
$NO_3^- + NO_2^-$	Nitrate + nitrite
NH_4^+	Ammonium
OMZ	Oxygen minimum zone
PEIs	Prince Edward Islands
Pi	Inorganic phosphate
POC	Particulate organic carbon
PO_4^{3-}	Phosphate
PT	Phosphorus transport
rTCA	Reductive Tricarboxylic Acid Cycle
RuBisCO	Ribulose-1,5-bisphosphate carboxylase-oxygenase
SASW	Subantarctic Surface Water
SAGs	Single-cell assembled genomes
SIO	Subantarctic Indian Ocean
SiO_4^{4-}	Silicate
SO	Southern Ocean
TCA	Tricarboxylic acid cycle

Abstract

The Southern Ocean represents an important carbon sink, and recent studies suggest that microbial communities contribute substantially to ecosystem functions via the biological pump. There is some evidence that microbial communities within the vicinity of the Prince Edward Islands (PEIs) are shaped by factors such as depth and the availability of nutrients. However, we lack a clear mechanistic understanding regarding the interplay between these environmental variables and their effects on microbial communities. Gene duplication is a known source of genetic variation in both macrofauna and microorganisms. Duplication events contribute to the evolution of genomes by providing additional genetic material for natural selection or neutral genetic drift to act on, causing evolution at a molecular level. Previous studies on suggest that gene duplications may be mechanisms used by microbial communities to adapt to extreme conditions. However, few studies have investigated gene duplications in marine microbiota, and we lack empirical insights regarding evolutionary dynamics.

Little is known about the role gene duplication events play in biogeochemical cycling. There are no reports from the Southern Ocean focused on paralogs responsible for these cycles in critical locations such as the PEIs and surrounding waters. This thesis aims to examine the extent and potential function of gene duplicates in marine microbial communities from the subantarctic Indian Ocean (SIO), specifically in waters surrounding the PEIs. We lack insights regarding the factors that influence microbial communities in these locations, and even less regarding the factors influencing gene duplications in these environments. The extent and putative functions of gene duplications in these microbial communities was evaluated, using shotgun metagenomic approaches. These findings

were correlated with extracellular enzymatic activity assays and nutrient analyses to determine the potential function of the paralogs, as well as the factors which determine their relative abundance.

The results revealed several gene duplication events within the PEIs marine microbial communities. Duplications varied significantly (statistical test and P value) across eleven bacterial and archaeal phyla, where Proteobacteria, Actinobacteria and Verrucomicrobiota genomes showed the highest levels of duplications. Paralogs with the genomic potential for carbon, nitrogen, and phosphorous metabolism, as well as methanogenesis were observed in our data. Previous studies of the PEIs and surrounding waters have revealed the remarkable capacity for biogeochemical cycling in the region, and several mechanisms linked to the cycling of carbon and nitrogen. Our data suggests that gene paralogs may play an important role in mediating these cycles, as well as in the cycling of phosphorous in the SIO.

This study provides the first insights regarding the extent of gene duplications in SIO microbial communities. The data provide insights regarding the factors which potentially influence the rate of gene duplication events in these environments. For the first time, we provide a clear understanding regarding the potential function of the gene paralogs observed in these microbial communities and their putative roles.

Chapter 1

Subantarctic Indian Ocean microbial gene duplications and their fates

1. Introduction

Gene duplication, the production of a copy of a single gene in an organism's genome, acts as a source of genetic innovation for the evolution of genomes (Álvarez-Lugo and Becerra 2021; Copley 2020; Gevers et al. 2004; Hooper and Berg 2003). Gene duplication may occur in response to changing conditions, where different selective pressures are experienced (Bergthorsson et al. 2007; Copley 2020; Kondrashov 2012), such as heat adaptation in *E. coli* (Riehle et al. 2001). These duplicated genes may diverge, developing novel functions, which allow an organism to adapt to its environment by expressing a new phenotype (Hooper and Berg 2003; Jordan et al. 2001). Duplication events play an important role in the evolution of secondary metabolism, where duplicates create the opportunity to produce more complex chemical analogues from pre-existing metabolites (Fewer and Metsä-Ketelä 2020; Moghe and Last 2015; Tawfik and Khersonsky 2010). These novel gene products may either have an enhanced or novel biological activity which provides the organism with a selective advantage, facilitating the adaptation to environmental stressor (Fewer and Metsä-Ketelä 2020). Gene duplicates may increase an organism's metabolic efficiency (Copley 2020; Tawfik and Khersonsky 2010) and/or gene dosage of beneficial enzymes (Conant and Wolfe 2008). Gene duplicates also play a role in microbial diversity, by providing genetic variation within and between species (Virdi and Sachdeva 2005). In this way, gene copies may be responsible for large changes in an organism and, in some cases, result in sufficient genetic diversity to allow for the emergence of entirely new lineages (Serres et al. 2009; Virdi and Sachdeva 2005).

Studies on gene duplications tend to focus on eukaryotes, such as humans (Li et al. 2001), zebrafish (Postlethwait et al. 2000), *Arabidopsis thaliana* (The *Arabidopsis* Genome 2000) and other land plants (Gómez-Anduro et al. 2011; Guo et al. 2016; Magadum et al. 2013; Rensing 2014). Although several studies have reported gene duplication events in prokaryotes (Gevers et al. 2004), most have focused on land microbes or pathogens (Jordan et al. 2001; Virdi and Sachdeva 2005). The few studies on gene duplications in marine environments tend to focus on individual species, such as Thaumarchaeota (Sheridan et al. 2020) and *Planctomycetes* (Kim et al. 2016), rather than spatial or depth trends of gene duplications. Gene duplications may constitute a large proportion of bacterial genomes, which could result in gene duplications contributing to bacterial evolution (Gevers et al. 2004).

Marine microbes have specialised genomes, which have evolved through processes such as gene duplication (Serres et al. 2009), horizontal gene transfer (HGT), gene loss, gene mutations, genome rearrangements (Fraser-Liggett 2005) and polyploidy (Jones and Baxter 2017). Gene duplication occurs when one gene gives rise to two genes and both cannot be operationally distinguished (Magadum et al. 2013). Horizontal gene transfer is the transfer of genetic material from one strain or species to another (Baltrus 2013). Gene loss is the deletion of a gene through genetic instability caused by the insertion of insertion sequence (IS) elements or other mobile DNA within a gene (Moran and Mira 2001). Gene mutations are any alteration in the nucleotide sequence of a genome, regardless of phenotypic change (Arber 2000). Genome rearrangements are a set of mutations that change the order of the gene content in a genome (Hartmann et al. 2018). Polyploidy is possessing multiple sets of chromosomes as a result of an organism's whole genome

being duplicated (Van de Peer et al. 2021). Marine microbes can have a high genetic diversity due to their high turnover, environmental complexity and large biomass (Whitman et al. 1998). These organisms make up communities that carry out important ecosystem functions within the ocean that include carbon sequestration and the cycling of elements such as iron, nitrogen, sulphur and silica (Glöckner et al. 2012). Marine microbes are particularly important in the sequestration of atmospheric carbon, where approximately half of global carbon sequestration is ascribed to microbes (Worden et al. 2015). A large proportion of carbon sequestration is carried out in the South Indian Ocean (SIO) (Glöckner et al. 2012; Stirnimann et al. 2021).

The SIO is an important carbon sink, despite limited marine microbial growth due to the generally low iron and trace element concentrations (DeVries 2014; Glöckner et al. 2012; Sunda and Huntsman 1997). The Subantarctic Ocean, including the Subantarctic (SFZ) and Polar Frontal Zones (PFZ), is the largest carbon sink south of 30° latitude (DeVries 2014). Microbial communities here are subject to extreme environmental conditions including variation in salinity and depth, and thus pressure, temperature, and the availability of both oxygen and light (Phoma et al. 2018). Phoma and colleagues (2018) suggest that these environmental factors may be responsible for 35% of the community variation. However, the SIO is largely understudied in terms of microbial community structure and function (Hermes et al. 2019; Phoma and Makhalanyane 2021; Phoma et al. 2018). Given the importance of microbes as mediators of biogeochemical cycles, it is important to study the SIO and to elucidate the mechanisms used for survival by the microorganisms found in this extreme environment (Arrigo 2005; Falkowski et al. 2008; Karl 2007; Phoma et al. 2018). These mechanisms may include, for instance, gene

duplication events (Bratlie et al. 2010; Gevers et al. 2004; Kondrashov 2012). Gene duplications may lead to the formation of novel proteins or metabolites (Fewer and Metsä-Ketelä 2020), genetic diversity (Virdi and Sachdeva 2005) and an increased gene dosage for dosage sensitive proteins (Moghe and Last 2015). These duplications may be particularly useful in facilitating the adaptation of microorganisms to extreme stressors in environments such as the SIO.

Here, I will review the extent of gene duplications in microorganisms, specifically marine microbes. I will critically assess the role of these duplicated genes on the survival and adaptation of these species to the extreme environments.

2. Gene duplication

Early evidence suggests that gene duplications provide genetic variation (Anderson and Roth 1977; Hughes 1994; Ohno 1970) for natural selection or neutral genetic drift to act on (Copley 2020; Lauer and Gresham 2019; Magadum et al. 2013). Gene duplication can be defined as the production of one or more copies of segment of DNA, from a few base pairs to whole chromosomes (Bell 2023). This increased genetic variation is thought to result in evolution at a molecular level (Lauer and Gresham 2019; Serres et al. 2009). Certain duplicated gene classes are preferentially retained (Davis and Petrov 2004; Gevers et al. 2004). Gevers et al. (2004) analysed 106 bacterial genomes and found that genes involved in transcription, defence mechanisms, energy production and energy conversion are preferentially duplicated. The study found ABC-type transporters, transcriptional regulators and dehydrogenases are gene families that contain the most paralogs, showing evidence of preferential duplication (Gevers et al. 2004). Gene duplicates that belong to functional classes that aid an organism in surviving in a specific

environment may also be preferentially retained (Bratlie et al. 2010; Zhang et al. 2016). This was seen in the paranome of mycobacteria, where several gene duplicates responsible for lipid transport and metabolism were present (Gevers et al. 2004). These are just a few of the roles that gene duplicates play in the evolution of microbial genomes, others include increasing metabolic efficiency (Copley 2020; Tawfik and Khersonsky 2010), aiding in microbial diversity (Serres et al. 2009; Virdi and Sachdeva 2005), providing antibiotic resistance (Chirakul et al. 2019) and niche specialisation (Kim et al. 2016).

2.1. The role of duplicated genes

Gene copies of enzymes may diverge in function from each other via mutations that cause a change in their specificity, as seen in β-ketoacyl-CoA synthase (KCS) in plants (Guo et al. 2016). Duplicates may also diverge in function from one another by the addition of another domain to one of the copies, as observed in the ribose repressor (RbsR) and the periplasmic protein for ribose transport (RbsB) genes in *E. coli* (Serres et al. 2009). Gene copies have the capacity to gain new functions, as long as one of the copies retains its original function, and this creates opportunities for metabolic diversification (Moghe and Last 2015). This additional genetic material can accumulate mutations that allow the gene copy to diverge in function from the original gene (Lynch and Conery 2000), leading to the development of microbial diversity.

Gene duplication plays an important role in facilitating diversity in microbial communities (Kamneva et al. 2012; Rivas et al. 2014; Serres et al. 2009). Duplication results in new gene families (Finet et al. 2013), and or new strains or species (Virdi and Sachdeva 2005). A clear example is demonstrated in studies on pathogenic microbes (Virdi and Sachdeva

2005). Gene duplication events may cause the expansion and divergence of protein families, specifically enzymes (Serres et al. 2009). These superfamilies include the Short Chain Dehydrogenase Reductase (SDR) family and the crotonase family, which have high amino acid conservation in the acyl-CoA-binding and the catalytic site (Serres et al. 2009). Small changes in the sequences of these duplicated enzyme genes resulted in the expansion of each pathway and biochemical capability of each organism. Duplication of the G protein of the respiratory syncytial virus (RSV) was implicated in the emergence of two new strains (BA and ON1) (Eshaghi et al. 2012; Parvez and Parveen 2017; Trento et al. 2003). The production of new gene families, strains or species through gene duplication events is dependent on the fate of the gene duplicates, specifically whether the duplicated gene gains a new function or not.

2.2. The fate of gene duplicates

Gene duplicates may undergo many different fates, such as nonfunctionalization, neofunctionalization or subfunctionalization (Force et al. 1999; Lynch and Conery 2000; Prince and Pickett 2002), then the gene duplicates can become fixed or lost within a population (Figure 1). Nonfunctionalization is when the gene copy becomes silenced by degenerative mutations (Force et al. 1999; Lynch and Conery 2000), such as point or frameshift mutations (Mira et al. 2001), or degenerates to a pseudogene (Prince and Pickett 2002). Lynch and Conery (2000) analysed the fate of duplicated genes in species from different kingdoms and found that most of the gene duplicates were silenced during nonfunctionalization rather than becoming fixed. The authors suggested that gene duplicates that become fixed play a role in gene dosage in an organism with high gene

dosage requirements. On the other hand, neofunctionalization and subfunctionalization mutations are assumed to be rare (Prince and Pickett 2002).

Gene duplicates are able to avoid deletion or redundancy through neofunctionalization and subfunctionalization (Figure 1) (Force et al. 1999; Lynch and Conery 2000; Ohno 1970; Prince and Pickett 2002). Neofunctionalization is when a gene copy acquires a novel function that is beneficial via alterations in the coding or regulatory sequences of the gene copy (Prince and Pickett 2002). The gene copy with the new function can be preserved by natural selection (Force et al. 1999; Lynch and Conery 2000) or random genetic drift (Mira et al. 2001). Subfunctionalization is where both gene copies lose part of their original function after acquiring complementary loss-of-function mutations, reducing their total combined capacity to that of a single copy of the ancestral gene (Force et al. 1999; Lynch and Conery 2000). Each copy is then subjected to a different selective pressure (Prince and Pickett 2002). There may be several gene duplication events which may occur within a genome depending on the rate of duplication, with gene duplicates experiencing different fates within the same genome.

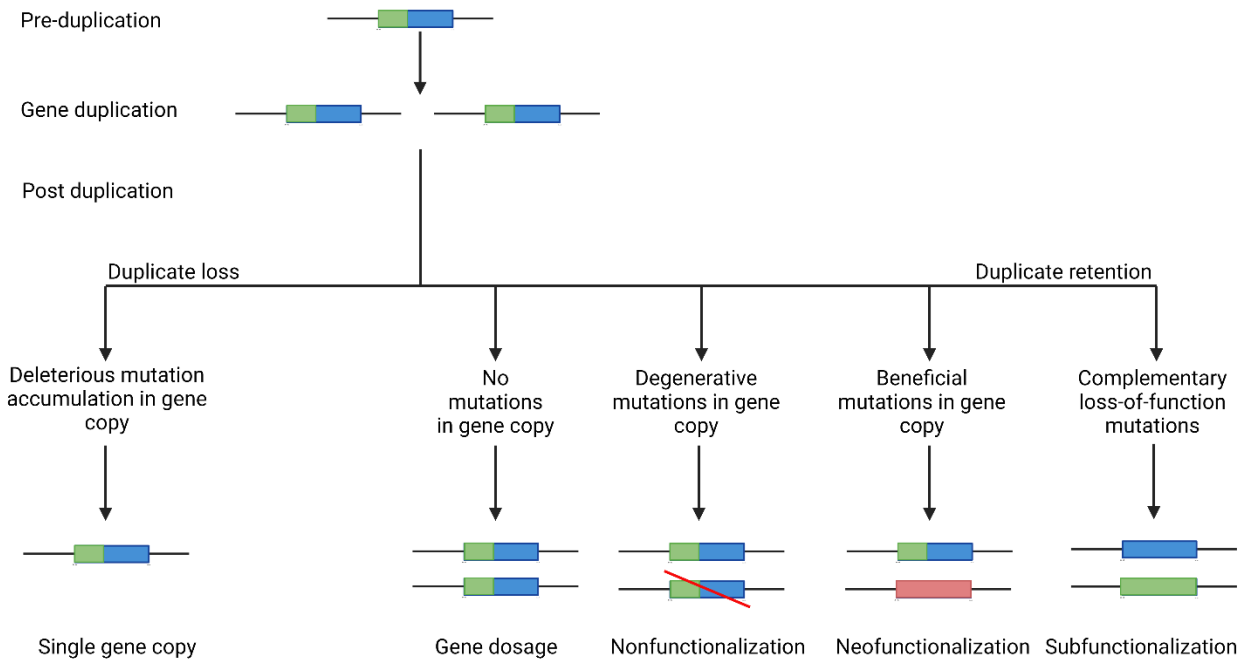

Figure 1: The fate of duplicated genes based on an image by Voordeckers and Verstrepen (2015). Most duplicates will be lost through the accumulation of deleterious mutations. Duplicates can be retained in subfunctionalization, neofunctionalization, nonfunctionalization or if they play a role in gene dosage.

2.3. The rate of gene duplication

The rate of gene duplication within a genome varies from species to species (Gevers et al. 2004; Li et al. 2001; Postlethwait et al. 2000; The *Arabidopsis* Genome 2000). The human genome is made up of 15% gene duplicates (Li et al. 2001), the zebrafish genome is made up of 20% gene duplicates (Postlethwait et al. 2000) and the genome of *Arabidopsis thaliana* is made up of 31% gene duplicates (The *Arabidopsis* Genome 2000). The proportion of gene duplicates may also vary between different strains of the same species, as observed in a study on *E. coli* that showed a 30% difference in the proportion of gene duplicates between the pathogenic *E. coli* O157:H7 and the non-

pathogenic laboratory strain *E. coli* K-12 (Blattner et al. 1997; Riehle et al. 2001). The number of gene duplicates seen in a genome is also dependant on the number of generations that have passed, where the rate of duplication is measured as duplications/gene/generation (Lipinski et al. 2011).

There are numerous studies on the rate of gene duplication events in bacteria that compare the rates between each species (Copley 2020; Gevers et al. 2004; Hooper and Berg 2003; Lynch and Conery 2000). The extent of gene duplication and its contribution towards adaptation has been compared amongst different bacteria in a study by Gevers et al. (2004). The authors searched 106 bacterial genomes for duplicated genes, also known as paralogs. The study found that all 106 genomes exhibited significant paralogy, where duplications accounted for an average of 23.5% ± 8.7% of a specie's genome. *Rickettsia conorii* exhibited the lowest paralogy (7% of its protein-coding genes) and *Streptomyces coelicolor* A3(2) exhibited the highest (41% of its protein-coding genes) (Gevers et al. 2004). From these results the authors confirmed that bacterial genome size is correlated with the number of paralogous genes it contains (Gevers et al. 2004). The extent of gene duplication, and its contribution towards adaptation, has been studied in individual taxa as well, such as in Thaumarchaeota (Sheridan et al. 2020), *Burkholderia pseudomallei* (Chirakul et al. 2019) and *Planctomycetes* (Kim et al. 2016).

2.4. Examples of gene duplications

Gene duplications are responsible for niche specialisation in Thaumarchaeota (Sheridan et al. 2020). The duplication of genes responsible for carbohydrate transport and coenzyme metabolism in Thaumarcheota were found to facilitate niche specialisation in this ammonia-oxidising archaeal phylum (Sheridan et al. 2020) (Table 1). The study used

metagenomic approaches to suggest that 77% of the gene gains, which occurred in the group of Thaumarchaeota, were a result of gene duplications. Whereas, 17% of duplications occurred via intra-LGT and 6% occurred via inter-LGT and *de novo* formation (Sheridan et al. 2020). Duplications were observed in gene families responsible for transcription, signal transduction, carbohydrate transport and coenzyme metabolism (Sheridan et al. 2020). In these instances, the copy numbers of the duplicated genes were more than 10-fold higher (Sheridan et al. 2020). Gene duplication events occur in other Kingdoms, such as bacteria (Chirakul et al. 2019), plants (Guo et al. 2016; The *Arabidopsis* Genome 2000) and animals (Li et al. 2001; Postlethwait et al. 2000).

There is evidence that gene duplications play a role in bacterial responses to antibiotic compounds from different environments, as seen in *Burkholderia pseudomallei* (Chirakul et al. 2019) (Table 1). A common mechanism of ceftazidime (CAZ) resistance used by the causative agent of melioidosis (White et al. 1989), *B. pseudomallei*, is the duplication of its *penA* gene that encodes PenA class A β-lactamase (Rholl et al. 2011; Viberg et al. 2017). Chirakul et al. (2019) analysed a *B. pseudomallei* isolate from a patient with clinically significant CAZ resistance, which was acquired through a gene duplication event. They found that PenA protein levels were significantly higher in the CAZ-resistant strain than CAZ-susceptible strains. The CAZ-susceptible strains were found to possess a single, unstable, copy of the *penA* gene (Chirakul et al. 2019). In contrast, the CAZ-resistant strain possessed 10-11 copies of the gene (Chirakul et al. 2019). The authors concluded that the presence of multiple copies of the *penA* gene, along with increased β-lactamase activity in the CAZ-resistant strain, was sufficient to provide clinically significant

CAZ resistance (Chirakul et al. 2019). Thus, gene duplication events provide clinically significant CAZ resistance in *B. pseudomallei.*

Table 1: Studies focused on gene duplications in different organisms, as well as the references of each study and main findings.

Authors (year)	Species	Main findings
Guo et al. (2016)	Plants	β-ketoacyl-CoA synthase (KCS) in plants was duplicated from a single copy in the common ancestor of all plants (Merchant et al. 2007; Prochnik et al. 2010) to up to 30 copies in angiosperms. Subfunctionalization and neofunctionalization via mutations are responsible for the divergence in catalytic specificity of KCS enzymes.
Riehle et al. (2001)	*Escherichia coli*	Adaptation to heat was acquired through the duplication of genes involved in heat-stress, such as the *rpoS*, *nlpD*, *pcm* and *surE* genes that were duplicated in three out of the six heat-adapted derived strains, as well as the duplication of the *poxB* gene that encodes a pyruvate oxidase.
Kim et al. (2016)	*Planctomycetes*	Duplication events in genes responsible for cell wall polysaccharide degradation and encoding sulfatase gene families are responsible for the survival of three undescribed *Planctomycetes* species on red algae. Polysaccharide degradation allows the bacteria to use the cell walls of the algae as a carbon source for their survival.
Chirakul et al. (2019)	*Burkholderia pseudomallei*	*B. pseudomallei* acquired resistance to ceftazidime via the duplication of the *penA* gene that encodes PenA class A β-lactamase. This duplication event provided the strain with increased β-lactamase activity, providing ceftazidime resistance.

2.5. Gene duplicates in marine microbes

Fewer studies have focused on the trends in gene duplication events in marine microbes, while the studies that do exist on the topic generally focus on duplications within a species or phylum (Kim et al. 2016; Sheridan et al. 2020). A study conducted on three undescribed *Planctomycetes* found on *Porphyra umbilicalis* (Blouin and Brawley 2012) found multiple gene duplications that could be responsible for the survival of the bacteria on the red

algae (Kim et al. 2016) (Table 1). All three genomes showed expansions of protein-coding gene families, such as GH10 xylanases comprising the largest protein-coding family expansion in only two of the three *Planctomycete* genomes. Multiple copies of polysaccharide degrading genes were found in the three unclassified genomes, including genes encoding κ-carrageenases the main enzyme responsible for carrageenan degradation in the cell walls of other red algae species (Kim et al. 2016). The three *Planctomycetes* also possessed copies of genes implicated in the degradation of fucans and alginates in brown algal cells, with one of the genomes possessing multiple copies of fucosidases, while the other two possessed multiple copies of alginate lyases (Kim et al. 2016). The variation of the number of gene copies, which are responsible for the degradation of different polysaccharides in the cell walls of different algae within the same area, suggests that the three *Planctomycetes* isolates are adapted to different niches via gene duplications (Bondoso et al. 2014). The results act as an example of how gene duplication allows a species to adapt to its environment (Kim et al. 2016).

3. Adaptation

Adaptation is a biological process where an organism changes its phenotype to circumvent and thrive in its environment, and to exploit new niches for survival (Hooper and Berg 2003). There is evidence that adaptations may be a result of new selective pressures, imposed by a new or changed environment, favouring nucleotide substitutions which generate novel gene functions if found in protein-coding regions of the gene (Irwin 1995; Irwin and Wilson 1990; Stewart et al. 1987). The small genome size and tight packing of the DNA in bacterial cells is an adaptation, which promotes replication efficiency and competitiveness (Andersson and Kurland 1998). There are many

mechanisms that have allowed organisms to adapt to their environments over their evolutionary histories (Fraser-Liggett 2005; Jones and Baxter 2017; Schönknecht et al. 2014). These methods include the horizontal gene transfer (HGT) of a gene from one species to another, the duplication of an already existing gene, the rearrangement of an organism's genome, gene loss (Fraser-Liggett 2005) and the duplication of an organism's entire genome in polyploidy (Jones and Baxter 2017).

3.1. Mechanisms of adaptation in microbial communities

Microbial genomes are dynamic, such that they are under the influence of many factors including gene loss, genome rearrangements, HGT, gene duplications (Fraser-Liggett 2005) and polyploidy (Jones and Baxter 2017). Vertical gene transfer influences an organisms adaptation that is passed from generation to generation and allows for the adaptation to sudden changes (Schönknecht et al. 2014). Gene loss, through genetic instability caused by the insertion of insertion sequence (IS) elements or other mobile DNA within a gene (Moran and Mira 2001), may cause an organism's phenotype to change. For instance, the increase in virulence in some *Listeria* species has been attributed to such gene loss (den Bakker et al. 2010). Horizontal gene transfer contributes to the evolution of microbial genomes by introducing new genes into a genome, which may be from another species or organism that is not the individual's parent (Koonin et al. 2001; Long 2001). This may result in rapid adaptations, the emergence of new strains and generate diversity amongst different strains (Boucher et al. 2003; Fraser-Liggett 2005). Horizontal gene transfer results in the acquisition of new functions and gene duplications result in the divergence of gene functions in some cases (Fraser-Liggett 2005). The potential to acquire or replace a gene in an organism with that from another

organism decreases as the phylogenetic distance between them increases (Ochman 2001). Polyploidy may result in a change in an organism's genome (Soltis et al. 2015; Van de Peer et al. 2017). An example of this is polyploidy providing photoprotection in halophilic archaea that experience high levels of UV radiation that can cause DNA lesions (Breuert et al. 2006; Jones and Baxter 2017; Ludt and Soppa 2018).

Gene duplications are a common evolutionary response to changing selection pressures in bacteria (Gevers et al. 2004; Kondrashov 2012). A specific example is when the alteration of gene expression is not sufficient to overcome the changing conditions (Kondrashov 2012; Riehle et al. 2001). This mechanism of adaptation is reversible, and therefore acts as a primitive regulatory mechanism in bacteria (Riehle et al. 2001). Gene duplications thus contribute to the evolution of microbial genomes (Moghe and Last 2015; Virdi and Sachdeva 2005).

4. Evolution of microbial genomes

Microbes have evolved to adapt to extreme environments that range from cold deserts in Antarctica (Le et al. 2016) to deep-sea hydrothermal vents (Anderson et al. 2017). Bacterial genome sizes are relatively similar, with only up to a hundred-fold difference in genome size being observed compared to eukaryote genomes for example, that vary by five orders of magnitude(Bennett and Moran 2013; Casjens 1998; Han et al. 2013; Land et al. 2015). This disparity in genome size translates into a variation in organismal complexity, since a larger bacterial genome contains more genes than a smaller genome (Mira et al. 2001). Their genomes are able to evolve through different mechanisms, such as HGT (Selander and Levin 1980; Woese 2000) and gene duplications (Jordan et al. 2001). Both HGT and gene duplication can create intra-genome homologs, which are

homologs that arose within the genome (Gevers et al. 2004). These events usually only increase bacterial genome size if they confer a benefit to the organism, otherwise the new gene could be lost via deletional bias or genetic drift (Mira et al. 2001). Genes can then be lost via whole or multiple gene deletion events or by the accumulation of point or frame-shift mutations that inhibit the gene from functioning, causing it to become a pseudogene that can be eroded by small deletion events (Mira et al. 2001). Deletions have been shown to be more common than insertions, leading to deletional biases which reduce the genome size of bacteria (Mira et al. 2001). Genetic drift and deletion events cause a reduction in the size of bacterial genomes, while gene duplications and HGT events cause an increase in the size of bacterial genomes (Mira et al. 2001). It is important to understand gene duplications and HGT in an organism in order to understand the evolution of its genome (Gevers et al. 2004).

Horizontal gene transfer can transfer homologous as well as novel genes between species (Mira et al. 2001), which allows for the evolution of a specific lineage (Ochman 2001). Even one transfer event of a gene with a novel phenotype is sufficient to change a benign bacteria into a pathogenic bacteria (Groisman and Ochman 1994; Hacker and Kaper 2000). The transfer can be responsible for a large portion of an organism's genome, such as 16% of *E. coli* genome made up of HGT genes (Lawrence and Ochman 1998).

5. Microbial communities of the Subantarctic Indian Ocean

Microbial communities of the SIO can be broken up into communities belonging to different depths, including the surface water layer (Surface), the fluorescence maximum layer (F-max), the mesopelagic zone (MESO), the oxygen minimum zone (OMZ), the

epipelagic mixing layer (ML) and the deep layer (Pesant et al. 2015). The surface water layer is described as the layer 3 to 7 m below the surface (Pesant et al. 2015). The fluorescence maximum layer where the maximum abundance of plankton bearing chlorophyll pigments exists, or the layer containing plankton with higher chlorophyll content because of living in a dark environment (Pesant et al. 2015). The mesopelagic zone is described as being between 200 and 1000 m depths (Pesant et al. 2015), while the deep layer is 10 m above the seafloor (Phoma et al. 2018). The distribution of different phyla of marine microbes in the SIO are depth-dependant (Phoma et al. 2018).

Many factors affect the distribution and abundance of different genera across oceanic microbial communities, such as temperature, latitude (Swan et al. 2013) and depth (Phoma and Makhalanyane 2021; Sunagawa et al. 2015). Microbes living in the ocean have great genetic diversity due to their high turnover, environmental complexity and large biomass, with as much as 10^4 - 10^6 cells per milliliter seawater (Flemming and Wuertz 2019; Sunagawa et al. 2015; Whitman et al. 1998). Surface ocean microbial communities are characterized by abundant lineages such as Cyanobacteria and Alphaproteobacteria (Li et al. 2020; Pajares et al. 2020; Swan et al. 2013; Zhuang et al. 2018), as well as ubiquitous bacterial lineages such as Marine Group A, Verrucomicrobia, Actinobacteria, Bacteroidetes, and Proteobacteria (Swan et al. 2013). The distribution of these taxa is correlated to surface temperatures and global latitudes (Ibarbalz et al. 2019; Sunagawa et al. 2015; Swan et al. 2013). This affects their abundance, location relative to the coast and geographic distances between sampling stations. Microbial species' taxonomic and functional richness increases from the ocean surface to the mesopelagic zone (Sunagawa et al. 2015). Species richness in the also increases with an increase in

temperature towards the equator (Ibarbalz et al. 2019; Sunagawa et al. 2015). Microbial communities in the SIO are heterogeneous and spatially structured (Phoma et al. 2018).

The Subantarctic PEIs are located within the vicinity of Antarctic Circumpolar Current (ACC) in the Southern Ocean (Güller et al. 2020; Venkatachalam et al. 2019). These islands, located in the south-west Indian sector of the Southern Ocean, are approximately 2 000 km south of the east coast of South Africa (Belkin and Gordon 1996; Venkatachalam et al. 2019). The PEIs and the ACC are positioned within the Antarctic Polar Frontal Zone (APFZ), which links northern warm subtropical waters to southern cold Subantarctic waters (Ansorge and Lutjeharms 2002; Belkin and Gordon 1996; Venkatachalam et al. 2019). Due to legislative amendments and the goal of protecting the unique species, habitats and ecosystem processes biodiversity, the islands are regarded as core marine protected area (MPA) (Kirkman et al. 2021; Kirkman et al. 2016; Lombard et al. 2007). Previous studies have shown some evidence confirming that the region may facilitate monitoring and predictive capacity helping to advise future management on climate change in this important carbon sink, the SIO (Glöckner et al. 2012; Lombard et al. 2007).

The SIO is largely understudied, especially with regards to microbial community structure and function (Hermes et al. 2019; Phoma and Makhalanyane 2021; Phoma et al. 2018). The SIO represents an ecosystem where microbial communities experience extreme environmental conditions including variation in depth, salinity, and the availability of both oxygen and light (Bryant et al. 2016; Marambio et al. 2017; Phoma et al. 2018). The SIO is an important ecosystem because it possesses low concentrations of biologically available iron, making iron a limiting nutrient in this part of the ocean (Glöckner et al. 2012;

Sunda and Huntsman 1997). The Agulhas Current in the SIO runs from 20 km off the South African coast east to 25°S to 40°S, where the current conducts an inter-ocean exchange of heat, salt and biota between the Indian, Atlantic and Southern Oceans (Lutjeharms 2006). The environment is responsible for 35% of the marine microbial community variation in waters surrounding the PEIs (Phoma et al. 2018). These waters are thought to possess low species diversity, with high endemism indicating the adaptation of marine microbial species to the region (Ansorge et al. 2017). Dominant taxa include Proteobacteria, followed by Actinobacteria, Bacteroidetes, Firmicutes and Cyanobacteria (Li et al. 2020; Phoma et al. 2018). Surface microbial communities in the SIO experience short-term environmental selection, such as temperature and surface currents, while microbial communities in the deep layer experience extended effects such as water mass residence, reduced ventilation and slow moving connectivity through ocean circulation (Phoma et al. 2018).

The supply of nutrients strongly influences microbial communities of the SIO in terms of composition and primary production, which in turn affect biogeochemical cycles through nutrient cycling and carbon sequestration (Henley et al. 2020). Water masses formed in the SIO set the biochemistry of the global thermocline that supplies nutrients to surface waters, influencing primary production and carbon sequestration of oceans around the world (Moore et al. 2018; Moreau et al. 2015). Primary production in the high nutrient low chlorophyll Southern Ocean is limited by iron and silica concentrations due to their limited supply in summer, while primary production is limited by light in winter (Boyd et al. 1999; de Baar et al. 1990). Enzymatic activities and primary production are also limited by macronutrients that are supplied by the Circumpolar Deep Water (CDW) mass, such as

nitrate, phosphate and silicic acid (Moore et al. 2018). Primary production is already declining with increased anthropogenic climate change (Ansorge et al. 2017), possibly due to the seawater temperatures surrounding the PEIs increasing by 1.4°C since the 1950s (Rouault et al. 2005). Further impacts on ecosystem functioning have been predicted (Le Roux and McGeoch 2008; Smith 2002).

Another factor influencing lagoonal plankton communities in the PEIs marine microbiomes is the Island Mass Effect (IME) (Elliott et al. 2013; Stirnimann et al. 2021), which conserves these communities. The IME refers to the advantages of being in close proximity to an island mass, including the supply of nutrients from freshwater runoff from the PEIs that increase DOM levels (Boden 1988; Phoma and Makhalanyane 2021). This is due to the exchange and merge of the two fronts in the region, the Subantarctic and Antarctic fronts, exchanging characteristics and limiting the gradation between the fronts in terms of nutrient concentrations and physical properties (Boden 1988). This phenomenon can cause enhanced primary production around the islands due to physical and chemical properties that are the result of the proximity to the land masses. These factors include tidal flows and eddies that enhance mixing of surrounding waters (Perissinotto and Rae 1990; Soares et al. 2020), run-off from the island leaking nutrients into these waters and the exchange of nutrients (Smith and Froneman 2008), salinity and temperature with the lagoonal system near the islands.

6. Oceanic microbial community functions

Microbial communities play an important role in global geochemical cycles (Arrigo 2005; Falkowski et al. 2008; Pesant et al. 2015), making them an essential part of their ecosystems and in ecosystems dependent on these nutrients. It is thus important to study

these oceanic microbial ecosystems to further understand their and the ocean's role in this biogeochemical cycling. For this reason, there have been numerous expeditions around the world's oceans (Ansorge et al. 2017; Kemeny et al. 2016), such as the Tara Oceans expedition that sampled 210 ecosystems in 20 biogeographic provinces (Pesant et al. 2015). Microbial communities belonging to many diverse environments, including extreme environments like rock-hosted subseafloor crustal aquifers (Tully et al. 2018), have been shown to play a role in the global marine geochemical cycles (Glöckner et al. 2012; Mitra et al. 2014; Worden et al. 2015; Zehr and Kudela 2011).

Taxa with known chemolithoautotrophic lifestyles, such as carbon fixing and sulphur-oxidizing Oceanospirillales, Thaumarchaeota and anammox Planctomycetes, were identified in the SIO (Phoma et al. 2018; Swan et al. 2011; Venkatachalam et al. 2019). Chemolithoautotrophs use carbon dioxide in the deep ocean as a carbon source for metabolic processes, as well as inorganic compounds such as nitrogen and sulphur as energy sources for these metabolic processes (Ver Eecke et al. 2012; Zeng et al. 2021). Phoma and colleagues (2018) suggested Gammaproteobacteria as a class that is responsible for carbon metabolism in the SIO(Zeng et al. 2021). These chemolithoautotrophic Gammaproteobacteria were found to use inorganic compounds, such as sulphur or nitrate compounds, as an energy source for the fixation of oceanic carbon dioxide (Glaubitz et al. 2013; Vasquez-Cardenas et al. 2015; Zehr and Kudela 2011). Gammaproteobacteria dominate the nutrient rich marine environment in the SIO, suggesting that this class is adapted to this specialised region (Phoma et al. 2018).

Microbial communities are important because they make up the majority of the biomass in the ocean, contribute significantly to global biogeochemical cycles (Arrigo 2005;

Falkowski et al. 2008; Graham et al. 2016; Karl 2007) and are a crucial part of marine food webs (Del Giorgio and Duarte 2002). Marine microbial communities can consist of four different groups, including bacteria and archaea (prokaryotes), single-celled eukaryotes and viruses (Fraser-Liggett 2005). These communities play a crucial role in global carbon dioxide sequestration (Arrigo 2005; Smith Jr et al. 2009), reducing the effect of human activities on global climate change (Glöckner et al. 2012). Marine microbes are responsible for half of global primary production, where carbon dioxide and water are used to produce chemical energy stored in organic compounds (Henley et al. 2020; Worden et al. 2015). These organic compounds are used by individuals in higher trophic levels, thus marine microbes contribute to the marine food chain (Halo et al. 2016). Organic compounds that are not used up by higher organisms sink to the ocean floor where they are stored for thousands of years (Halo et al. 2016; Volk and Hoffert 1985). This makes up the 'biological pump' (Volk and Hoffert 1985). Microbial communities belonging to many diverse environments, including extreme environments such as rock-hosted subseafloor crustal aquifers (Tully et al. 2018), have been shown to play a role in the global marine geochemical cycles (Falkowski et al. 2008; Pérez Castro et al. 2021; Smith Jr et al. 2009). Carbon, nitrogen, sulfur, phosphorous, iron and silica cycling is carried out by bacterioplankton in the ocean (Glöckner et al. 2012; Montoya et al. 2004; Swan et al. 2011; Zhang et al. 2015). These biogeochemical cycling processes are linked, where the end product of one cycle is the used in the beginning of another cycle (Falkowski et al. 2008; Li et al. 2018; Ver Eecke et al. 2012).

6.1. Nitrogen cycling

Nitrogen is an essential limiting factor in the growth of marine microorganisms and higher organisms as it is present in a limited supply in the ocean column as different forms that are not all biologically available (Elser et al. 2007; Moore et al. 2013; Zehr and Kudela 2011). Because nitrogen transformation is carried out by microorganisms in the ocean, it is important to understand these processes on a gene-level and to understand the biological capabilities of marine microorganisms (Zehr and Kudela 2011; Zhang et al. 2020). Nitrogen is typically found in the form of dissolved dinitrogen gas and dissolved organic nitrogen (DON) in the surface ocean (Voss et al. 2013; Zhang et al. 2020). Dinitrogen is available to a small set of Archaea and Bacteria that convert it into biologically available ammonium (Kuypers et al. 2018; Zehr and Kudela 2011; Zhang et al. 2020). Nitrogen is also found in oxidized forms, such as nitrate, in coastal waters where deep water is brought to the surface or where riverine inputs are present (Zehr and Kudela 2011; Zhang et al. 2020). Nitrogen-containing dissolved organic matter is largely understudied in the ocean (McCarthy and Bronk 2008), however it is present in the form of urea and amino acids that can be readily assimilated by auto- and heterotrophic organisms (Baker et al. 2009).

6.2. Carbon cycling

Carbon dioxide (CO_2) fixation in the ocean is performed by diverse groups of photosynthetic eukaryotes (protists) and cyanobacteria (Mitra et al. 2014; Worden et al. 2015). Carbon fixation is carried out by marine microorganisms (Fischer et al. 2016; Li et al. 2018), where dissolved carbon dioxide (CO_2) from the atmosphere is converted into oxygen and products that can be used by other microorganisms such as particulate organic carbon (POC) and dissolved organic carbon (DOC) (Falkowski et al. 1998;

Glöckner et al. 2012). Energy for carbon fixation is mostly provided by sunlight, but can come from chemical compounds in places where sunlight is limited, such as the OMZ or hydrothermal vent systems (Fischer et al. 2016). Under anaerobic conditions CO_2 is converted to methane which is converted back into CO_2 via methanotrophic bacteria (Orcutt et al. 2011; Zeng et al. 2021). Cyanobacteria fix CO_2 into organic carbon which is then taken up by grazers and predators, to be released in fecal pellets as particulate or dissolved organics (Perissinotto et al. 2007; Soares et al. 2020). Dissolved organics are also released during viral lysis and the death of microorganisms (Glöckner et al. 2012; Worden et al. 2015). Organic matter that is not used up by microorganisms in the upper surface layers sinks to the sea floor (Capone and Hutchins 2013; Johnson et al. 2020), where it is buried and converted to natural gas, petroleum, kerogen and bitumen over geological timescales (Glöckner et al. 2012). Scientists are able to monitor and study genes responsible for carrying out these biogeochemical nutrient cycles looking at the genomes of marine microbial communities (Fraser-Liggett 2005).

7. Techniques used to study microbial genomes and gene duplications

Many oceanic prokaryotes are difficult to culture (Joint et al. 2010; López-Pérez et al. 2020), emphasizing the importance of genomic approaches. Sequencing the full-length 16S rRNA gene sequences show the community composition of a sample, including relative abundances as well as the specific species that make up that community (Bonk et al. 2018; Tully et al. 2018). However, this data does not reveal insights into the biology or interactions between species in a community (Fraser-Liggett 2005). Metabarcoding is a technique that allows researchers to assess the community composition of an environmental sample, where a Polymerase Chain Reaction (PCR) is used to amplify

DNA extracted from an environmental sample and specific targets are sequenced according to the target kingdom (Djurhuus et al. 2017; Francioli et al. 2021). This approach allows researchers to look at organisms from different trophic levels within the sample, allowing them to study biodiversity over time and space (Cermeño et al. 2010). Another technique that can be used to study individual microbes is the generation of single-cell assembled genomes (SAGs) (Cahn and Piel 2021). This approach has several disadvantages that include a DNA amplification step could result in amplification biases and cells could be difficult to separate if they occur in biofilms (Clingenpeel et al. 2015), resulting in fragmented and incomplete assemblies (Stepanauskas 2012). Metagenomics is an important genomics-based tool that allows researchers to understand microbial biology, because it allows them to study complex microbial communities and to probe these same communities (Fraser-Liggett 2005; Grossart et al. 2020).

7.1 Metagenomics

Metagenomic approaches allow researchers to characterize the microbial community within an environmental sample, as well as to examine the functional potential of the community members (Kieser et al. 2020; Shakya et al. 2019). Metagenomic studies give insights into the way microbial communities are shaped and how these communities are able to adapt to their environments (Cowan et al. 2015), which can include extreme environments such as rock-hosted subseafloor crustal aquifers (Tully et al. 2018), hypolithic hot (Namib Desert), cold (Antarctica) and arid deserts (Le et al. 2016), as well as subseafloor deep-sea hydrothermal vents (Anderson et al. 2017). It is a powerful tool that allows researchers to make comparisons between different samples from the same or different studies (Nayfach and Pollard 2016). To date, more than 513,120 full publicly

available metagenomes (Metagenomics RAST server; http://metagenomics.anl.gov; as of April 2023). Metagenome-assembled genomes (MAGs) are assembled from metagenomes and allow organism-specific analyses from metagenomes (Kieser et al. 2020).

The construction of complete or near-complete MAGs from microbial taxa that make up a specific community allows unique metabolic features to be assigned to a specific species beyond that of taxonomic prediction (Frioux et al. 2020; Podell et al. 2020). An advantage of metagenomic approaches is the small quantity of sample needed, where a few grams of water is sufficient to generate gigabases of nucleic acid sequences (Ansorge et al. 2017). This is specifically useful when bioprospecting microbial communities for bioactive secondary metabolites, bioactive pathways and enzymes that possess medical or industrial significance (Ansorge et al. 2017; Wani et al. 2022). Assembly-based metagenome pipelines are used to assemble metagenomes from a sample (Kieser et al. 2020; Norouzi-Beirami et al. 2021). These pipelines are advantageous when a sample contains uncultivated species, because it does not rely on a reference genome to map the raw reads against (Kieser et al. 2020). Despite the abovementioned advantages of metagenomic approaches, there are also some limitations to the method. Metagenomics can be a limited approach since MAGs are not complete genomes, but partial genomes recovered from their respective environments (Anderson et al. 2017; Chen et al. 2020). This could lead one to believe that there are no gene duplicates of a certain gene, when the recovered genome might not contain that specific duplication (Anderson et al. 2017).

7.2 Studying marine microbiomes with metagenomic approaches

Metagenomic studies have been used to study the patterns of genomic variation within populations, such as in the microbial communities that live in the subseafloor of deep-sea hydrothermal vents (Anderson et al. 2017). The aim of the study was to use metagenomes to infer the evolutionary histories of the two vent populations. The authors found between the two vent populations, one possessed less non-synonymous mutations and hence was under stronger purifying selection than the other population. The other population contained more unique genes in the cellular membrane, hence the authors concluded the two populations experienced different evolutionary pressures. This study showed how metagenomic approaches can be used to study microbial population diversity and differentiation, and how the evolutionary history of microbial communities can be inferred from genomic studies (Anderson et al. 2017; Fadiji and Babalola 2020).

8. Pangenomes

The genome of a species can be divided into two parts: the core and accessory genomes (Fraser-Liggett 2005; McCarthy and Fitzpatrick 2019). The core genome is the set of genes that is shared by every strain, while the accessory genome is the rest of the genes found in the different isolates of the species (McInerney et al. 2017; Steinberg et al. 2022). The accessory genome includes dispensable genes found in some but not all isolates as well as strain-specific genes found in a single strain of the species (Fraser-Liggett 2005; McCarthy and Fitzpatrick 2019; McInerney et al. 2017). Bacterial genomes are made up primarily of core sequences with a consistent G+C content and codon usage among them (Hacker and Kaper 2000). These core sequences carry housekeeping genes, which are responsible for the basic functioning of bacteria including ribosomal-RNA and metabolic

protein encoding genes (Hacker and Kaper 2000; Steinberg et al. 2022). The rest of bacterial genomes are made up of genes acquired through processes such as HGT and gene duplications (Brockhurst et al. 2019; Pang and Lercher 2019). Accessory sequences differ in their codon usage and G+C content, and have been shown to code accessory functions such as antibiotic resistance in *E. coli* K-12 (Lawrence and Ochman 1998). The accessory genome is responsible for genetic diversity amongst different strains of a species as well as allowing different species to adapt to and utilise different environments (Brockhurst et al. 2019; Fraser-Liggett 2005).

An example of a pangenomic analysis is the study on 31 *Prochlorococcus* isolate genomes and 74 SAGs to determine the underlying causes of niche specialization by Delmont and Eren (2018). *Prochlorococcus* is a photosynthetic genus of bacteria that is abundant in euphotic zones of low latitude marine systems (Chisholm et al. 1988; Rusch et al. 2010). The genus plays a large role in the fixation of carbon from the atmosphere (Flombaum et al. 2013). Delmont and Eren (2018) linked metagenomic studies of surface marine microbes (Acinas et al. 2021) with a pangenomic analysis of *Prochlorococcus* genomes to infer associations between individual gene clusters and their potential role in niche specialization and fitness. The *Prochlorococcus c*ore genome included 766 genes relating to DNA repair, while the accessory genome included 2215 genes relating to sugar utilization, amino acid synthesis, membrane biogenesis, lipopolysaccharide biosynthesis, cell wall synthesis, urea synthesis, secretion, protein structure and biosynthetic pathways (Delmont and Eren 2018). Core DNA repair genes are necessary for survival in surface waters that are exposed to higher levels of UV radiation (Jeffrey et al. 1996).

9. Summary

Gene duplication is an important mode of microbial evolution throughout the history of living organisms (Moghe and Last 2015; Virdi and Sachdeva 2005). These events can be responsible for the evolution of microbial genomes resulting in novel metabolites (Fewer and Metsä-Ketelä 2020), novel species (Virdi and Sachdeva 2005), adaptation to a specific environment (Sheridan et al. 2020), the expansion of gene families (Finet et al. 2013) and increasing the levels of an enzyme (Moghe and Last 2015). Gene duplicates have been implicated in the evolution of different kingdoms from plants (Guo et al. 2016; The *Arabidopsis* Genome 2000) and animals (Li et al. 2001; Postlethwait et al. 2000) to bacteria (Chirakul et al. 2019) and archaea (Sheridan et al. 2020). Genes responsible for allowing a species to adapt to a specific environment are found within the organisms' accessory genomes.

Bacteria are thought to dominate the world's oceans, making up a large portion of ocean biomass (Whitman et al. 1998). Marine bacteria are capable of achieving this through their great genetic diversity due to their high turnover, environmental complexity and large biomass (Whitman et al. 1998). These microorganisms carry out important ecosystem functions within the ocean that include carbon sequestration, providing the basis of ocean food webs and the cycling of elements such as iron, nitrogen, sulphur and silica (Glöckner et al. 2012). This is particularly important in the most important iron-limited carbon sink in the ocean, the SIO (Glöckner et al. 2012).

The SIO represents an extreme environment limited by the concentration of iron, however it remains an important carbon sink (Glöckner et al. 2012). Few studies have been conducted in the SIO in terms of characterizing microbial community structure and

function (Phoma et al. 2018). Studies on marine microbial communities are limited by access to these communities through extended research voyages and a large number of marine prokaryotes being difficult to culture (Arnosti et al. 2014; Joint et al. 2010; López-Pérez et al. 2020). It is thus important to expand our knowledge on marine microbial communities in the SIO and understand the mechanisms used for survival by the microorganisms living in this extreme environment through the use of metagenomic approaches.

10. Problem statement

Microbial communities contribute substantially to ecosystem services in the oceans. However, despite the substantial advances linked to elucidating the phylogeny and function of microbes, we lack mechanistic insights. Studies in microbial ecology have provided some insights regarding phylogeny and functional attributes, yet fundamental questions regarding the evolution of marine microbes remain unclear. For instance, the extent of marine microbial gene duplications in the Subantarctic Indian Ocean at Marion Island and between Marion Island and South Africa is largely understudied. Given the known importance of these environments, it is reasonable to predict that microbial communities play disproportionate roles in influencing biogeochemical cycling. Yet, there few studies on these microbiomes, resulting in a gap in our understanding regarding microbial community contributions. It is important to reduce the knowledge deficit regarding the extant and the fate of gene duplications in SIO microbial communities, as well as if these gene duplications are, responsible for the adaptation of these microbes to the extreme environments in the SIO. It is important to study the possible impact gene duplication events have on the adaptation of marine microbial communities to the already

changing environmental conditions in the SIO. The PEIs form part of a MPA that can be used as a model to predict microbial response to climate change due to their isolation and legal protection.

11. Hypothesis, aims and Objectives

11.1. Hypothesis:

We predict that marine microbes have, as yet, extensive gene duplications. Further, we predict that these genes encode important enzymes with functions that allow the species to adapt to their environment in the Subantarctic Indian Ocean.

11.2. Aim:

To elucidate the extent of gene duplications in microbial communities from Subantarctic Indian Ocean waters.

11.3. Objectives:

1. To produce metagenome-assembled genomes (MAGs) of SIO microbial community members.

2. To identify significant and assign function to putative paralogs within Medium and High quality MAGs.

3. To determine the pangenome of the MAGs and where the paralogs lie within it.

12. Uncategorized References

Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, Royo-Llonch M, Paoli L, Sunagawa S, Hingamp P. 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun. Biol. 4(1)**:** 1-15. https://doi.org/10.1038/s42003-021-02112-2.

Álvarez-Lugo A, Becerra A. 2021. The role of gene duplication in the divergence of enzyme function: a comparative approach. Front. Genet. 12(https://doi.org/10.3389/fgene.2021.641817.

Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber JA. 2017. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat. Commun. 8(1)**:** 1-11. https://doi.org/10.1038/s41467-017-01228-6.

Anderson RP, Roth JR. 1977. Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. 31(1)**:** 473-505. https://doi.org/10.1146/annurev.mi.31.100177.002353.

Andersson SG, Kurland CG. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6(7)**:** 263-268. https://doi.org/10.1016/S0966-842X(98)01312-2.

Ansorge I, Lutjeharms J. 2002. The hydrography and dynamics of the ocean environment of the Prince Edward Islands (Southern Ocean). J. Mar. Syst. 37(1-3)**:** 107-127. https://doi.org/10.1016/S0924-7963(02)00198-7.

Ansorge IJ, Skelton P, Bekker A, de Bruyn NP, Butterworth D, Cilliers P, Cooper J, Cowan DA, Dorrington R, Fawcett S. 2017. Exploring South Africa's southern frontier: A 20-year vision for polar research through the South African National Antarctic Programme. S. Afr. J. Sci. 113(5-6)**:** 1-7. http://dx.doi.org/10.17159/sajs.2017/a0205

Arber W. 2000. Genetic variation: molecular mechanisms and impact on microbial evolution. FEMS Microbiol. Rev. 24(1)**:** 1-7. https://doi.org/10.1111/j.1574-6976.2000.tb00529.x.

Arnosti C, Bell C, Moorhead D, Sinsabaugh R, Steen A, Stromberger M, Wallenstein M, Weintraub M. 2014. Extracellular enzymes in terrestrial, freshwater, and marine environments: perspectives on system variability and common research needs. Biogeochemistry. 117(1)**:** 5-21. https://doi.org/10.1007/s10533-013-9906-5.

Arrigo KR. 2005. Marine microorganisms and global nutrient cycles. Nature. 437(7057)**:** 349-355. https://doi.org/10.1038/nature04159.

Baker KM, Gobler CJ, Collier JL. 2009. Urease gene seqeunces from algae and heterotrophic bacteria in anexic and nonanexic phytoplankton cultures. J. Phycol. 45(3)**:** 625-634. https://doi.org/10.1111/j.1529-8817.2009.00680.x.

Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends Ecol. Evol. 28(8)**:** 489-495. https://doi.org/10.1016/j.tree.2013.04.002.

Belkin IM, Gordon AL. 1996. Southern Ocean fronts from the Greenwich meridian to Tasmania. J. Geophys. Res. Oceans. 101(C2)**:** 3675-3696. https://doi.org/10.1029/95JC02750.

Bell DB, 2023. Duplication [internet]. Available from: https://www.genome.gov/genetics-glossary/Duplication

Bennett GM, Moran NA. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biol. Evol. 5(9)**:** 1675-1688. https://doi.org/10.1093/gbe/evt118.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. . 104(43)**:** 17004-17009. https://doi.org/10.1073/pnas.0707158104.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF. 1997. The complete genome sequence of *Escherichia coli* K-12. Science. 277(5331)**:** 1453-1462. https://doi.org/10.1126/science.277.5331.1453.

Blouin N, Brawley S. 2012. An AFLP-based test of clonality in widespread, putatively asexual populations of *Porphyra umbilicalis* (Rhodophyta) in the Northwest Atlantic with an in silico analysis for bacterial contamination. Mar. Biol. 159(12)**:** 2723-2729. https://doi.org/10.1007/s00227-012-2029-z.

Boden BP. 1988. Observations of the island mass effect in the Prince Edward Archipelago. Polar Biol. 9(1)**:** 61-68. https://doi.org/10.1007/BF00441765.

Bondoso J, Balague V, Gasol JM, Lage OM. 2014. Community composition of the Planctomycetes associated with different macroalgae. FEMS Microbiol. Ecol. 88(3)**:** 445-456. https://doi.org/10.1111/1574-6941.12258.

Bonk F, Popp D, Harms H, Centler F. 2018. PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls. J. Microbiol. Methods. 153(139-147. https://doi.org/10.1016/j.mimet.2018.09.015.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbø CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37(1)**:** 283-328. https://doi.org/10.1146/annurev.genet.37.050503.084247.

Boyd P, LaRoche J, Gall M, Frew R, McKay RML. 1999. Role of iron, light, and silicate in controlling algal biomass in subantarctic waters SE of New Zealand. J. Geophys. Res. Oceans. 104(C6)**:** 13395-13408. https://doi.org/10.1029/1999JC900009.

Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genom. 11(1)**:** 1-17. https://doi.org/10.1186/1471-2164-11-588.

Breuert S, Allers T, Spohn G, Soppa J. 2006. Regulated polyploidy in halophilic archaea. PloS One. 1(1)**:** e92. https://doi.org/10.1371/journal.pone.0000092.

Brockhurst MA, Harrison E, Hall JP, Richards T, McNally A, MacLean C. 2019. The ecology and evolution of pangenomes. Curr. Biol. 29(20)**:** R1094-R1103. https://doi.org/10.1016/j.cub.2019.08.012.

Bryant JA, Aylward FO, Eppley JM, Karl DM, Church MJ, DeLong EF. 2016. Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre. ISME J. 10(6)**:** 1308-1322. https://doi.org/10.1038/ismej.2015.221

Cahn JK, Piel J. 2021. Opening up the Single‐Cell Toolbox for Microbial Natural Products Research. Angew. Chem., Int. Ed. 60(34)**:** 18412-18428. https://doi.org/10.1002/anie.201900532.

Capone DG, Hutchins DA. 2013. Microbial biogeochemistry of coastal upwelling regimes in a changing ocean. Nat. Geosci. 6(9)**:** 711-717. https://doi.org/10.1038/ngeo1916.

Casjens S. 1998. The diverse and dynamic structure of bacterial genomes. Ann. Rev. Genet. 32(1)**:** 339-377.

Cermeño P, De Vargas C, Abrantes F, Falkowski PG. 2010. Phytoplankton biogeography and community stability in the ocean. PloS One. 5(4)**:** e10037. https://doi.org/10.1371/journal.pone.0010037.

Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. Genome Res. 30(3)**:** 315-333. http://doi:10.1101/gr.258640.119.

Chirakul S, Somprasong N, Norris MH, Wuthiekanun V, Chantratita N, Tuanyok A, Schweizer HP. 2019. *Burkholderia pseudomallei* acquired ceftazidime resistance due to gene duplication and amplification. Int. J. Antimicrob. Agents. 53(5)**:** 582-588. https://doi.org/10.1016/j.ijantimicag.2019.01.003.

Chisholm S, Olson R, Zettler E, Goericke R, Waterbury J, Welschmeyer N. 1988. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. Nature. 334(340-343. http://dx.doi.org/10.1038/334340a0.

Clingenpeel S, Clum A, Schwientek P, Rinke C, Woyke T. 2015. Reconstructing each cell's genome within complex microbial communities—dream or reality? Front. Microbiol. 5(1-6. https://doi.org/10.3389/fmicb.2014.00771.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9(12)**:** 938-950. https://doi.org/10.1038/nrg2482.

Copley SD. 2020. Evolution of new enzymes by gene duplication and divergence. FEBS J. 287(7)**:** 1262-1283. https://doi.org/10.1111/febs.15299.

Cowan DA, Ramond J-B, Makhalanyane TP, De Maayer P. 2015. Metagenomics of extreme environments. Curr. Opin. Microbiol. 25(97-102. https://doi.org/10.1016/j.mib.2015.05.005.

Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2(3)**:** e55. https://doi.org/10.1371/journal.pbio.0020055.

de Baar HJ, Buma AG, Nolting RF, Cadée GC, Jacques G, Tréguer PJ. 1990. On iron limitation of the Southern Ocean: experimental observations in the Weddell and Scotia Seas. Mar. Ecol. Prog. Ser. 105-122.

Del Giorgio PA, Duarte CM. 2002. Respiration in the open ocean. Nature. 420(6914)**:** 379-384. https://doi.org/10.1038/nature01165.

Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ. 6(e4320. https://doi.org/10.7717/peerj.4320.

den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M. 2010. Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics. 11(1)**:** 10.1186/1471-2164-11-688.

DeVries T. 2014. The oceanic anthropogenic CO_2 sink: Storage, air‐sea fluxes, and transports over the industrial era. Glob. Biogeochem. Cycles. 28(7)**:** 631-647. https://doi.org/10.1002/2013GB004739.

Djurhuus A, Port J, Closek CJ, Yamahara KM, Romero-Maraccini O, Walz KR, Goldsmith DB, Michisaki R, Breitbart M, Boehm AB. 2017. Evaluation of filtration and DNA extraction methods for environmental DNA biodiversity assessments across multiple trophic levels. Front. Mar. Sci. 4(314.

Elliott KT, Cuff LE, Neidle EL. 2013. Copy number change: evolving views on gene amplification. Future Microbiol. 8(7)**:** 887-899. https://doi.org/10.2217/fmb.13.53.

Elser JJ, Bracken ME, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, Ngai JT, Seabloom EW, Shurin JB, Smith JE. 2007. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol. Lett. 10(12)**:** 1135-1142. https://doi.org/10.1111/j.1461-0248.2007.01113.x.

Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, Low DE, Gubbay JB. 2012. Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. PloS One. 7(3): e32807. https://doi.org/10.1371/journal.pone.0032807.

Fadiji AE, Babalola OO. 2020. Metagenomics methods for the study of plant-associated microbial communities: a review. J. Microbiol. Methods. 170(105860. https://doi.org/10.1016/j.mimet.2020.105860.

Falkowski PG, Barber RT, Smetacek V. 1998. Biogeochemical controls and feedbacks on ocean primary production. Science. 281(5374): 200-206. https://doi.org/10.1126/science.281.5374.200.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science. 320(5879): 1034-1039. https://doi.org/10.1126/science.1153213.

Fewer DP, Metsä‐Ketelä M. 2020. A pharmaceutical model for the molecular evolution of microbial natural products. FEBS J. 287(7): 1429-1449. https://doi.org/10.1111/febs.15129.

Finet C, Berne-Dedieu A, Scutt CP, Marlétaz F. 2013. Evolution of the ARF gene family in land plants: old domains, new tricks. Mol. Biol. Evol. 30(1): 45-56. https://doi.org/10.1093/molbev/mss220.

Fischer WW, Hemp J, Johnson JE. 2016. Evolution of oxygenic photosynthesis. Annu. Rev. Earth Planet. Sci. 44(647-683. https://doi.org/10.1146/annurev-earth-060313-054810.

Flemming H-C, Wuertz S. 2019. Bacteria and archaea on Earth and their abundance in biofilms. Nat. Rev. Microbiol. 17(4): 247-260. https://doi.org/10.1038/s41579-019-0158-9.

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WK, Lomas MW, Veneziano D. 2013. Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc. Natl. Acad. Sci. 110(24): 9824-9829. https://doi.org/10.1073/pnas.1307701110.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151(4): 1531-1545.

Francioli D, Lentendu G, Lewin S, Kolb S. 2021. DNA metabarcoding for the characterization of terrestrial microbiota—pitfalls and solutions. Microorganisms. 9(2): 361. https://doi.org/10.3390/microorganisms9020361.

Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res. 15(12): 1603-1610. https://doi.org/10.1101/gr.3724205.

Frioux C, Singh D, Korcsmaros T, Hildebrand F. 2020. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput. Struct. Biotechnol. J. 18(1722-1734. https://doi.org/10.1016/j.csbj.2020.06.028.

Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12(4): 148-154. https://doi.org/10.1016/j.tim.2004.02.007.

Glaubitz S, Kießlich K, Meeske C, Labrenz M, Jürgens K. 2013. SUP05 dominates the gammaproteobacterial sulfur oxidizer assemblages in pelagic redoxclines of the central Baltic and Black Seas. Appl. Environ. Microbiol. 79(8)**:** 2767-2776. https://doi.org/10.1128/AEM.03777-12.

Glöckner FO, Gasol JM, McDonough N, Calewaert J-B. 2012. Marine microbial diversity and its role in ecosystem functioning and environmental change. European Science Foundation.

Gómez-Anduro G, Ceniceros-Ojeda EA, Casados-Vázquez LE, Bencivenni C, Sierra-Beltrán A, Murillo-Amador B, Tiessen A. 2011. Genome-wide analysis of the beta-glucosidase gene family in maize (*Zea mays* L. *var B73*). Plant Mol. Biol. 77(1)**:** 159-183. https://doi.org/10.1007/s11103-011-9800-2.

Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A, Beman J, Abell G, Philippot L, Prosser J. 2016. Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? Front. Microbiol. 7(214. https://doi.org/10.3389/fmicb.2016.00214.

Groisman EA, Ochman H. 1994. How to become a pathogen. Trends Microbiol. 2(8)**:** 289-294.

Grossart HP, Massana R, McMahon KD, Walsh DA. 2020. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnol. Oceanogr. 65(S2-S20. https://doi.org/10.1002/lno.11382.

Güller M, Puccinelli E, Zelaya DG. 2020. The Antarctic Circumpolar Current as a dispersive agent in the Southern Ocean: evidence from bivalves. Mar. Biol. 167(10)**:** 143. https://doi.org/10.1007/s00227-020-03746-2.

Guo H-S, Zhang Y-M, Sun X-Q, Li M-M, Hang Y-Y, Xue J-Y. 2016. Evolution of the KCS gene family in plants: the history of gene duplication, sub/neofunctionalization and redundancy. Mol. Genet. Genom. 291(2)**:** 739-752. https://doi.org/10.1007/s00438-015-1142-3.

Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Ann. Rev. Microbiol. 54(1)**:** 641-679.

Halo I, Dorrington R, Bornman T, De Villiers S, Fawcett S. 2016. South Africa in the Antarctic Circumnavigation Expedition: A multi-institutional and interdisciplinary scientific project. S. Afr. J. Sci. 112(9-10)**:** 1-4. https://dx.doi.org/10.17159/sajs.2016/a0173.

Han K, Li Z-f, Peng R, Zhu L-p, Zhou T, Wang L-g, Li S-g, Zhang X-b, Hu W, Wu Z-h. 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Sci. Rep. 3(1)**:** 2101. https://doi.org/10.1038/srep02101.

Hartmann T, Middendorf M, Bernt M. 2018. Genome rearrangement analysis: cut and join genome rearrangements and gene cluster preserving approaches. Comp. Genom.: Methods Protoc. 261-289. https://doi.org/10.1007/978-1-4939-7463-4_9.

Henley SF, Cavan EL, Fawcett SE, Kerr R, Monteiro T, Sherrell RM, Bowie AR, Boyd PW, Barnes DK, Schloss IR. 2020. Changing biogeochemistry of the Southern Ocean and its ecosystem implications. Front. Mar. Sci. 7(581. https://doi.org/10.3389/fmars.2020.00581.

Hermes JC, Masumoto Y, Beal L, Roxy MK, Vialard J, Andres M, Annamalai H, Behera S, D'Adamo N, Feng M. 2019. A sustained ocean observing system in the Indian

Ocean for climate related scientific knowledge and societal needs. Front. Mar. Sci. 355. https://doi.org/10.3389/fmars.2019.00355.
Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. 20(6)**:** 945-954.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. B: Biol. Sci. 256(1346)**:** 119-124. https://doi.org/10.1098/rspb.1994.0058.
Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, Coelho LP, Endo H, Gasol JM, Gregory AC. 2019. Global trends in marine plankton diversity across kingdoms of life. Cell. 179(5)**:** 1084-1097. e1021. https://doi.org/10.1016/j.cell.2019.10.008.
Irwin DM. 1995. Evolution of the bovine lysozyme gene family: changes in gene expression and reversion of function. J. Mol. Evol. 41(3)**:** 299-312.
Irwin DM, Wilson AC. 1990. Concerted evolution of ruminant stomach lysozymes. Characterization of lysozyme cDNA clones from sheep and deer. J. Biol. Chem. 265(9)**:** 4944-4952.
Jeffrey WH, Pledger RJ, Aas P, Hager S, Coffin RB, Von Haven R, Mitchell DL. 1996. Diel and depth profiles of DNA photodamage in bacterioplankton exposed to ambient solar ultraviolet radiation. Mar. Ecol. Prog. Ser. 137(283-291. http://dx.doi.org/10.3354/meps137283.
Johnson WM, Longnecker K, Kido Soule MC, Arnold WA, Bhatia MP, Hallam SJ, Van Mooy BA, Kujawinski EB. 2020. Metabolite composition of sinking particles differs from surface suspended particles across a latitudinal transect in the South Atlantic. Limnol. Oceanogr. 65(1)**:** 111-127. https://doi.org/10.1002/lno.11255.
Joint I, Mühling M, Querellou J. 2010. Culturing marine bacteria–an essential prerequisite for biodiscovery. Microb. Biotechnol. 3(5)**:** 564-575. https://doi.org/10.1111/j.1751-7915.2010.00188.x.
Jones DL, Baxter BK. 2017. DNA repair and photoprotection: Mechanisms of overcoming environmental ultraviolet radiation exposure in halophilic archaea. Front. Microbiol. 8(1882. https://doi.org/10.3389/fmicb.2017.01882.
Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. 11(4)**:** 555-565.
Kamneva OK, Knight SJ, Liberles DA, Ward NL. 2012. Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol. Evol. 4(12)**:** 1375-1390. https://doi.org/10.1093/gbe/evs113.
Karl DM. 2007. Microbial oceanography: paradigms, processes and promise. Nat. Rev. Microbiol. 5(10)**:** 759-769. https://doi.org/10.1038/nrmicro1749.
Kemeny PC, Weigand M, Zhang R, Carter BR, Karsh KL, Fawcett SE, Sigman DM. 2016. Enzyme‐level interconversion of nitrate and nitrite in the fall mixed layer of the Antarctic Ocean. Global Biogeochem. Cy. 30(7)**:** 1069-1085. https://doi.org/10.1002/2015GB005350.
Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA. 2020. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinform. 21(1)**:** 1-8. https://doi.org/10.1186/s12859-020-03585-4.

Kim JW, Brawley SH, Prochnik S, Chovatia M, Grimwood J, Jenkins J, LaButti K, Mavromatis K, Nolan M, Zane M, Schmutz J, Stiller JW, Grossman AR. 2016. Genome analysis of Planctomycetes inhabiting blades of the red alga *Porphyra umbilicalis*. PLoS One. 11(3)**:** https://doi.org/10.1371/journal.pone.0151883.

Kirkman S, Mann B, Sink K, Adams R, Livingstone T, Mann-Lang J, Pfaff M, Samaai T, van der Bank M, Williams L. 2021. Evaluating the evidence for ecological effectiveness of South Africa's marine protected areas. Afr. J. Mar. Sci. 43(3)**:** 389-412.

Kirkman SP, Yemane DG, Lamont T, Meyer MA, Pistorius PA. 2016. Foraging behavior of subantarctic fur seals supports efficiency of a marine reserve's design. PLoS One. 11(5)**:** e0152370.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B: Biol. Sci. 279(1749)**:** 5048-5057. https://doi.org/10.1098/rspb.2012.1108.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55(1)**:** 709-742. https://doi.org/10.1146/annurev.micro.55.1.709.

Kuypers MM, Marchant HK, Kartal B. 2018. The microbial nitrogen-cycling network. Nat. Rev. Microbiol. 16(5)**:** 263-276. https://doi.org/10.1038/nrmicro.2018.9.

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T. 2015. Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genom. 15(141-161. https://doi.org/10.1007/s10142-015-0433-4.

Lauer S, Gresham D. 2019. An evolving view of copy number variants. Curr. Genet. 65(6)**:** 1287-1295. https://doi.org/10.1007/s00294-019-00980-0.

Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. 95(16)**:** 9413-9417.

Le PT, Makhalanyane TP, Guerrero LD, Vikram S, Van de Peer Y, Cowan DA. 2016. Comparative metagenomic analysis reveals mechanisms for stress response in hypoliths from extreme hyperarid deserts. Genome Biol. Evol. 8(9)**:** 2737-2747.

Le Roux PC, McGeoch MA. 2008. Rapid range expansion and community reorganization in response to warming. Glob. Chang. Biol. 14(12)**:** 2950-2962. https://doi.org/10.1111/j.1365-2486.2008.01687.x.

Li J, Gu X, Gui Y. 2020. Prokaryotic diversity and composition of sediments from Prydz Bay, the Antarctic peninsula region, and the Ross Sea, southern ocean. Front. Microbiol. 11(783. https://doi.org/10.3389/fmicb.2020.00783.

Li W-H, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. Nature. 409(6822)**:** 847-849.

Li Y, Tang K, Zhang L, Zhao Z, Xie X, Chen C-TA, Wang D, Jiao N, Zhang Y. 2018. Coupled carbon, sulfur, and nitrogen cycles mediated by microorganisms in the water column of a shallow-water hydrothermal ecosystem. Front. Microbiol. 9(2718. https://doi.org/10.3389/fmicb.2018.02718.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in Caenorhabditis elegans. Curr. Biol. 21(4)**:** 306-310. http://dx.doi.org/10.1016%2Fj.cub.2011.01.026.

Lombard A, Reyers B, Schonegevel L, Cooper J, Smith-Adao L, Nel D, Froneman PW, Ansorge IJ, Bester M, Tosh C. 2007. Conserving pattern and process in the

Southern Ocean: designing a marine protected area for the Prince Edward Islands. Antarctic Science. 19(1): 39-54. https://doi.org/10.1017/S0954102007000077.
Long M. 2001. Evolution of novel genes. Curr. Opin. Genet. Dev. . 11(6): 673-680. https://doi.org/10.1016/S0959-437X(00)00252-5.
López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-Valera F. 2020. The evolutionary success of the marine bacterium SAR11 analyzed through a metagenomic perspective. Msystems. 5(5): e00605-00620. https://doi.org/10.1128/mSystems.00605-20.
Ludt K, Soppa J. 2018. Influence of origin recognition complex proteins on the copy numbers of three chromosomes in *Haloferax volcanii*. J. Bacteriol. 200(17): e00161-00118. https://doi.org/10.1128/JB.00161-18.
Lutjeharms JR. 2006. The Agulhas Current. Springer.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290(5494): 1151-1155. 10.1126/science.290.5494.1151.
Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. J. Genet. 92(1): 155-161. https://doi.org/10.1007/s12041-013-0212-8.
Marambio J, Mendez F, Ocaranza P, Rodriguez J, Rosenfeld S, Ojeda J, Murcia S, Terrados J, Bischof K, Mansilla A. 2017. Seasonal variations of the photosynthetic activity and pigment concentrations in different reproductive phases of *Gigartina skottsbergii* (Rhodophyta, Gigartinales) in the Magellan region, sub-Antarctic Chile. J. Appl. Phycol. 29(721-729. https://doi.org/10.1007/s10811-016-0913-5.
McCarthy CG, Fitzpatrick DA. 2019. Pan-genome analyses of model fungal species. Microb. Genom. 5(2): https://doi.org/10.1099%2Fmgen.0.000243.
McCarthy M, Bronk D. 2008. Analytical methods for the study of nitrogen. In: Nitrogen in the marine environment. San Diego: Academic Press. p. 1219-1275.
McInerney JO, McNally A, O'connell MJ. 2017. Why prokaryotes have pangenomes. Nat. Microbiol. 2(4): 1-5. https://doi.org/10.1038/nmicrobiol.2017.40.
Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 318(5848): 245-250. https://doi.org/10.1126/science.1143609.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17(10): 589-596. https://doi.org/10.1016/S0168-9525(01)02447-7.
Mitra A, Flynn KJ, Burkholder JM, Berge T, Calbet A, Raven JA, Granéli E, Glibert PM, Hansen PJ, Stoecker DK. 2014. The role of mixotrophic protists in the biological carbon pump. Biogeosciences. 11(4): 995-1005. https://doi.org/10.5194/bg-11-995-2014.
Moghe GD, Last RL. 2015. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol. 169(3): 1512-1523. https://doi.org/10.1104/pp.15.00994.
Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. 2004. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. Nature. 430(7003): 1027-1031. https://doi.org/10.1038/nature02824.

Moore C, Mills M, Arrigo K, Berman-Frank I, Bopp L, Boyd P, Galbraith E, Geider R, Guieu C, Jaccard S. 2013. Processes and patterns of oceanic nutrient limitation. Nat. Geosci. 6(9)**:** 701-710. https://doi.org/10.1038/ngeo1765.

Moore JK, Fu W, Primeau F, Britten GL, Lindsay K, Long M, Doney SC, Mahowald N, Hoffman F, Randerson JT. 2018. Sustained climate warming drives declining marine biological productivity. Science. 359(6380)**:** 1139-1143. https://doi.org/10.1126/science.aao6379.

Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. 2(12)**:** 1-12. https://doi.org/10.1186/gb-2001-2-12-research0054.

Moreau S, Mostajir B, Bélanger S, Schloss IR, Vancoppenolle M, Demers S, Ferreyra GA. 2015. Climate change enhances primary production in the western Antarctic Peninsula. Glob. Change Biol. 21(6)**:** 2191-2205. https://doi.org/10.1111/gcb.12878.

Nayfach S, Pollard KS. 2016. Toward accurate and quantitative comparative metagenomics. Cell. 166(5)**:** 1103-1116. https://doi.org/10.1016/j.cell.2016.08.007.

Norouzi-Beirami MH, Marashi S-A, Banaei-Moghaddam AM, Kavousi K. 2021. CAMAMED: a pipeline for composition-aware mapping-based analysis of metagenomic data. NAR Genom. Bioinform. 3(1)**:** lqaa107. https://doi.org/10.1093/nargab/lqaa107.

Ochman H. 2001. Lateral and oblique gene transfer. Curr. Opin. Genet. Dev. 11(6)**:** 616-619.

Ohno S. 1970. Evolution by gene duplication. . London-New York: Allen & Unwin.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. 2011. Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiol. Mol. Biol. Rev. 75(2)**:** 361-422. https://doi.org/10.1128/MMBR.00039-10.

Pajares S, Varona-Cordero F, Hernández-Becerril DU. 2020. Spatial distribution patterns of bacterioplankton in the oxygen minimum zone of the tropical mexican pacific. Microb. Ecol. 80(519-536. https://doi.org/10.1007/s00248-020-01508-7.

Pang TY, Lercher MJ. 2019. Each of 3,323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. Proc. Natl. Acad. Sci. 116(1)**:** 187-192. https://doi.org/10.1073/pnas.1718997115.

Parvez MK, Parveen S. 2017. Evolution and emergence of pathogenic viruses: past, present, and future. Intervirology. 60(1-2)**:** 1-7. https://doi.org/10.1159/000478729.

Pérez Castro S, Borton MA, Regan K, Hrabe de Angelis I, Wrighton KC, Teske AP, Strous M, Ruff SE. 2021. Degradation of biological macromolecules supports uncultured microbial populations in Guaymas Basin hydrothermal sediments. ISME J. 15(12)**:** 3480-3497. https://doi.org/10.1038/s41396-021-01026-5.

Perissinotto R, Mayzaud P, Nichols P, Labat J. 2007. Grazing by *Pyrosoma atlanticum* (Tunicata, Thaliacea) in the south Indian Ocean. Mar. Ecol. Prog. Ser. 330(1-11. https://doi.org/10.3354/meps330001.

Perissinotto R, Rae CD. 1990. Occurrence of anticyclonic eddies on the Prince Edward Plateau (Southern Ocean): effects on phytoplankton biomass and production. Deep Sea Res. Part I Oceanogr. Res. Pap. 37(5)**:** 777-793. https://doi.org/10.1016/0198-0149(90)90006-H.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R. 2015. Open science resources for the discovery and analysis of *Tara* Oceans data. Sci. Data. 2(1)**:** 1-16.

Phoma BS, Makhalanyane TP. 2021. Depth-dependent variables shape community structure and functionality in the Prince Edward Islands. Microb. Ecol. 81(2)**:** 396-409. https://doi.org/10.1007/s00248-020-01589-4.

Phoma S, Vikram S, Jansson JK, Ansorge IJ, Cowan DA, Van de Peer Y, Makhalanyane TP. 2018. Agulhas Current properties shape microbial community diversity and potential functionality. Sci. Rep. 8(1)**:** 1-12. https://doi.org/10.1038/s41598-018-28939-0.

Podell S, Blanton JM, Oliver A, Schorn MA, Agarwal V, Biggs JS, Moore BS, Allen EE. 2020. A genomic view of trophic and metabolic diversity in clade-specific *Lamellodysidea* sponge microbiomes. Microbiome. 8(1)**:** 1-17. https://doi.org/10.1186/s40168-020-00877-y.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan Y-L, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res. 10(12)**:** 1890-1902.

Prince VE, Pickett FB. 2002. Splitting pairs: The diverging fates of duplicated genes. Nat. Rev. Genet. 3(11)**:** 827-837. 10.1038/nrg928.

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science. 329(5988)**:** 223-226. https://doi.org/10.1126/science.1188800.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. Curr. Opin. Plant Biol. 17(43-48. https://doi.org/10.1016/j.pbi.2013.11.002.

Rholl DA, Papp-Wallace KM, Tomaras AP, Vasil ML, Bonomo RA, Schweizer HP. 2011. Molecular investigations of PenA-mediated β-lactam resistance in *Burkholderia pseudomallei*. Front. Microbiol. 2(1389-1399. https://doi.org/10.3389/fmicb.2011.00139.

Riehle MM, Bennett AF, Long AD. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. Proc. Natl. Acad. Sci. 98(2)**:** 525-530. https://doi.org/10.1073/pnas.98.2.525.

Rivas AJ, Labella AM, Borrego JJ, Lemos ML, Osorio CR. 2014. Evidence for horizontal gene transfer, gene duplication and genetic variation as driving forces of the diversity of haemolytic phenotypes in *Photobacterium damselae* subsp. *damselae*. FEMS Microbiol. Lett. 355(2)**:** 152-162. https://doi.org/10.1111/1574-6968.12464.

Rouault M, Mélice JL, Reason CJ, Lutjeharms JR. 2005. Climate variability at Marion Island, southern ocean, since 1960. J. Geophys. Res. Oceans. 110(C5)**:** https://doi.org/10.1029/2004JC002492.

Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. 2010. Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. Proc. Natls. Acad. Sci. 107(37)**:** 16184-16189. https://doi.org/10.1073/pnas.1009513107.

Schönknecht G, Weber AP, Lercher MJ. 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays. 36(1)**:** 9-20. https://doi.org/10.1002/bies.201300095.

Selander R, Levin B. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science. 210(4469)**:** 545-547. 10.1126/science.6999623.
Serres MH, Kerr AR, McCormack TJ, Riley M. 2009. Evolution by leaps: gene duplication in bacteria. Biol. Direct. 4(1)**:** 1-17. https://doi.org/10.1186/1745-6150-4-46.
Shakya M, Lo C-C, Chain PS. 2019. Advances and challenges in metatranscriptomic analysis. Front. Genet. 10(904. https://doi.org/10.3389/fgene.2019.00904.
Sheridan PO, Raguideau S, Quince C, Holden J, Zhang L, Williams TA, Gubry-Rangin C. 2020. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat. Commun. 11(1)**:** 1-12. https://doi.org/10.1038/s41467-020-19132-x.
Smith Jr K, Ruhl H, Bett B, Billett D, Lampitt R, Kaufmann R. 2009. Climate, carbon cycling, and deep-ocean ecosystems. Proc. Natl. Acad. Sci. 106(46)**:** 19211-19218. https://doi.org/10.1073/pnas.0908322106.
Smith V. 2002. Climate change in the sub-Antarctic: an illustration from Marion Island. Clim. Change. 52(3)**:** 345-357. https://doi.org/10.1023/A:1013718617277

Smith VR, Froneman PW. 2008. Nutrient dynamics in the vicinity of the Prince Edward Islands. In: The Prince Edward Islands: Land-Sea Interactions in a Changing Ecosystem. Stellenbosch: AFRICAN SUN MeDIA. p. 165-179.
Soares MA, Bhaskar PV, Anilkumar N, Naik RK, George JV, Mishra RK, Dessai DG. 2020. Characteristics of particulate organic matter within the photic water column: A case study across the fronts in the indian sector of the Southern Ocean. Deep Sea Res. Part I Oceanogr. Res. Pap. 178(104851. https://doi.org/10.1016/j.dsr2.2020.104851.
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. Curr. Opin. Genet. Dev. . 35(119-125. https://doi.org/10.1016/j.gde.2015.11.003.
Steinberg AP, Lin M, Kussell E. 2022. Core genes can have higher recombination rates than accessory genes within global microbial populations. Elife. 11(e78533. https://doi.org/10.7554/eLife.78533.
Stepanauskas R. 2012. Single cell genomics: an individual look at microbes. Curr. Opin. Microbiol. 15(5)**:** 613-620. https://doi.org/10.1016/j.mib.2012.09.001.
Stewart C-B, Schilling JW, Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature. 330(6146)**:** 401-404. 10.1038/330401a0.
Stirnimann L, Bornman TG, Verheye HM, Bachèlery ML, Van der Poel J, Fawcett SE. 2021. Plankton community composition and productivity near the Subantarctic Prince Edward Islands archipelago in autumn. Limnol. Oceanogr. 66(12)**:** 4140-4158. https://doi.org/10.1002/lno.11949.
Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A. 2015. Structure and function of the global ocean microbiome. Science. 348(6237)**:**
Sunda WG, Huntsman SA. 1997. Interrelated influence of iron, light and cell size on marine phytoplankton growth. Nature. 390(6658)**:** 389-392. https://doi.org/10.1038/37093.
Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML. 2011. Potential for chemolithoautotrophy

among ubiquitous bacteria lineages in the dark ocean. Science. 333(6047)**:** 1296-1300.
Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonźalez JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc. Natl. Acad. Sci. U.S.A. 110(28)**:** 11463-11468. 10.1073/pnas.1304246110.
Tawfik DS, Khersonsky O. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. 79(471-505. https://doi.org/10.1146/annurev-biochem-030409-143718.
The *Arabidopsis* Genome I. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 408(6814)**:** 796-815. 10.1038/35048692.
Trento A, Galiano M, Videla C, Carballal G, García-Barreno B, Melero JA, Palomo C. 2003. Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. J. Gen. Virol. 84(11)**:** 3115-3120. https://doi.org/10.1099/vir.0.19357-0.
Tully BJ, Wheat CG, Glazer BT, Huber JA. 2018. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. ISME J. 12(1)**:** 1-16.
Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. The Plant Cell. 33(1)**:** 11-26. https://doi.org/10.1093/plcell/koaa015.
Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18(7)**:** 411-424. https://doi.org/10.1038/nrg.2017.26.
Vasquez-Cardenas D, Van De Vossenberg J, Polerecky L, Malkin SY, Schauer R, Hidalgo-Martinez S, Confurius V, Middelburg JJ, Meysman FJ, Boschker HT. 2015. Microbial carbon metabolism associated with electrogenic sulphur oxidation in coastal sediments. ISME J. 9(9)**:** 1966-1978. https://doi.org/10.1038/ismej.2015.10.
Venkatachalam S, Matcher GF, Lamont T, van den Berg M, Ansorge IJ, Dorrington RA. 2019. Influence of oceanographic variability on near‐shore microbial communities of the sub‐Antarctic Prince Edward Islands. Limnol. Oceanogr. 64(1)**:** 258-271. https://doi.org/10.1002/lno.11035.
Ver Eecke HC, Butterfield DA, Huber JA, Lilley MD, Olson EJ, Roe KK, Evans LJ, Merkel AY, Cantin HV, Holden JF. 2012. Hydrogen-limited growth of hyperthermophilic methanogens at deep-sea hydrothermal vents. Proc. Natl. Acad. Sci. 109(34)**:** 13674-13679. https://doi.org/10.1073/pnas.1206632109.
Viberg LT, Sarovich DS, Kidd TJ, Geake JB, Bell SC, Currie BJ, Price EP, Parkhill J. 2017. Within-host evolution of *Burkholderia pseudomallei* during chronic infection of seven Australasian cystic fibrosis patients. mBio. 8(2)**:** 00356-00317. https://doi.org/10.1128/mBio.00356-17.
Virdi JS, Sachdeva P. 2005. Genetic diversity of pathogenic microorganisms: Basic insights, public health implications and the Indian initiatives. Curr. Sci. 113-123.

Volk T, Hoffert MI. 1985. Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean‐driven atmospheric CO2 changes. The carbon cycle and atmospheric CO2: natural variations Archean to present. 32(99-110.

Voordeckers K, Verstrepen KJ. 2015. Experimental evolution of the model eukaryote *Saccharomyces cerevisiae* yields insight into the molecular mechanisms underlying adaptation. Curr. Opin. Microbiol. 28(1-9. https://doi.org/10.1016/j.mib.2015.06.018.

Voss M, Bange HW, Dippner JW, Middelburg JJ, Montoya JP, Ward B. 2013. The marine nitrogen cycle: recent discoveries, uncertainties and the potential relevance of climate change. Philos. Trans. R. Soc. B: Biol. Sci. 368(1621)**:** 20130121. https://doi.org/10.1098/rstb.2013.0121.

Wani A, Rahayu F, Kadarwati F, Suhara C, Singh R, Dhanjal D, Akhtar N, Mir T, Chopra C. Year. Metagenomic screening strategies for bioprospecting enzymes from environmental samples. In: IOP Conf. Ser.: Earth Environ. Sci.; 2022; IOP Publishing. p. 012003

White NJ, Chaowagul W, Wuthiekanun V, Dance DAB, Wattanagoon Y, Pitakwatchara N. 1989. Halving the mortality of severe meloidosis by ceftazidime. The Lancet. 334(8665)**:** 697-701. **https://doi.org/10.1016/S0140-6736(89)90768-X**.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. 95(12)**:** 6578-6583.

Woese CR. 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. 97(15)**:** 8392-8396.

Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. 2015. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 347(6223)**:** 1257594. https://doi.org/10.1126/science.1257594.

Zehr JP, Kudela RM. 2011. Nitrogen cycle of the open ocean: from genes to ecosystems. Annu. Rev. Mar. Sci. 3(197-225. https://doi.org/10.1146/annurev-marine-120709-142819.

Zeng X, Alain K, Shao Z. 2021. Microorganisms from deep-sea hydrothermal vents. Mar. Life sci. Technol. 3(204-230. https://doi.org/10.1007/s42995-020-00086-4.

Zhang CL, Xie W, Martin-Cuadrado A-B, Rodriguez-Valera F. 2015. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. Front. Microbiol. 6(1108. https://doi.org/10.3389/fmicb.2015.01108.

Zhang X, Liu X, He Q, Dong W, Zhang X, Fan F, Peng D, Huang W, Yin H. 2016. Gene turnover contributes to the evolutionary adaptation of *Acidithiobacillus caldus*: insights from comparative genomics. Front. Microbiol. 7(1960. https://doi.org/10.3389/fmicb.2016.01960.

Zhang X, Ward BB, Sigman DM. 2020. Global nitrogen cycle: critical enzymes, organisms, and processes for nitrogen budgets and dynamics. Chem. Rev. 120(12)**:** 5308-5351. https://doi.org/10.1021/acs.chemrev.9b00613.

Zhuang GC, Peña‐Montenegro TD, Montgomery A, Hunter KS, Joye SB. 2018. Microbial metabolism of methanol and methylamine in the Gulf of Mexico: insight into marine carbon and nitrogen cycling. Environ. Microbiol. 20(12)**:** 4543-4554. https://doi.org/10.1111/1462-2920.14406.

Chapter 2

Materials and methods

1. Sampling

1.1. Sample collection from the Subantarctic Indian Ocean (SIO)

Samples were collected at four depths, including '5 meters' (surface samples collected from the underway system), the epipelagic zone (fluorescence maximum, 'Fmax'), the oxygen minimum zone ('OMZ') and approximately 10m above the seafloor ('Deep') to sample a homogenous ecosystem. Surface water samples (10 L) were collected from the crossroads line (CR) and S-transect between South Africa and Marion Island aboard *RV SA Agulhas II* (Figure 1) at 5m using the underway water collection system in duplicate at six and five stations, respectively (n=22). The underway water collection system pumps seawater from 5m beneath the ship to onboard the ship for collection. Water samples were further collected at three different depths, including F-max, OMZ and Deep, from nine stations along the T-transect (n=25) and five stations along the X-M-transect (n=10) around Marion Island (Figure 2) using a 24 x 12 L Niskin bottle CTD rosette multi-sampler (Sea-Bird SBE-911 plus V2 CTD system; Sea-Bird Electronics, Inc., USA). The Niskin bottles are used to sample water at depths specified by the operator, while the CTD sensors capture metadata such as conductivity (S/m), temperature (°C) and depth (m). Samples were labelled according to the sampling station, where 'S' and 'CR' represent those taken between South Africa and Marion Island, respectively. 'T' and 'X' / 'MT' represent samples from the 'T' and 'X-M' transects, respectively, taken around the Prince Edward Islands. 1 L and 4 x 15 mL of seawater was collected at each station for nutrient analysis and extracellular enzyme activity assays, respectively. For each of the samples collected along these transects, we collected properties that relate to salinity (PSU), temperature (°C), conductivity (S/m), Oxygen (ml/L), pressure (db) and depth (m) which were measured by rosette multi-sampler sensor. Coordinates of each sample were

recorded (Table 2) and a map of the sampling stations was created using ggplot2 (version 3.3.6) (Figure 2) (Wickham 2016) in the R statistical package version R 3.3.0+ and R Studio version 2022.07.2+576 (R Core Team 2013).

Table 2: The coordinates of the stations that were sampled, as well as the sample type (CTD or underway) and transect the station belongs to. S = S-transect; CR = Crossroads transect; T = T-transect; X-M = X-M-transect.

Station	Sample Type	Transect	Latitude (S)	Longitude (E)
S01	Underway	S	31°00’404”	22°02’815”
S02	Underway	S	41°29’710”	29°03’304”
S03	Underway	S	39°50’299”	26°29’005”
S04	Underway	S	43°55’848”	33°00’304”
S05	Underway	S	45°44’343”	36°01’913”
CR01	Underway	CR	45°88’919”	30°56’147”
CR03	Underway	CR	43°59’774”	29°13’587”
CR05	Underway	CR	42°07’083”	28°09’652”
CR07	Underway	CR	39°56’629”	26°02’062”
CR11	Underway	CR	37°46’047”	24°32’520”
CR17	Underway	CR	35°37’256”	23°09’286”
T01	CTD	T	45°75’000”	37°30’000”
T02	CTD	T	45°59’961”	37°18’015”
T03	CTD	T	46°11’014”	37°27’632”
T05	CTD	T	46°44’988”	37°18’035”
T07	CTD	T	47°15’059”	37°18’069”
T08A	CTD	T	47°44’983”	37°17’039”
T08B	CTD	T	47°59’965”	37°18’018”
T09	CTD	T	46°45’165”	38°29’926”
T11	CTD	T	46°44’937”	38°06’116”
X04	CTD	X-M	46°40’792”	38°00’026”
M1-01	CTD	X-M	46°51’583”	37°54’010”
M1-07	CTD	X-M	46°40’202”	37°53’987”
M3-01	CTD	X-M	46°47’986”	37°41’992”
X02	CTD	X-M	46°48’003”	37°48’007”

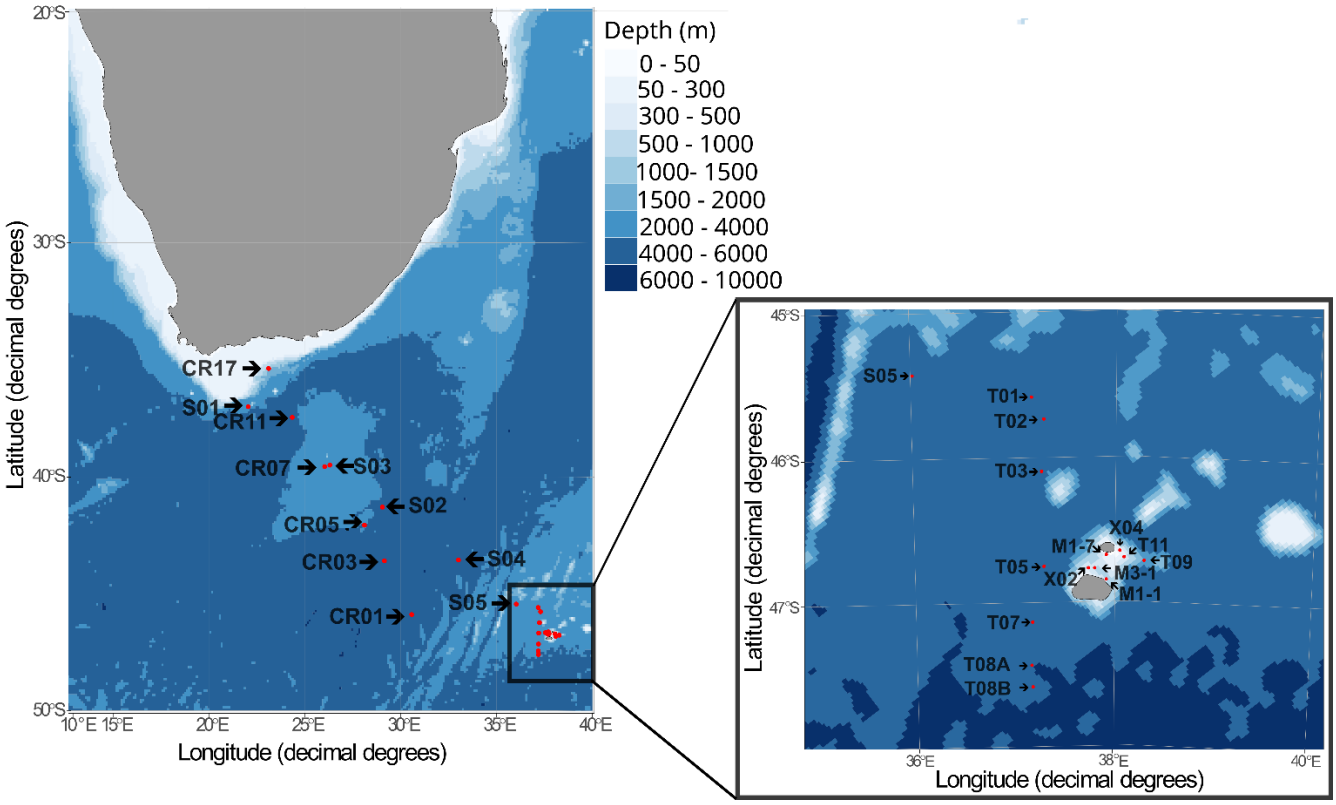

Figure 2: Map of the Prince Edward Islands showing the sampling locations and study area. Water was collected from 14 stations around the PEI's in the 'T-transect' and the 'X-M-transect', as well as 11 stations between Cape Town, South Africa and the PEI's. The red points represent the stations that were sampled. The maps were created using Ocean Data View version 5.6.2 (Schlitzer 2023).

1.1.1. Processing of samples

Five litres of water from each of the 'T-transect', the 'X-M transect' and the 'Crossroads line' samples were filtered immediately, using a 0.45 µm-pore cellulose acetate membrane filter (Sterivex, location) using a vacuum system to capture the particle associated (PA) microbial communities. Filtrates were then filtered using a 0.2 µm

Sterivex filter to capture the free-living (FL) communities (Anderson et al. 2017; Tully et al. 2018). Subsequently, ten litres of water from each of the S-transect samples were filtered using the same protocol. From these filtered samples, the filtrate was spiked with 500 µL iron chloride to enrich for viruses (10 g/L), shaken thoroughly by hand, then incubated in the dark at room temperature for 1 hour, then filtered through a 1.0 µm Sterivex filter. The filters were stored at -20ºC with 1 mL RNALater (ThermoFishcer, USA) during the voyage transport, then at -80ºC until needed (Le et al. 2016). A total 1 L and four 15 mL of seawater samples were collected at each station for nutrient analysis and extracellular enzyme activity assays, respectively. 50 µL of 25% formaldehyde (v/v) was added to each of the four 15 mL water samples collected for extracellular enzyme activity assays, and these were stored at -80ºC.

1.2. Nutrient composition analysis

1.2.1. Nutrient measurements

One litre samples were sent to the University of Cape Town Marine Biogeochemistry Lab (Cape Town, South Africa) for nutrient analysis. The samples were tested for nitrate (NO_3^-), nitrite (NO_2^-), nitrate + nitrite (NO_3^- + NO_2^-), ammonium (NH_4^+), silicate (SiO_4^{4-}) and phosphate (PO_4^{3-}). Measurements were determined using a AA500 AutoAnalyzer (SEAL Analytical Ltd., UK) and auto-analysis protocols as described previously by Grasshof (1976). The detection limit was 0.1 µM for nitrate and nitrite, and silicate, 0.05 µM for phosphate, and 0.01 µM for nitrite. Ammonium concentrations were measured using the fluorometric method as described by Holmes et al. (1999), with a detection limit of 0.01 µM.

1.3. Extracellular enzyme activity assays

1.3.1. Extracellular enzyme activity test assays

To determine potential extracellular enzymatic activities in each sample, water samples collected from each site (4 x 15 ML) were assayed. We assayed four enzymes, β-glucosidase (BG), β-xylosidase (BX), β-N-acetyl-glucosaminidase (NAG) and alkaline phosphatase (AP), respectively. Potential extracellular enzyme activities were assayed using fluorescence, where the substrate analogues used were linked to the fluorescent molecule 4-methylumbelliferone (MUB). Test assays were performed using a dilution series (50, 100, 200, 400 and 500 µM) of each of the four substrate-analogues for each extracellular enzyme assay from a 500 µM stock solution, of each, to determine both the saturation concentration and the ideal incubation time for each assay. The stock solutions were prepared in 50 mL certifuge tubes (Corning life sciences, USA) by dissolving 9.934 mg 4-MUB-N-acetyl-β-D-glucosaminidase dihydrate, 7.707 mg 4-MUB-β-D-xylopyranoside, 8.460 mg 4-MUB-β-D-glucopyranosidase or 6.400 mg 4-MUB-phosphate in 50 mL dissolved water for each assay. The substrate-analogues were attached to 4-methylumbelliferone (4-MUB), which releases fluorescence when it is cleaved by the target enzymes (Hoarfrost and Arnosti 2017; Sinsabaugh et al. 2014).

Four replicate wells were used for each sample, for each assay, by adding 200 µL seawater to black 96-well flat bottom plates (Thermoscientific, USA), then adding 50 µL of the substrate-analogue. Three controls were used in each plate including the substrate control where 200 µL of buffer was added instead of sample, a sample control where 50 µL seawater was added instead of the substrate-analogue and a quench standard where 50 µL MUB was added instead of the substrate-analogue. The plates were then incubated in the dark at 4°C and fluorescence was measured at different times on a Spectramax®

Paradigm Multimode Microplate reader (Molecular Devices, USA) with an emission and excitation wavelength of 445 nm and 360 nm, respectively. The ideal concentration and incubation time were defined as the peak fluorescence, which was determined by plotting fluorescence values at different time points for each of the dilution concentrations.

1.3.2. Extracellular enzyme activity measurements

The potential extracellular enzymatic activities were determined using the optimal concentration and incubation times, as determined in the test assay (Table 3). The same method for the test assays was used to determine potential extracellular enzymatic activities. Enzyme activities were calculated using the following formula (Jackson et al. 2013):

$$Enzyme\ activity\ = \frac{(Sample\ fluorescence - \ Sample\ control)}{(Quench\ standard/\ 0.5\ mol)(0.2\ ml)(incubation\ time)}$$

= nmoles.hr^{-1}.ml^{-1}

Table 3: The specific substrates, concentrations of substrates and incubation times used for each of the extracellular enzyme activity assays.

Enzyme	Substrate	Concentration (µM)	Incubation time (hours)
β-Glucosidase (BG)	4-MUB-β-D-glucopyranoside	500	1
β-Xylosidase (BX)	4-MUB-β-D-xylopyranoside	300	0.25
Alkaline Phosphatase (AP)	4-MUB-phosphate	200	11
Chitinase (NAG)	4-MUB-N-acetyl-β-D-glucosaminide dihydrate	300	0.5

1.4. DNA extraction and agarose gel electrophoresis

DNA was extracted from the filters using the Power Soil DNA extraction kit (Qiagen, USA), following the protocol detailed by the manufacturer with a few minor modifications. A third of each membrane filter was cut into small pieces, using sterile forceps. The DNA was eluted with 50ul of the elution buffer. The concentration of each sample was determined using a Qubit Fluorometer and the 1x High sensitivity (HS) DNA Assay kit (ThermoFischer, USA). A total of 2 µL of the sample was added to 198 µL of the HS Assay working solution, while 10 µL of each standard was added to 190 µL of the HS Assay working solution. The concentrations were then measured using the Qubit 4 Fluorometer (ThermoFishcer, USA).

The quality of the DNA was confirmed using a 0.8% (w/v) agarose gel electrophoresis as detailed previously (Sambrook 1989). The gel was prepared by dissolving 0.32 g agarose powder in 40 mL 1 X TAE buffer (10 mM Tris pH 7.4, 1 mM Na_2.EDTA.2H_2O pH 8.0). A total of 2 µL of each sample was loaded onto the gel with 1 µL GelRed alongside the GeneRuler 1kb DNA ladder (0.5 µg/µL) (ThermoScientific, USA). Electrophoresis was performed at 80 V for 90 minutes and the gel was visualised using a UV transilluminator (BioRad, Germany).

2. Metagenome-assembled genomes (MAGs) curation

2.1. Shotgun metagenomic sequencing

To construct metagenomes from the microbial communities sampled in the SIO, DNA extracts from the 0.4 µm and the 0.2 µm Sterivex filters were pooled for each station. A total of 20 µL of each DNA sample was then sent to Admera Health Biopharma Services

(New Jersey, USA) for sequencing on an Illumina 2x150 instrument (Illumina, USA). The sequencing company performed quality control (QC) on the DNA extracts, where the DNA concentration was determined using a Qubit 2.0 DNA HS Assay (Life Technologies, Grand Island, NY). The DNA quality was determined using a 1% (w/v) standard agarose gel and/or Tapestation genomic DNA Assay (Agilent Technologies, CA, USA). Metagenomic libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina, USA). The libraries were sequenced on the Illumina HiSeq X instrument.

2.2. MAGs assembly

We constructed metagenomic assembled genomes (MAGs) from each sample using the ATLAS-metagenome framework version 2.6a4 (Kieser et al. 2020; White et al. 2017). Briefly, raw reads from 57 metagenomes were quality filtered using the BBTools suite (Bushnell 2014; Kieser et al. 2020). Next, quality filtered reads were assembled using metaSPAdes version 3.15.3 (Nurk et al. 2017) and MAGs were generated using both metabat2 version 0.24.1 (Kang et al. 2019) and maxbin2 version 2.2.7 (Wu et al. 2016). The resultant bins were combined, refined and de-replicated using DAS_Tool version 1.1.3 (Sieber et al. 2018) and dREP (Olm et al. 2017) de-replicated MAGs were assessed for quality and completeness using CheckM version 1.1.3 (Parks et al. 2015). Using previously defined standards, MAGs with completeness > 90% and contamination < 5%, and those with completeness of ≥ 50% and contamination of < 10% were classified as high and medium quality, respectively.

2.3. MAGs classification

The taxonomic identities of medium and high quality MAGs were inferred using the Genome Database Taxonomy toolkit (GTDBK-Tk) version 1.7.0 (Chaumeil et al. 2019).

GToTree version 1.6.11 (Lee 2019) was used to generate phylogenetic trees for 228 high and medium quality MAGs including the 58 archaeal and the 170 bacterial genomes. GToTree uses hidden Markov model (HMM) files containing single-copy genes in a bacteria and archaea to infer evolutionary relationships at the genome-level (Lee 2019). The search was based on at least 30% of the single-copy genes being present in each MAG. The phylogenetic trees were visualised and annotated using the online tool Interactive Tree of Life (iTol) version 5 (Letunic and Bork 2021), where the MAGs were divided into their individual phyla.

3. Gene duplication analysis and assignment of putative paralogs

3.1. Gene duplication analysis

To assess gene duplication events, the protein sequences of the 228 medium to high quality MAGs were predicted using Prodigal version 2.6.3 (Hyatt et al. 2010) with -p meta, -a, and -c parameters. The parameter, '-c', was used in Prodigal to include only complete protein sequences which possess both a start and a stop codon. Following this, a BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE =Proteins) all-against-all search with the upper E-value cut-off of 1×10^{-20} (He and Zhang 2005) was conducted on each of the MAGs using the protein sequence as the subject and query sequences for each MAG to identify putative paralogs. The BLASTp self-hits were filtered using a Bitscore of ≥100 (Gevers et al. 2004; Tekaia and Dujon 1999).

3.2. Assigning function to the putative paralogs

To assess the putative functions of gene duplicates, MAGs were grouped based on their assigned phyla. These MAGs were then separated according to the sampling depth. KofamScan version 102.0 (Aramaki et al. 2020) was used to assign functions to the

putative paralogs from the BLASTp search. The protein sequences of the putative paralogs were searched against the KOfam HMM profile database using their Bitscore values to search for sequence similarity between the genes and functional orthologs, KEGG Orthology (KO) groups, in the database (Aramaki et al. 2020). Only high scoring assignments, which are KO numbers assigned to each gene that have scores above the threshold E-value of 0.01, were included in further analyses.

KEGGDecoder version 1.3 (Graham et al. 2018) was used to determine the completeness of biogeochemical cycling pathways the paralogs were assigned to. The putative paralogs from each MAG were grouped into their separate phyla, which were split into the different sampling depths. KEGGDecoder version 1.3 facilitates the visualisation of the presence of annotated metabolic pathways and shows pathway completeness on a scale between 0 and 1 (Graham et al. 2018).

A heatmap of the completeness of biogeochemical cycling pathways, the paralogs were affiliated with, was generated in RStudio Desktop version 2022.02.2+485 (R Core Team 2013) using the pheatmap software package available on Github (https://github.com/raivokolde/pheatmap.git). Only pathways pertaining to the cycling of carbon, nitrogen, phosphorus, and methane were included.

Shapiro-Wilk's test (Shapiro and Wilk 1965) was first used for the number of putative paralogs per depth, station, and phylum to determine if the number of putative paralogs for each of these parameters was normally distributed. Levene's test (Levene 1960) was performed using the car package version 3.1-0 for the number of putative paralogs per depth, station, and phylum. This analysis was used to determine if the variances in the number of putative paralogs, for each of these parameters, was homogenous. A Kruskal-

Wallis test was used to determine if differences between the samples per depth, station and phylum were significant, since the number of putative paralogs followed an abnormal distribution (Kruskal and Wallis 1952). Boxplots, showing the number of putative paralogs per depth, position and Phylum, were generated using the ggplot2 version 3.3.6 package software and the ggboxplot function (Wickham 2016). A bar graph of the number of putative paralogs per station was generated using the ggplot2 and the ggplot function (Wickham 2016). A bar graph of the number of paralogs per station belonging pathways pertaining to the cycling of carbon, nitrogen, phosphorus, and methane was generated using the ggplot2 and the ggplot function (Wickham 2016)

4. Statistical analyses

To determine statistically significant differences between samples, statistical analyses were performed using the R statistical package version R 3.3.0+ and R Studio version 2022.07.2+576 (R Core Team 2013). To determine if the extracellular enzyme activity, concentrations of nutrients and the number of putative paralogs for each parameter (depth, station and position) was normally distributed. We used the R statistical package version R 3.3.0+ and the Shapiro-Wilk's test (Shapiro and Wilk 1965). Levene's test (Levene 1960) was performed for each nutrient, extracellular enzyme assay and the number if putative paralogs per depth, station, and position (inter-island or open ocean) using car package version 3.1-0 to determine variance for each parameter was homogenous. A Kruskal-Wallis test was used to determine the significance of differences between the samples per depth, station and position, since the nutrient concentrations, extracellular enzyme activities and the number of putative paralogs were not normally distributed or variances in concentration, activities and number of putative paralogs were

not homogenous (Kruskal and Wallis 1952). Boxplots of the different nutrient concentrations, enzyme activities and number of putative paralogs per depth and position were generated using the ggplot2 version 3.3.6 package software and the ggboxplot function (Wickham 2016). A bar graph of the nutrient concentrations, extracellular enzyme activities and number of putative paralogs per station was generated using ggplot2 and the ggplot function (Wickham 2016). A bar graph of the number of putative paralogs per phylum was also generated using ggplot2 and the ggplot function (Wickham 2016).

To determine the correlation between the number of paralogs encoding an extracellular enzyme and enzyme activity, Spearman's correlation was calculated using the R statistical package version R 3.3.0+ and R Studio version 2022.07.2+576 (R Core Team 2013). The ggscatterstat function in the ggstatsplot package version 0.11.0.9000 (Patil 2021) was used to calculate Spearman's correlation for the non-parametric data, as well as to plot scatterplots for each of the four extracellular enzymes (BG, BX, NAG and AP). AP and NAG activity outliers that were previously removed were also removed from the scatter plots and Spearman's correlation calculation. The ggscatterstat function in the ggstatsplot package version 0.11.0.9000 (Patil 2021) was also used to determine the relationship between the number of MAGs belonging to a phylum and the total number of paralogs possessed by the phylum.

5. A Genus-level pangenomic analysis of high and medium quality MAGs

Prokka version 1.14.5 (Seemann 2014) was used to determine the protein coding sequences of each of the MAGs. Then Roary: the Pan genome pipeline version 3.11.2 (Page et al. 2015) was used to calculate the pangenome of the 228 high and medium

quality MAGs, using their Prokka annotated protein sequences. A 75% sequence identity was used as the cut-off for core genes (Oliver et al. 2018; Plissonneau et al. 2018).

The pangenome was visualised in RStudio Desktop 2022.02.2+485 (R Core Team 2013) using the ggplot2 version 3.3.6 software package (Wickham 2016) to generate a bar graph of the number of paralogs contained in each pangenome component (core genes, soft-core genes, and accessory genes).

A genus-level pangenomic analysis of SAR324 was carried out using Roary: the Pan genome pipeline version 3.11.2 (Page et al. 2015) to calculate the pangenome of the 16 MAGs using a 75% sequence identity was used as the cut-off for core genes.

6. References

Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber JA. 2017. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat. Commun. 8(1)**:** 1-11. https://doi.org/10.1038/s41467-017-01228-6.

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. Bioinform. 36(7)**:** 2251-2252. https://doi.org/10.1093/bioinformatics/btz859.

Bushnell B. 2014. BBTools software package. URL http://sourceforge. net/projects/bbmap. 578(579.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinform. 36(6)**:** 1925-1927. https://doi.org/10.1093/bioinformatics/btz848.

Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12(4)**:** 148-154. https://doi.org/10.1016/j.tim.2004.02.007.

Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a globally-distributed bacterial phototroph. ISME J. 12(7)**:** 1861-1866. https://doi.org/10.1038/s41396-018-0091-3.

Grasshof K. 1976. Methods of seawater analysis. Verlag Chemie, Weinheim and New York.

He X, Zhang J. 2005. Gene complexity and gene duplicability. Curr. Biol. 15(11)**:** 1016-1021. https://doi.org/10.1016/j.cub.2005.04.035.

Hoarfrost A, Arnosti C. 2017. Heterotrophic extracellular enzymatic activities in the Atlantic Ocean follow patterns across spatial and depth regimes. Front. Mar. Sci. 4(200. https://doi.org/10.3389/fmars.2017.00200.

Holmes RM, Aminot A, Kérouel R, Hooker BA, Peterson BJ. 1999. A simple and precise method for measuring ammonium in marine and freshwater ecosystems. Can. J. Fish. Aquat. Sci. . 56(10)**:** 1801-1808. https://doi.org/10.1139/f99-128.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11(1)**:** 1-11. https://doi.org/10.1186/1471-2105-11-119.

Jackson CR, Tyler HL, Millar JJ. 2013. Determination of microbial extracellular enzyme activity in waters, soils, and sediments using high throughput microplate assays. J. Vis. Exp. JOVE. 80)**:** https://dx.doi.org/10.3791/50399.

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. https://doi.org/10.7717/peerj.7359.

Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA. 2020. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinform. 21(1)**:** 1-8. https://doi.org/10.1186/s12859-020-03585-4.

Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47(260)**:** 583-621. https://doi.org/10.1080/01621459.1952.10483441.

Le PT, Makhalanyane TP, Guerrero LD, Vikram S, Van de Peer Y, Cowan DA. 2016. Comparative metagenomic analysis reveals mechanisms for stress response in hypoliths from extreme hyperarid deserts. Genome Biol. Evol. 8(9)**:** 2737-2747.

Lee MD. 2019. GToTree: a user-friendly workflow for phylogenomics. Bioinform. 35(20)**:** 4162-4164. https://doi.org/10.1093/bioinformatics/btz188.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49(W1)**:** W293-W296. https://doi.org/10.1093/nar/gkab301.

Levene H. 1960. Robust tests for equality of variances. Contributions to probability and statistics. 278-292.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27(5)**:** 824-834. https://doi.org/10.1101/gr.213959.116.

Oliver A, Kay M, Cooper KK. 2018. Comparative genomics of cocci-shaped *Sporosarcina* strains with diverse spatial isolation. BMC Genom. 19(1)**:** 1-17. https://doi.org/10.1186/s12864-018-4635-8.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11(12)**:** 2864-2868.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinform. 31(22)**:** 3691-3693. https://doi.org/10.1093/bioinformatics/btv421.

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(7)**:** 1043-1055. https://doi.org/10.1101/gr.186072.114.

Patil I. 2021. Visualizations with statistical details: The'ggstatsplot'approach. J. Open Source Softw. 6(61)**:** 3167.
Plissonneau C, Hartmann FE, Croll D. 2018. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 16(1)**:** 1-16.
R Core Team. 2013 R: A language and environment for statistical computing, pp. R Foundation for

Statistical Computing, Vienna, Austria.
Sambrook J. 1989. Gel electrophoresis of DNA. Molecular cloning a laboratory manual. 6.36-36.45.
Ocean Data View [internet]. 2023. Available from: http://odv.awi.de
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinform. 30(14)**:** 2068-2069. https://doi.org/10.1093/bioinformatics/btu153.
Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). Biometrika. 52(3/4)**:** 591-611. https://doi.org/10.2307/2333709.
Sieber CM, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiol. 3(7)**:** 836-843. https://doi.org/10.1038/s41564-018-0171-1.
Sinsabaugh RL, Belnap J, Findlay SG, Shah JJF, Hill BH, Kuehn KA, Kuske CR, Litvak ME, Martinez NG, Moorhead DL. 2014. Extracellular enzyme kinetics scale with resource availability. Biogeochemistry. 121(287-304. https://doi.org/10.1007/s10533-014-0030-y.
Tekaia F, Dujon B. 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. J. Mol. Evol. 49(591-600. https://doi.org/10.1007/PL00006580.
Tully BJ, Wheat CG, Glazer BT, Huber JA. 2018. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. ISME J. 12(1)**:** 1-16.
White RA, Brown J, Colby S, Overall CC, Lee J-Y, Zucker J, Glaesemann KR, Jansson C, Jansson JK. 2017. ATLAS (Automatic Tool for Local Assembly Structures)-a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data. PeerJ Prepr. 5(e2843v2841. https://doi.org/10.7287/peerj.preprints.2843v1.
Wickham H. 2016. gplot2: Elegant graphics for data analysis. New York: Springer-Verlag New York.
Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinform. 32(4)**:** 605-607. https://doi.org/10.1093/bioinformatics/btv638.

Chapter 3

The extent and putative functions of gene duplicates in the Subantarctic Indian Ocean

1. Abstract

The Subantarctic Indian Ocean (SIO) is a vital carbon sink and a crucial ecosystem for the sequestration of atmospheric carbon dioxide (CO_2), with microbial communities, including bacteria and archaea, playing a pivotal role in carbon sequestration and the biological pump. However, there is a significant knowledge gap regarding the diversity and functional pathways of microbial communities involved in biogeochemical recycling, particularly in poorly studied marine environments such as the SIO. To address this gap, we investigated biogeochemical cycling and microbial communities in the SIO, focusing on the Subantarctic Prince Edward Islands (PEIs).

We used extracellular enzymatic assays for β-glucosidase (BG), β-xylosidase (BX), N-acetylglucosaminidase (NAG), and alkaline phosphatase (AP) to demonstrate evidence of biogeochemical cycling in the PEIs. We also constructed metagenome-assembled genomes (MAGs) to investigate the extent and potential function of gene duplications in the SIO. Our results show that the SIO supports a diverse consortium of microbial communities, with Proteobacteria, Actinobacteria, and Verrucomicrobiota possessing higher levels of gene duplications. Interestingly, several duplicated genes, including those encoding extracellular enzymes, appear to facilitate biogeochemical cycling in this extreme environment lacking sufficient nitrate, silicate, and phosphate.

We found several duplicated genes encoding enzymes involved in carbon, phosphorous, and nitrogen metabolism, as well as methanogenesis, with these duplications being phyla specific. Taxa affiliated with Proteobacteria, Bacteroidota, Planctomycetota, Thermoplasmatota, Cyanobacteria, and SAR324 showed higher counts of copied genes responsible for these biogeochemical cycles. Furthermore, a pangenomic analysis of

SAR324 revealed five core paralogs responsible for carbon and nitrogen metabolism. The location of gene duplicates suggests their role in the evolution of microbial lineages, with core genes potentially being critical to the evolution of the phylum, and several accessory genes responsible for phosphorous metabolism appearing to adapt specific SAR324 lineages to the PEIs.

2. Introduction

There is mounting evidence that microbial communities play a crucial role in biogeochemical cycling in the ocean (Glöckner et al. 2012; Phoma et al. 2018; Rush and Sinninghe Damsté 2017). Studies conducted in the Southern Ocean, an important carbon sink, have shown that despite oligotrophic conditions and trace element limitation, the region hosts diverse microbial communities (Glöckner et al. 2012; Phoma et al. 2018). However, the mechanisms and drivers of phylogenetic diversity and the implications of gene duplications on these communities remain largely unexplored (Hermes et al. 2019; Phoma et al. 2018). Moreover, the biogeographic trends and variability across different environments, including marine depths and distances from island masses, require further investigation.

Bacteria are ideal model organisms for studying adaptation and gene duplication events due to their relatively short generation times (Lenski and Travisano 1994), which facilitate comparisons between ancestral and derived organisms over several time intervals (Barrick et al. 2009; Lenski et al. 1991; Lenski and Travisano 1994; Riehle et al. 2001; Wiser et al. 2013). Several studies have largely focused on gene duplications at genome scales (Chirakul et al. 2019; Kim et al. 2016; Riehle et al. 2001). For instance, the study by Chirakul et al. (2019) showed that the duplication of a gene encoding PenA class A β-lactamase was responsible for increased β-lactamase activity, providing antibiotic resistance in *Burkholderia pseudomallei*. Although these studies have provided important insights regarding gene duplication events in specific genomes, their prevalence and association at community scale remain unclear. In addition, the biogeographic trends, variability across different environments (for instance, amongst different marine depths or

distances from island masses) remains largely unexplored (Elliott et al. 2013; Sheridan et al. 2020).

Gene duplication is an important mechanism that allows organisms to adapt to environmental selective pressures (Elliott et al. 2013; Kondrashov 2012). Gene duplicates contribute to the evolution of genomes by providing additional genetic material for natural selection or neutral genetic drift to act on, causing evolution at a molecular level (Ohno 1970; Raes and Van de Peer 2003; Serres et al. 2009). Consequently, gene copies may lead to genomic variation that results in the evolution of secondary metabolites (Fewer and Metsä-Ketelä 2020; Moghe and Last 2015), the gene dosage effect due to increased enzymatic activities (Conant and Wolfe 2008; Moghe and Last 2015), microbial diversity (Virdi and Sachdeva 2005), niche specialisation (Sheridan et al. 2020) and antibiotic resistance (Chirakul et al. 2019). There is some preliminary evidence that gene duplications, as a mode of adaptation, may be important in marine microbial communities (Kim et al. 2016; Sheridan et al. 2020). It is reasonable to predict that due to the confluence of environmental variables, marine waters in the South Indian Ocean (SIO) and the surrounding Subantarctic islands may support microbial communities with a propensity for duplication (Hooper and Berg 2003; Jordan et al. 2001; Lutjeharms 2006; Phoma et al. 2018; Riehle et al. 2001).

The Subantarctic PEIs are located within the vicinity of Antarctic Circumpolar Current (ACC) in the Southern Ocean. These islands, located in the south-west Indian sector of the Southern Ocean, are approximately 2 000 km south of the east coast of South Africa (Belkin and Gordon 1996; Venkatachalam et al. 2019). The PEIs and the ACC are positioned within the Antarctic Polar Frontal Zone (APFZ), which links northern warm

subtropical waters to southern cold Subantarctic waters (Ansorge and Lutjeharms 2002; Belkin and Gordon 1996; Venkatachalam et al. 2019). Due to legislative amendments and the goal of protecting the unique species, habitats and ecosystem processes biodiversity, the islands are regarded as core marine protected area (MPA) (Kirkman et al. 2021; Kirkman et al. 2016; Lombard et al. 2007). Previous studies have shown some evidence confirming that the region may facilitate monitoring and predictive capacity helping to advise future management on climate change in this important carbon sink, the SIO (Glöckner et al. 2012; Lombard et al. 2007).

Marine microbial communities are shaped by factors such as the availability of light (Cory et al. 2014; Phoma and Makhalanyane 2021) and hence depth (Phoma et al. 2018), salinity levels (Morrissey et al. 2014), temperature (Sunagawa et al. 2015), as well as nutrient availability (Henley et al. 2020; Phoma et al. 2018). There have been studies showing that the proximity to land masses may influence nutrient concentrations, specifically nitrogen, through a process known as the “Island Mass Effect” (IME) (Boden 1988; Seitzinger et al. 2005). The IME occurs when nutrients are supplied to surrounding waters through run-off from the islands, increasing primary productivity which results in phytoplankton blooms (Boden 1988; Treasure et al. 2015). This effect has specifically been observed in the PEIs, resulting in elevated productivity levels in waters that are generally limited in iron and silica (Glöckner et al. 2012; Henley et al. 2020; Perissinotto and Rae 1990). The convergence of physicochemical factors may act as a selective pressure for macrofauna and microbial communities. An example of this was seen in the PEIs, where depth-dependant physiochemical factors led to Gammaproteobacteria dominating these nutrient-rich waters (Phoma et al. 2018). The PEIs thus act as a good

model to study whether gene duplication events play a role in the adaptation of organisms to specialised environments.

Previous studies have reported the presence of several taxa with known chemolithoautotrophic lifestyles, such as carbon fixing and sulphur-oxidizing Oceanospirillales, Thaumarchaeota and anammox Planctomycetes in the SIO (Phoma et al. 2018). Chemolithoautotrophs use carbon dioxide from their environment, as a carbon source for metabolic processes, as well as inorganic compounds such as nitrogen and sulphur as energy resources for these metabolic processes (Farías et al. 2009; Hawley et al. 2014; Schunck et al. 2013; Swan et al. 2011). Gammaproteobacteria are largely responsible for carbon metabolism in PEI marine microbial communities (Phoma et al. 2018). These taxa may act as carbon sinks, or a source of organic matter, where they mediate the transfer fixed carbon along trophic levels (Pedler et al. 2014; Zehr and Kudela 2011). This fixation, by using inorganic compounds such as sulphur or nitrate compounds, provides a key source of energy and may drive oceanic carbon dioxide sequestration (GÜnter et al. 2008; Phoma et al. 2018; Swan et al. 2011; Zhao et al. 2020). However, despite the known importance of microbes as mediators of these processes, little is known regarding the role that gene duplication events may play in biogeochemical cycling (Fischer et al. 2016; Kondrashov 2012; Sánchez-Baracaldo et al. 2022). This knowledge deficit is especially true for understudied regions such as the PEIs where evolutionary processes including the presence of potential paralogs responsible for these cycles, remain unknown.

The specific factors which influence duplication events and their ecological trends are not well understood (Gevers et al. 2004; Rensing 2014). This is despite numerous studies on

gene duplications focused on specific bacterial species (Chirakul et al. 2019; Kim et al. 2016; Riehle et al. 2001). Given the potential implications that such duplications may have on biogeochemical recycling it is crucial to provide related baseline data. This is particularly important for understudied regions of the ocean such as the PEIs where microbial communities are thought to play disproportionate roles in the recycling of key nutrients. This study assesses the extent, and potential function, of gene duplicates in marine microbial communities from the SIO in waters surrounding the PEIs. The results are expected to fill the knowledge gap regarding the factors that influence microbial communities in this region, and the factors which influence gene duplications in these communities.

3. Results

3.1. Physiochemistry of the PEI

3.1.1. Physical characteristics

The CTD (conductivity, temperature and depth) rosette multi-sampler (Sea-Bird SBE-911 plus V2 CTD system; Sea-Bird Electronics, Inc., USA) measured several variables including salinity (PSU), temperature (°C), conductivity (S/m), Oxygen (ml/L), pressure (db) and depth (m) at three depths (Supplementary data, Table S2). Samples were retrieved from 14 inter-island stations belonging to the T-transect and the X-M-transect (n=35). In general, pressure increased with depth (Figure 3a) and showed a significant difference between depths (Table 4) (Kruskal-Wallis test, $p < 0.001$). A significant difference in temperature and conductivity between depths was observed (Table 4) (Kruskal-Wallis test, $p < 0.05$), with the highest temperatures (Figure 3b) and conductivity (Figure 3d) seen in Fmax samples. There was a significant difference (Kruskal-Wallis test,

p < 0.001) in the salinity of the samples for each depth (Table 4), with Deep samples found to have the highest salinity (Figure 3c). There was a significant difference in the levels of oxygen between depths (Kruskal-Wallis test, p < 0.001) (Table 4), with samples from Fmax having the highest levels of oxygen and samples from OMZ having the lowest levels of oxygen (Figure 3e). The depth of the samples ranged from 5 m, which were samples collected from the surface along the Crossroads transect (CR) and the S-transect, to a maximum depth of 4077,16 m for the sample collected at the deep depth at station T-08A. Our analysis excluded the seven inter-island stations which had no OMZ (Figure 3).

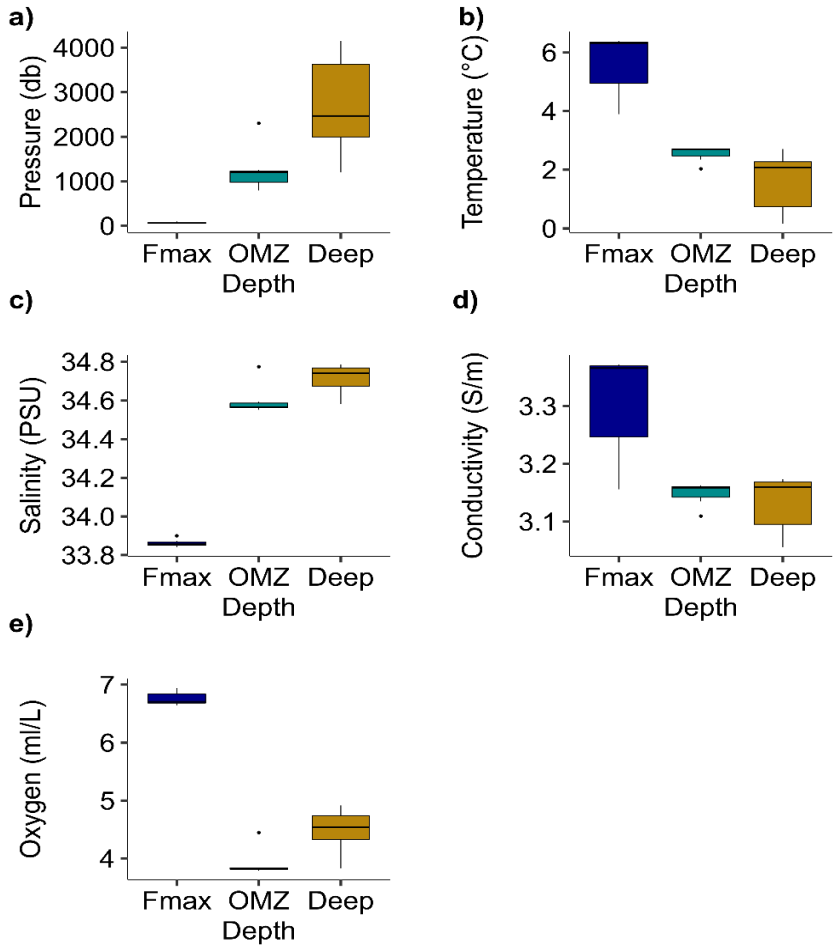

Figure 3: Boxplots of the physiochemical variables according to depth including a) pressure, b) temperature, c) salinity, d) electrical conductivity, e) oxygen measured at all

the inter-island stations on the ‘T-transect’ and the ‘X-M-transect’. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

Five water masses were sampled (Figure 4), including subantarctic Surface Water (SASW), Antarctic Surface Waters (AASW), Antarctic Intermediate Water (AAIW), Circumpolar Deep Water (CDW) and Antarctic Bottom Water (AABW) based on their temperature and salinity readings (Emery 2001). Antarctic surface waters are characterised by colder surface temperatures, while subantarctic surface waters are characterised by warmer temperatures and lower salinity measurements (Emery 2001).

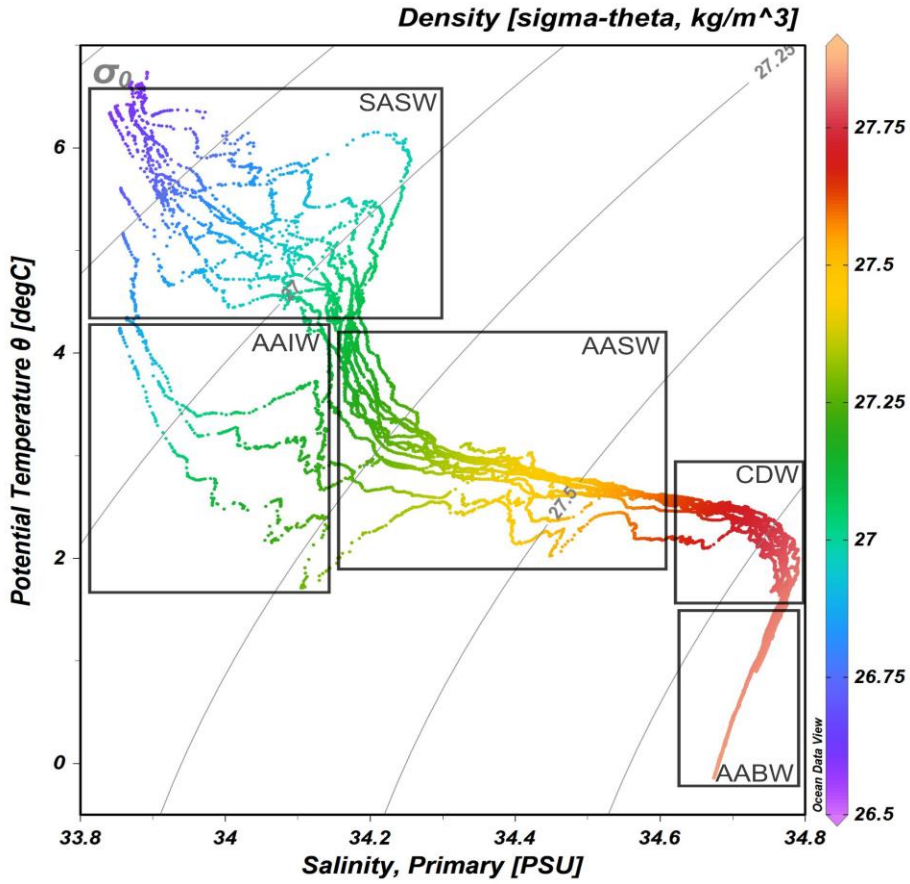

Figure 4: A temperature-salinity plot of stations showing the water masses sampled along the 'T-transect' (Plot by Gavin Tutt, Tesha Toolsee Marcel van den Berg, Oceans & Coasts Research at the Department of Forestry, Fisheries and the Environment). SASW = subantarctic Surface Water; AASW = Antarctic Surface Waters; AAIW = Antarctic Intermediate Water; CDW = Circumpolar Deep Water; AABW = Antarctic Bottom Water.

Table 4: Kruskal-Wallis non-parametric test of significance for each physical factor for each depth. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.

Physical factor	Chi-squared value	Degrees of freedom	P-value	Significance
Pressure	16.482	2	0.0002636	***
Salinity	15.636	2	0.0004024	***
Temperature	15.384	2	0.0004565	***
Oxygen	16.186	2	0.0003057	***
Conductivity	9.3729	2	0.009219	*

3.1.2. Chemical characteristics

The concentration of the nutrients including nitrate (NO_3^-), nitrite (NO_2^-), nitrate + nitrite ($NO_3^- + NO_2^-$), ammonium (NH_4^+), silicate (SiO_3^{2-}) and phosphate (PO_4^{3-}) (Supplementary data, Table S3), demonstrated a significant depth-dependant trend (Kruskal-Wallis test, $p < 0.001$) (Table 5). The concentration of nitrate (Figure 5a), nitrate + nitrite (Figure 5c), phosphate (Figure 5e) and silicate (Figure 5f) were highest in samples recovered from

OMZ, while those of nitrite (Figure 5b) and ammonium (Figure 5d) were highest in Fmax samples. Nitrate (Figure 5a), nitrate + nitrite (Figure 5c), phosphate (Figure 5e) and silicate (Figure 5f) concentrations were lowest from samples collected at 5 m below the surface, while nitrite (Figure 5b) and ammonium (Figure 5d) concentrations were lowest from OMZ inter-island samples. Nitrate and nitrite were combined (nitrate + nitrite) during the analysis with the SEAL AA500 AutoAnalyzer and auto-analysis protocols (Grasshof 1976). Fmax, OMZ and Deep samples were taken from the inter-island transects, while 5 m samples were taken from the open ocean transects.

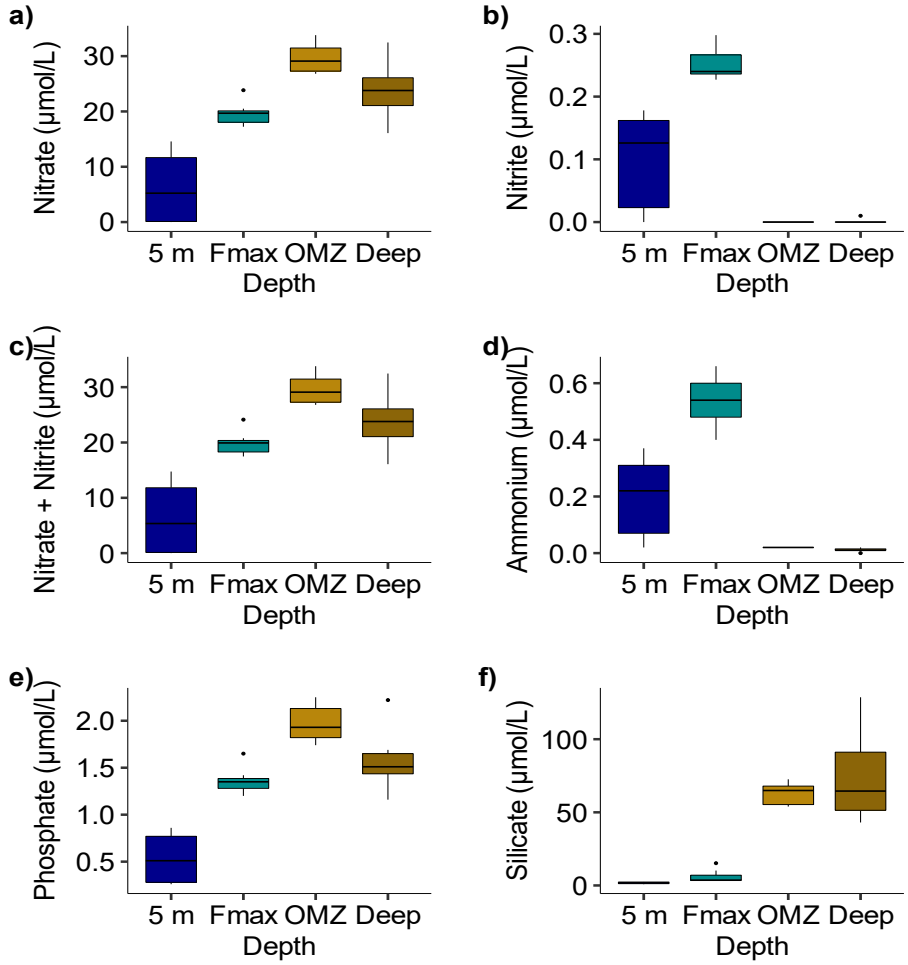

Figure 5: Boxplots showing nutrient concentrations at each depth for all samples for a) nitrate, b) nitrite, c) nitrate + nitrite measured as combined nitrate and nitrite, d)

ammonium, e) phosphate and f) silicate. Samples from the 'Crossroads line' were excluded since there were no nutrient concentration measurements for those samples. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

Table 5: Kruskal-Wallis non-parametric test of significance was applied to test the differences linked to each nutrient at each depth and station. In addition, the Wilcoxon test for significance was also applied to test differences in each nutrient at each position. The significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.

Nutrient	Parameter	Chi-squared value	Degrees of freedom	P-value	Significance
Nitrate	Depth	24.829	3	1.68E-05	***
	Station	26.955	18	7.99E-02	
	Position	10.045	1	3E-05	***
Nitrite	Depth	28.403	3	2.99E-06	***
	Station	7.7084	18	9.83E-01	
	Position	0.028246	1	0.5	
Nitrate + Nitrite	Depth	24.61	3	1.86E-05	***
	Station	26.966	18	7.96E-02	
	Position	10.045	1	3E-05	***
Ammonium	Depth	22.036	3	6.41E-05	***
	Station	11.072	18	8.91E-01	
	Position	0.0038505	1	0.4	
Phosphate	Depth	26.095	3	9.11E-06	***
	Station	27.878	18	6.39E-02	
	Position	12.523	1	7.2E-04	***
Silicate	Depth	27.121	3	5.55E-06	***
	Station	25.317	18	1.16E-01	
	Position	10.838	1	3E-05	***

A Wilcoxon test for significance confirmed a significant difference in the concentration of nitrate, and nitrate + nitrite ($p < 0.05$), as well as phosphate and silicate ($p < 0.001$) (Table 5) between open ocean and inter-island samples. Nitrate (Figure 6a), nitrate + nitrite (Figure 6c), phosphate (Figure 6e) and silicate (Figure 6f) are highest from the inter-island samples, while nitrite (Figure 6b) and ammonium (Figure 6d) concentrations are spread over open ocean and inter-island samples with inter-island samples having slightly higher concentrations that are not significantly different (Table 5).

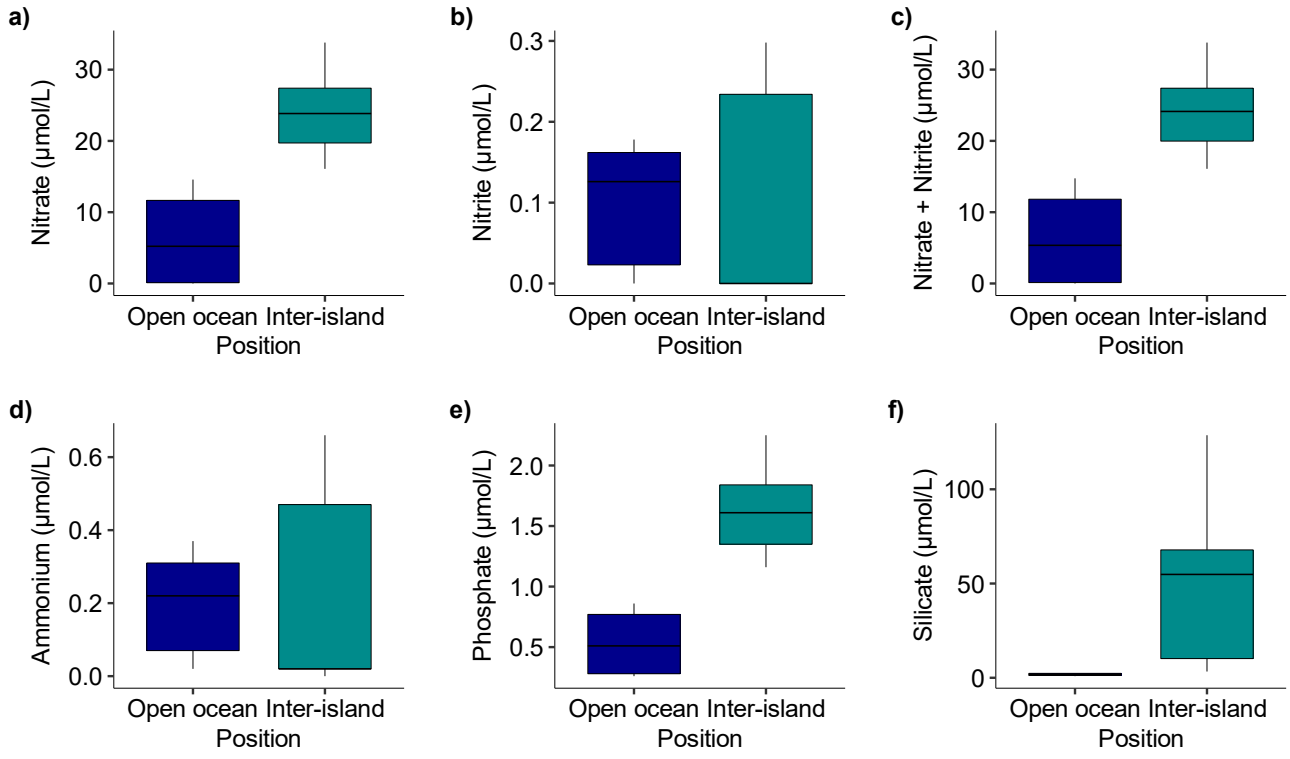

Figure 6: Boxplots showing nutrient concentrations at each position (open ocean and inter-island) for all samples for a) nitrate, b) nitrite, c) nitrate + nitrite measured as combined nitrate and nitrite, d) ammonium, e) phosphate and f) silicate. Samples from the 'Crossroads line' were excluded since there were no nutrient concentration

measurements for those samples. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

A significant difference was not observed in nutrient concentrations for each station (Table 5). However, samples from the T-transect and the X-M-transect had higher nitrate + nitrite (Figure 7a), ammonium (Figure 7b), phosphate (Figure 7c) and silicate (Figure 7d) concentrations. The T-transect and X-M-transect were sample between and around the PEIs, whereas samples from the S-transect were sampled from the open ocean.

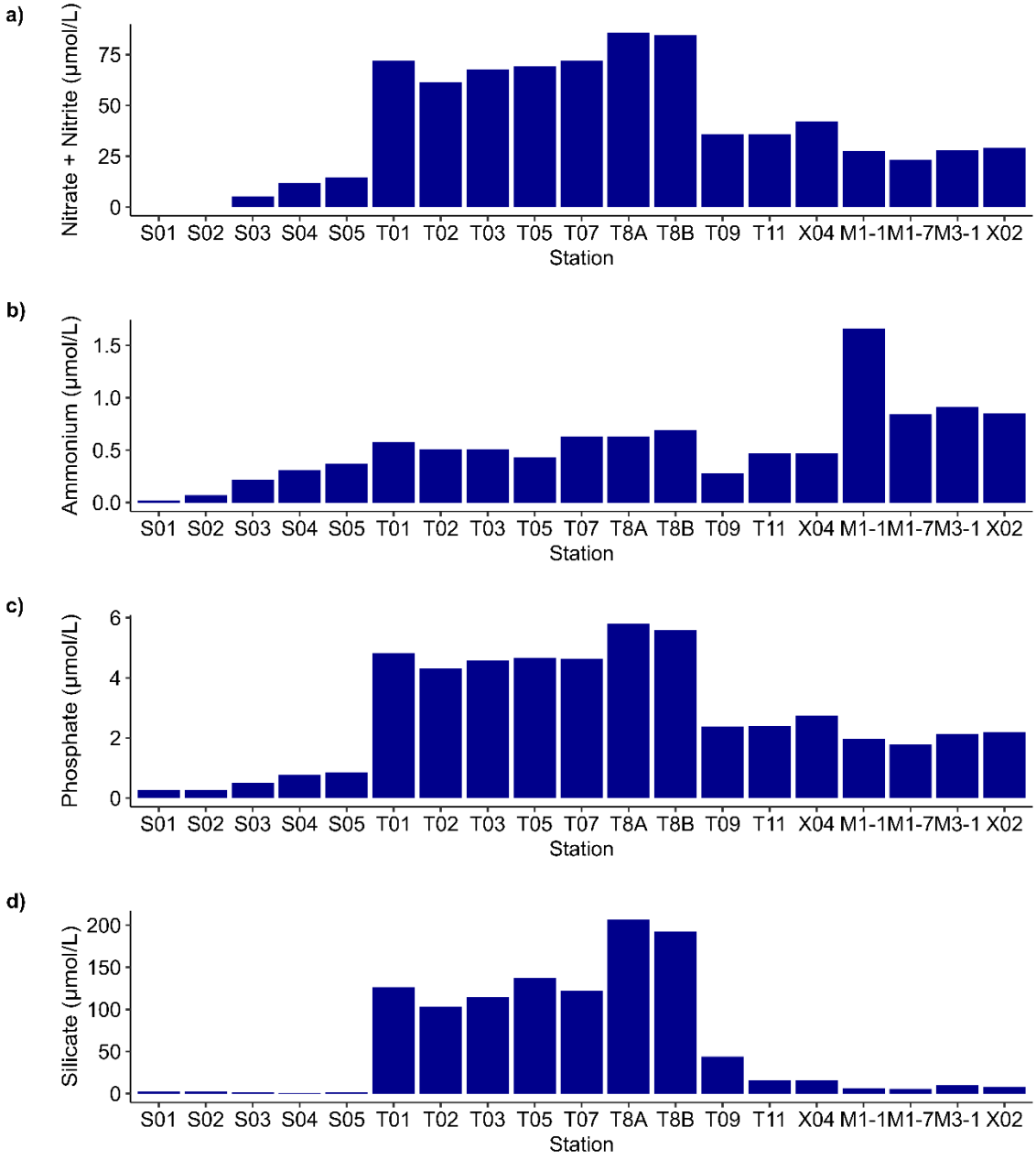

Figure 7: Bar graphs showing nutrient concentrations at each station for all samples for a) nitrate + nitrite measured as combined nitrate and nitrite, b) ammonium, c) phosphate and d) silicate. ‘S’ and ‘CR’ represent samples taken from the underway system between South Africa and Marion Island, while ‘T’ represents samples from the ‘T-transect’ and ‘X’ and ‘MT’ represent samples from the ‘X-M transect’ taken around the PEIs. Samples from

the ‘Crossroads line’ were excluded since there were no nutrient concentration measurements for those samples.

3.2. Extracellular enzymatic activity analysis

We assessed the extracellular enzymatic activities of four enzymes linked to carbon, nitrogen and phosphorus acquisition. Outliers linked to NAG activity from the inter-island stations MT1-1, T1, T2 and T8B, as well as open ocean stations S1, S2, S3, S4 and S5 were removed (Supplementary data, Table S4). Outliers linked to AP activity (Supplementary data, Table S4) from the inter-island stations MT1-7, T1, T2 and T5, as well as open ocean stations CR05 and CR11 were also removed.

The potential extracellular enzyme activities demonstrated a depth-dependent trend, with the exception of activities linked to BX (Figure 8b, Table 6) ($p < 0.05$). The highest mean BX activity was observed in Fmax samples, while the lowest was detected in Deep samples (Figure 8b). The activity of BG did not show a specific trend across the four depths (Figure 8a), while the highest mean NAG (Figure 8c) and AP activity (Figure 8d) was observed in OMZ samples. Potential extracellular enzymatic activities did not vary significantly between samples from the open-ocean and those collected from inter-island zones (Figure 9, Table 6). The mean BX (Figure 9b) and NAG activity (Figure 9c) was higher from inter-island samples, while mean BG (Figure 9a) and AP activity (Figure 9d) was higher from open ocean samples. The activity of NAG was significantly different for each station (Figure 10c, Table 6) ($p < 0.05$), with stations T8A and T03 with the highest NAG activity. Overall, stations from the T-transect had higher BG (Figure 10a), BX (Figure 10b) and AP activity (Figure 10d).

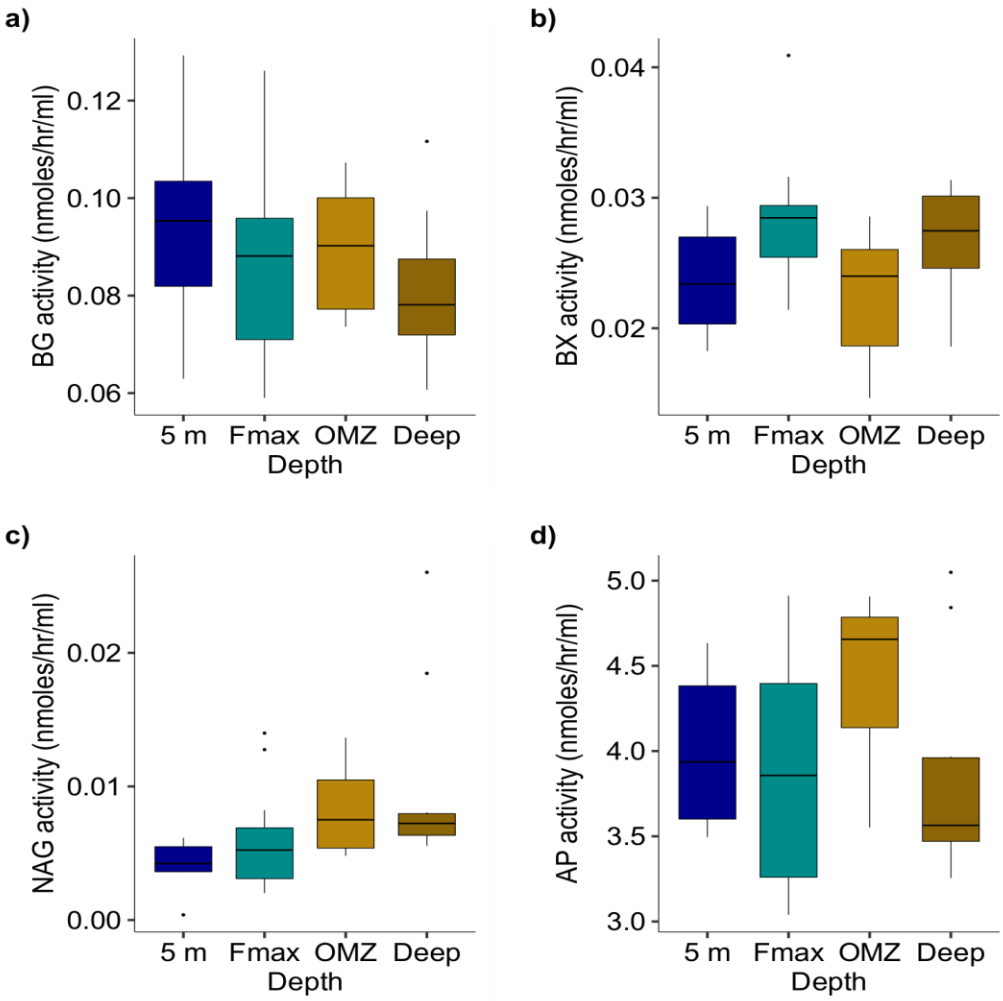

Figure 8: Boxplots showing potential extracellular enzyme activity at each depth for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

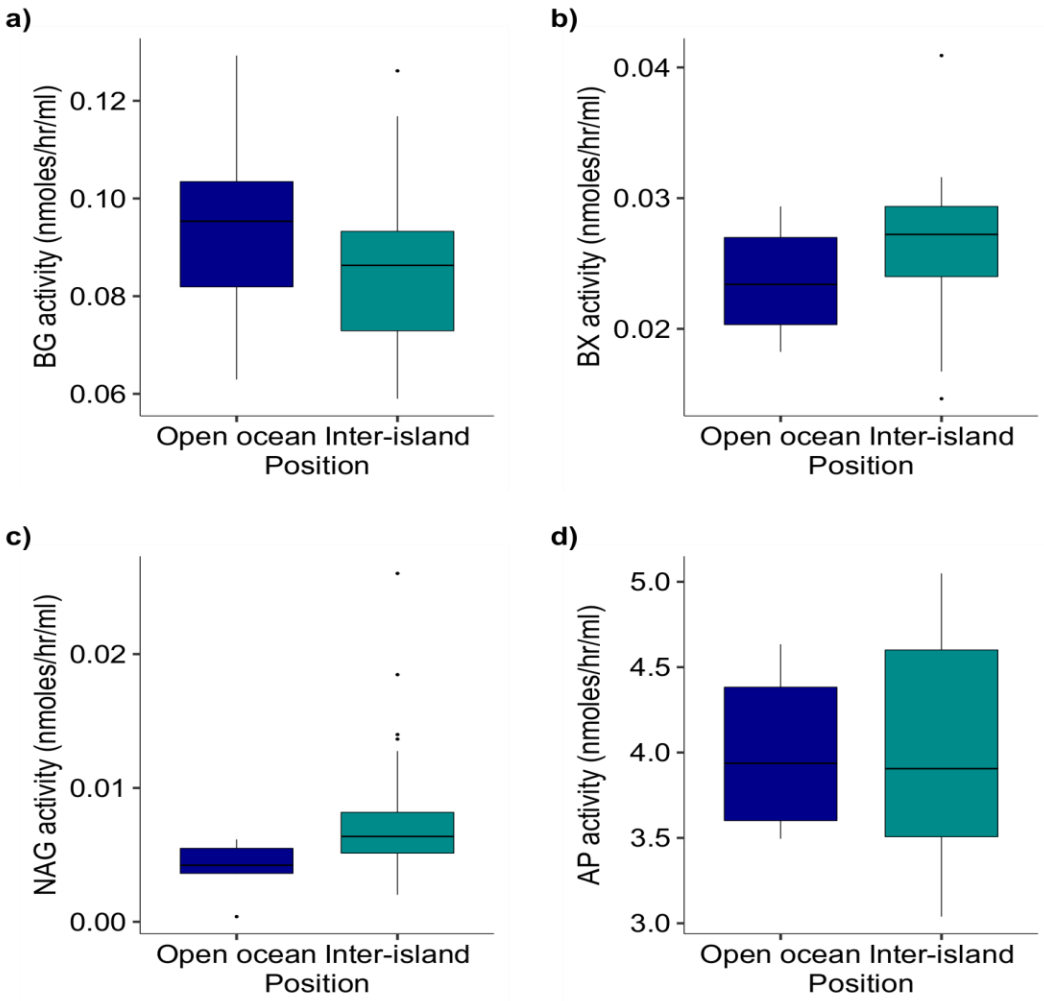

Figure 9: Boxplots showing potential extracellular enzyme activities at each position (open ocean and inter-island) for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

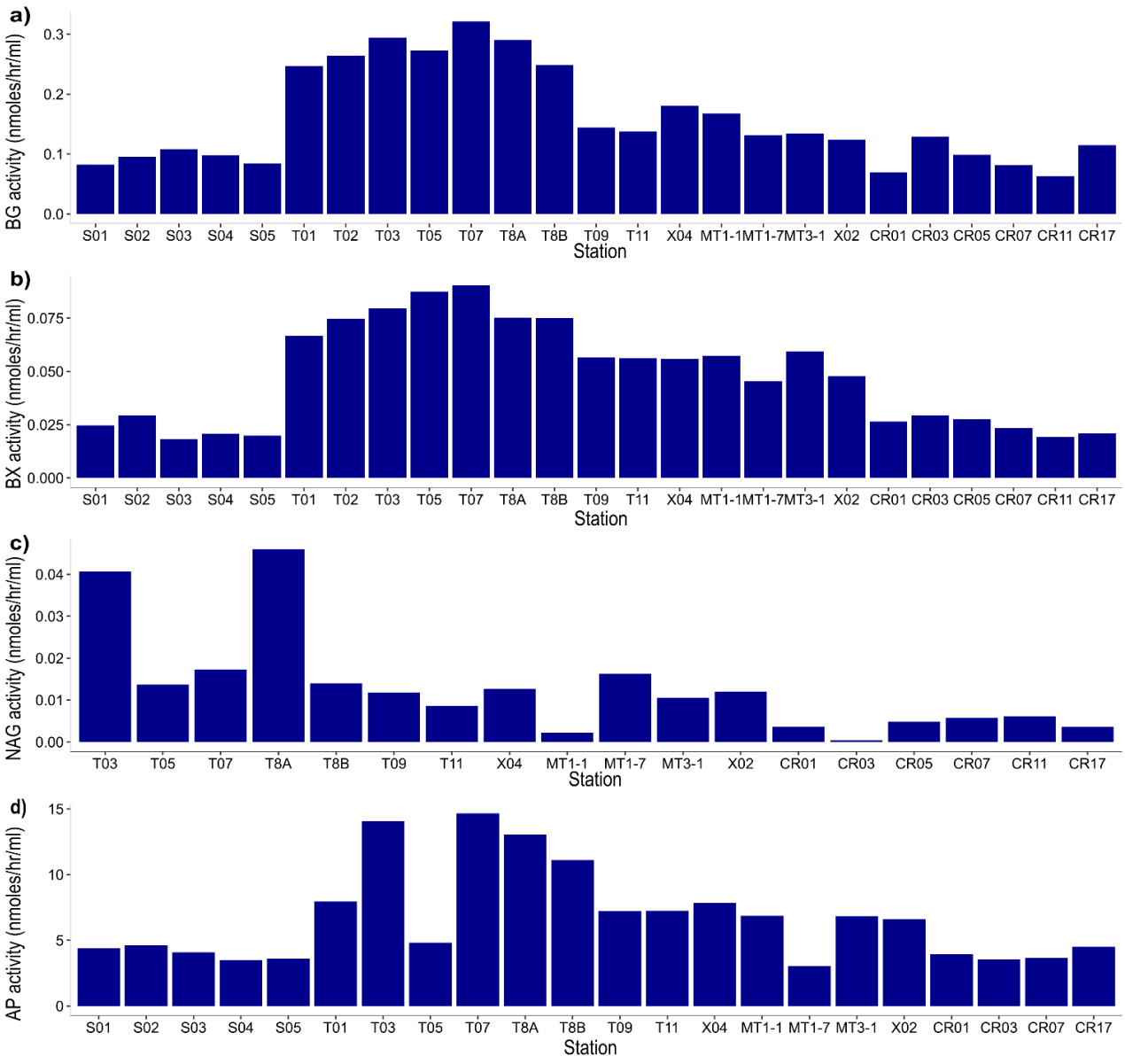

Figure 10: Bar graphs showing potential extracellular enzyme activities at each station for all samples for a) β-glucosidase (BG), b) β-xylosidase (BX), c) N-acetylglucosaminidase (NAG) and d) alkaline phosphatase (AP). ‘S’ and ‘CR’ represent samples taken from the underway system between South Africa and Marion Island, while ‘T’ represents samples from the ‘T-transect’ and ‘X’ and ‘MT’ represent samples from the ‘X-M transect’ taken around the PEIs. Outliers (Supplementary table S4) were excluded for N-acetylglucosaminidase and alkaline phosphatase.

Table 6: Kruskal-Wallis non-parametric test of significance for each extracellular enzyme activity assay for each depth and station, as well as Wilcoxon test of significance for each extracellular enzyme activity assay for each position. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.

Enzyme	Parameter	Chi-squared value	Degrees of freedom	P-value	Significance
β-Glucosidase (BG)	Depth	3.4611	3	3.26E-01	
	Station	34.672	24	7.34E-02	
	Position	1.7589	1	1.90E-01	
β-Xylosidase (BX)	Depth	9.5314	3	2.30E-02	*
	Station	22.298	24	5.62E-01	
	Position	3.6807	1	5.60E-02	
Alkaline phosphatase (AP)	Depth	1.6459	3	6.49E-01	
	Station	34.233	24	8.06E-02	
	Position	0.010611	1	8.10E-01	
N-acetylglucosaminidase (NAG)	Depth	1.1675	3	7.61E-01	
	Station	37.264	24	4.12E-02	*
	Position	0.33573	1	4.60E-01	

3.3. DNA extractions and shotgun metagenomic sequencing

Surface samples (5 m) were collected using the Underway system of the R.V. S.A. Agulhas II between Cape Town, South Africa, and the PEIs. A total of 11 locations were sampled in duplicate (n=22). Microbial communities were collected from three different depths, including Fmax, OMZ and Deep, around Marion Island. The samples were filtered

and nucleic acids were extracted from the filters. DNA samples were sent to Admera Health Biopharma Services (New Jersey, USA) for sequencing on an Illumina 2x150 instrument (Illumina, USA).

3.3.1. Metagenome-assembled genome curation

Shotgun metagenomic sequencing on an Illumina 2x150 instrument produced 6.66 Gb of sequencing data (Supplementary data, Table S1). Primary assembly resulted in 4 279 581 contigs with an average N50 of 16 377.84. Contigs were then assembled into scaffolds. 310 MAGs were produced after binning, however only the 42 high (completeness > 90% and contamination < 5%) and 186 medium quality (50% < completeness ≤ 90% and 5% ≤ contamination < 10%) MAGs were used in further analyses. Metagenomes were produced from 47 out of the 57 samples. Metagenomes were not produced for one CR03 sample, one S04 sample, one S03 sample, MT1-7 (Fmax), MT1-1 (Deep), MT1-1 (Fmax), X04 (Deep), T09 (Deep), T08B (Deep) and T05 (Deep).

Genome size ranged from 0.474 Mb in T59 (Thermoproteota) to 9.294 Mb in S09 (Cyanobacteria). The number of predicted genes ranged from 684 genes in T59 (Thermoproteota) to 13 662 genes in S09 (Cyanobacteria). The GC content of the MAGs ranged from 29.5% in M16 (Bacteroidota) to 70.1% in S07 (Actinobacteria), while the coding density ranged from 86.7% in T27 (Proteobacteria) to 96.5% in M16 (Bacteroidota) (Supplementary data, Table S1).

3.4. Phylogenetic diversity of the MAGs

The phylogenetic diversity, of the 25 sites in the SIO, was evaluated using GTDBK-Tk software. In total, we recovered 58 MAGs affiliated with the kingdom Archaea (Figure 11) and 170 MAGs linked with Bacteria (Figure 12) (Supplementary data, Table S5). A total of 11 phyla were observed in the dataset. Nine of the 11 phyla are bacterial, while the remaining two phyla are archaeal. Proteobacteria constituted a substantial proportion of the MAGs (86 MAGs), followed by Thermoplasmatota (39 MAGs), Verrucomicrobiota (23 MAGs), SAR324 (16 MAGs), Bacteroidota (14 MAGs), Actinobacteria (13 MAGs), Thermoproteota (9 MAGs), Cyanobacteria (8 MAGs), Marinisomatota (5 MAGs), Patescibacteria (3 MAGs) and Planctomycetota (2 MAGs).

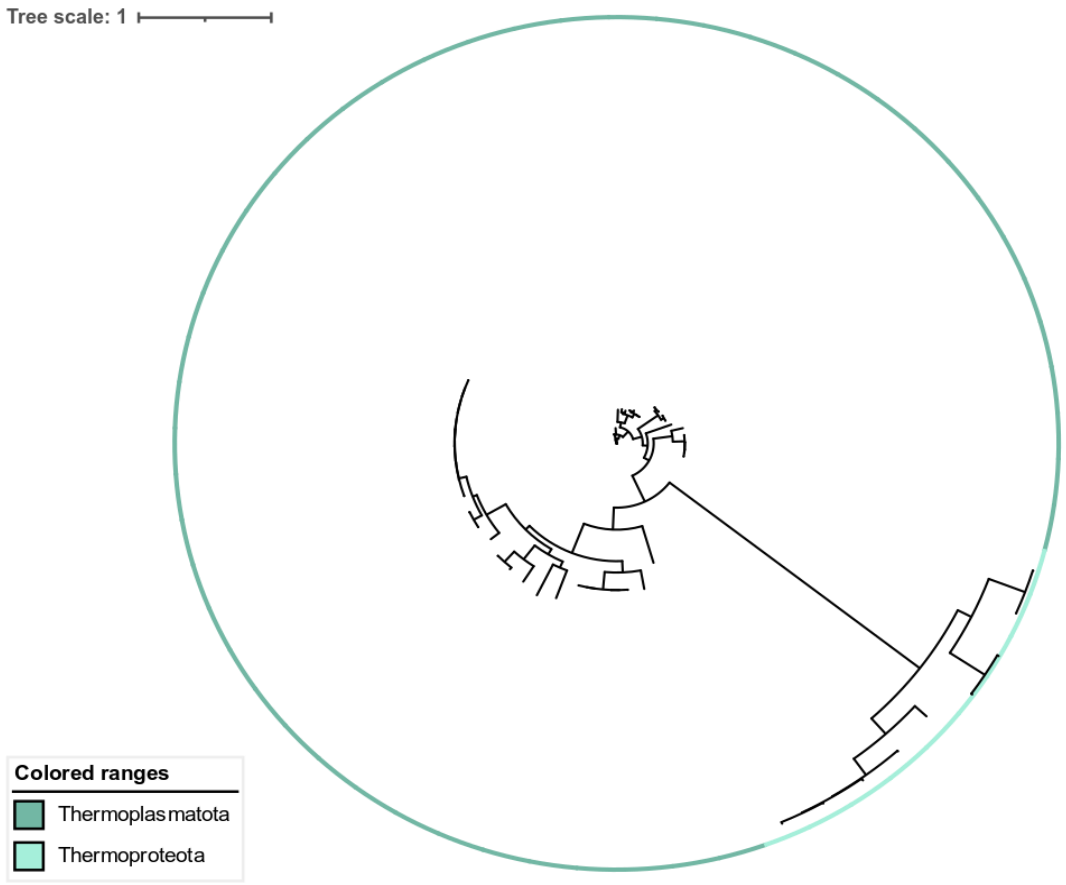

Figure 11: The phylogenetic placement of 58 reconstructed metagenome-assembled Archaeal genomes. A maximum-likelihood phylogenomic tree was built based on concatenated amino acid sequences GToTree. The scale represents one amino acid substitution per sequence position.

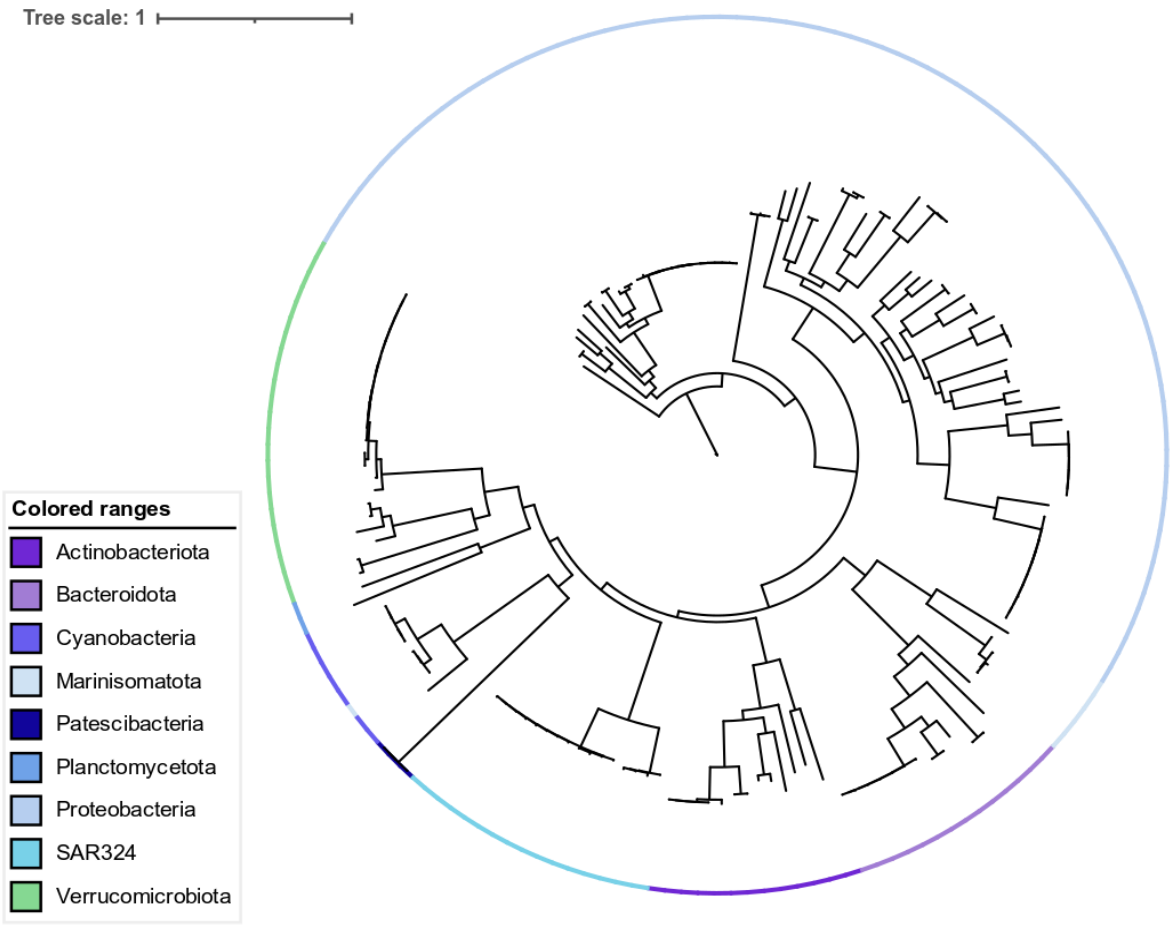

Figure 12: Phylogenetic placement of 170 reconstructed metagenome-assembled bacterial genomes. A maximum-likelihood phylogenomic tree was built based on concatenated amino acid sequences GToTree. The scale represents one amino acid substitution per sequence position.

3.5. Gene duplications

3.5.1. Putative paralogs

We assessed putative paralogs in high and medium quality genomes. In total, the 228 high and medium quality MAGs possessed 88 890 significant putative paralogs from a total of 487 476 predicted genes. Two MAGs (S09 and S22) retrieved from the 5 m S-transect samples were not included in the analysis since they were outliers with 5887 and 3716 putative paralogs, respectively (Figure 13, Figure 14 and Figure 15). S09 is a member of the Cyanobacteria phylum and was the only Cyanobacterium affiliated with *Prochlorococcus* in this study (Supplementary Table S5). S22 was affiliated with Gammaproteobacteria (Supplementary Table S5). S09 had the highest relative proportion of paralogs (43.09%) relative to the number of predicted genes, while C27 had the lowest relative proportion of paralogs (3.69%) (Supplementary Table S1). C27 was assigned to Patescibacteria (Supplementary table S5). A significant difference in the number of putative paralogs was observed for each phylum (Table 7, Figure 14) ($p < 0.001$). Verrucomicrobiota possessed the highest mean number of paralogs, followed by Actinobacteria, while Patescibacteria possessed the lowest number (Figure 14). A significant difference was not observed for each depth (Table 7, Figure 13a), nor each position (Table 7, Figure 13b). MAGs from Fmax have the highest mean putative paralogs, while MAGs from OMZ have the lowest mean putative paralogs (Figure 13a). MAGs from the inter-island stations had a slightly higher mean number of putative paralogs compared with MAGs from the open ocean stations (Figure 13b). The highest number of putative paralogs was observed at station T02, while the lowest was observed at station S04 (Figure 15). The relationship between the total number of paralogs and the

number of MAGs belonging to each phylum was determined by calculating the Spearman's correlation (Figure S1). The Spearman's rank correlation coefficient for the scatterplot is 0.86.

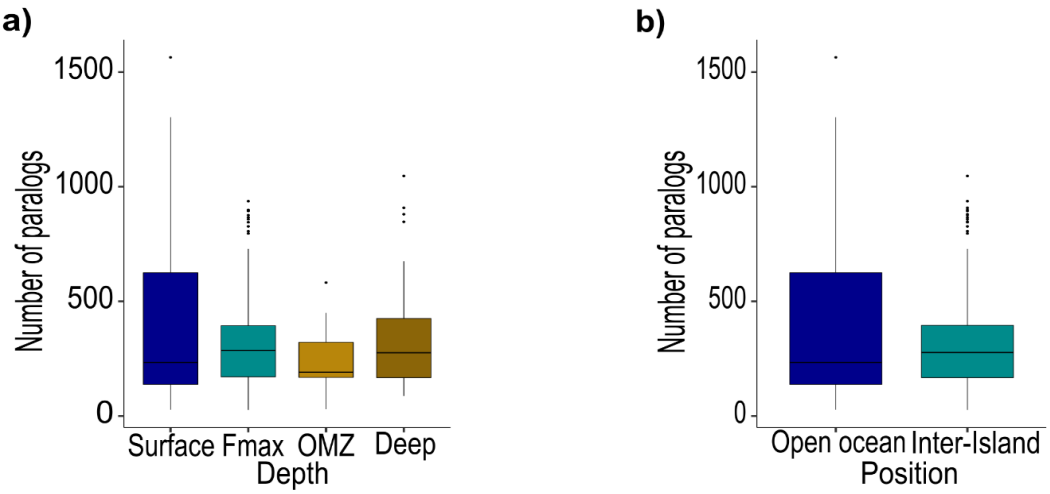

Figure 13: Boxplots of the number of paralogs according to a) depth and b) position (open ocean or inter-island). Two outliers from 5 m in the open ocean were removed. Whiskers on each boxplot indicate the minimum and maximum values and solid black lines indicate the median value. Top and bottom vertical box limits show the 25th and 75th percentile, respectively.

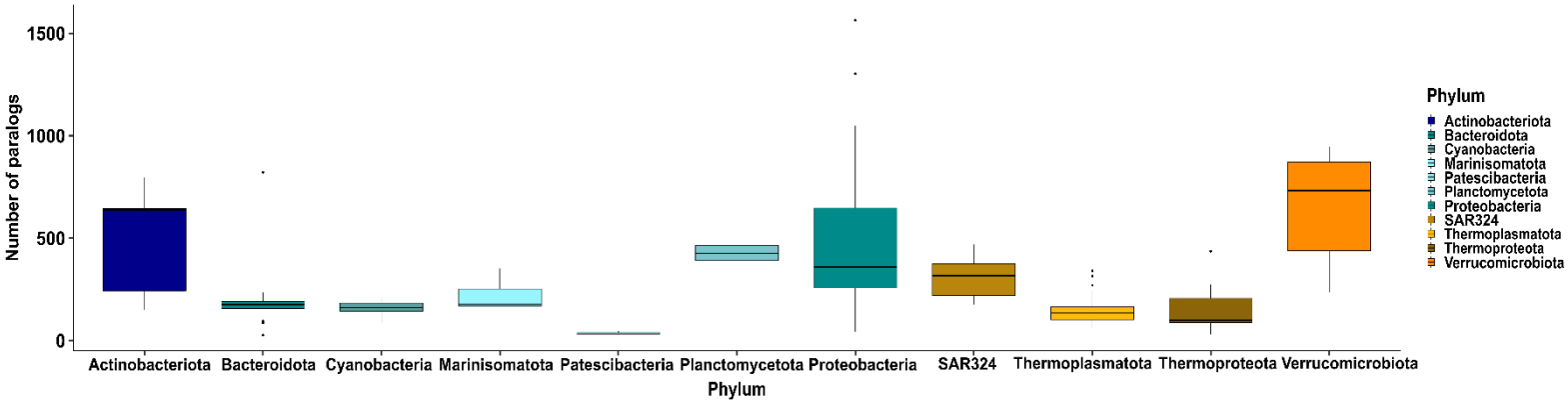

Figure 14: Boxplot showing the number of paralogs for all samples for each phylum. One Cyanobacterium and one Proteobacterium were removed because they were outliers.

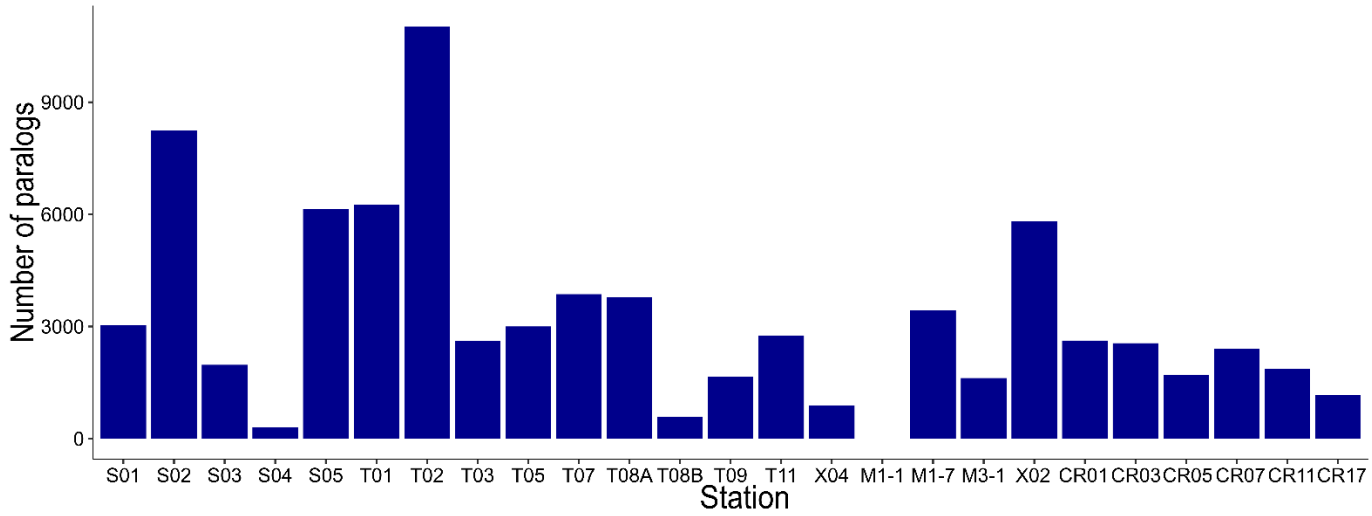

Figure 15: Bar graph showing the number of paralogs for all samples for each depth. ‘S’ and ‘CR’ represent samples taken from the underway system between South Africa and Marion Island, while ‘T’ represents samples from the ‘T-transect’ and ‘X’ and ‘MT’ represent samples from the ‘X-M transect’ taken around the PEIs. One sample from S01 and S02 were removed because they were outliers.

Table 7: Kruskal-Wallis non-parametric test of significance for the number of paralogs for each depth, phylum and station. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.

Parameter	Chi-squared value	Degrees of freedom	P-value	Significance
Depth	2.7918	3	0.4249	
Phylum	126.28	10	2.2E-16	***
Station	23	23	0.4608	

3.5.2. Duplicated gene functions

We determined the number of paralogs found in the MAGs, per station (Supplementary data, Table S1). Specifically, paralogs which encode enzymes associated with the extracellular enzymatic activity assays were determined. The number of paralogs found in the MAGs which encode each enzyme varied amongst the stations, but did not differ significantly (Table 8). In total, 12 paralogs amongst the MAGs were found to encode BG in samples from stations T01 and T02, while no paralogs encoding BG were observed at stations S04, T08B, MT1-1 and CR17 (Figure 16a). Paralogs responsible for BX were mostly absent from the MAGs at each station. However, station CR01 had the highest number of paralogs responsible for the enzyme with a total of three paralogs (Figure 16b). Paralogs encoding NAG were present in MAGS retrieved from 13 stations, with those from T02 possessing seven paralogs which encode NAG (Figure 16c). Paralogs encoding AP were present in MAGs at all stations except T08B, MT1-1 and CR07 (Figure 16d). MAGs from station T02 had the highest number of paralogs which encode AP with a total

of 43 paralogs (Figure 16d). The number of paralogs responsible for AP was highest in the T-transect.

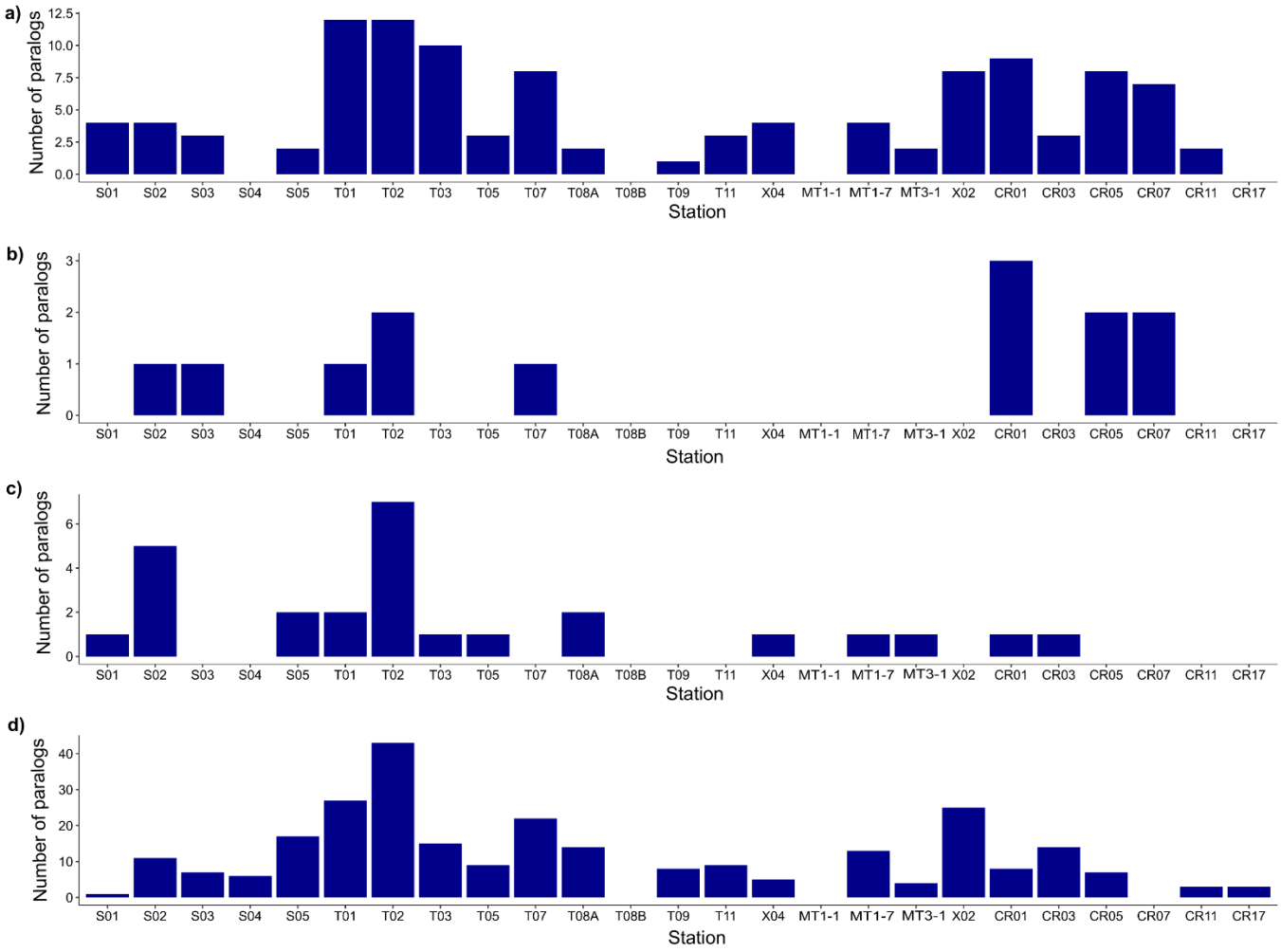

Figure 16: Bar graphs showing the number of paralogs for all samples for each extracellular enzyme sampled for a) β-glucosidase (BG), b) β-xylosidase (BX), c) chitinase (NAG) and d) alkaline phosphatase (AP). ‘S’ and ‘CR’ represent samples taken from the underway system between South Africa and Marion Island, while ‘T’ represents samples from the ‘T-transect’ and ‘X’ and ‘MT’ represent samples from the ‘X-M transect’ taken around the PEIs.

The study by Acinas et al. (2021), as well as KEGGDecoder version 1.3 (https://github.com/bjtully/BioData/blob/master/KEGGDecoder/KOALA_definitions.txt) was used to select important marine biogeochemical cycles. These include carbon, phosphorus and nitrogen metabolism, methanogenesis and methane oxidation (Figure 17). Paralogs were found in carbon, phosphorus and nitrogen metabolism, as well as methanogenesis cycles (Figure 17) (Supplementary data, Table S6).

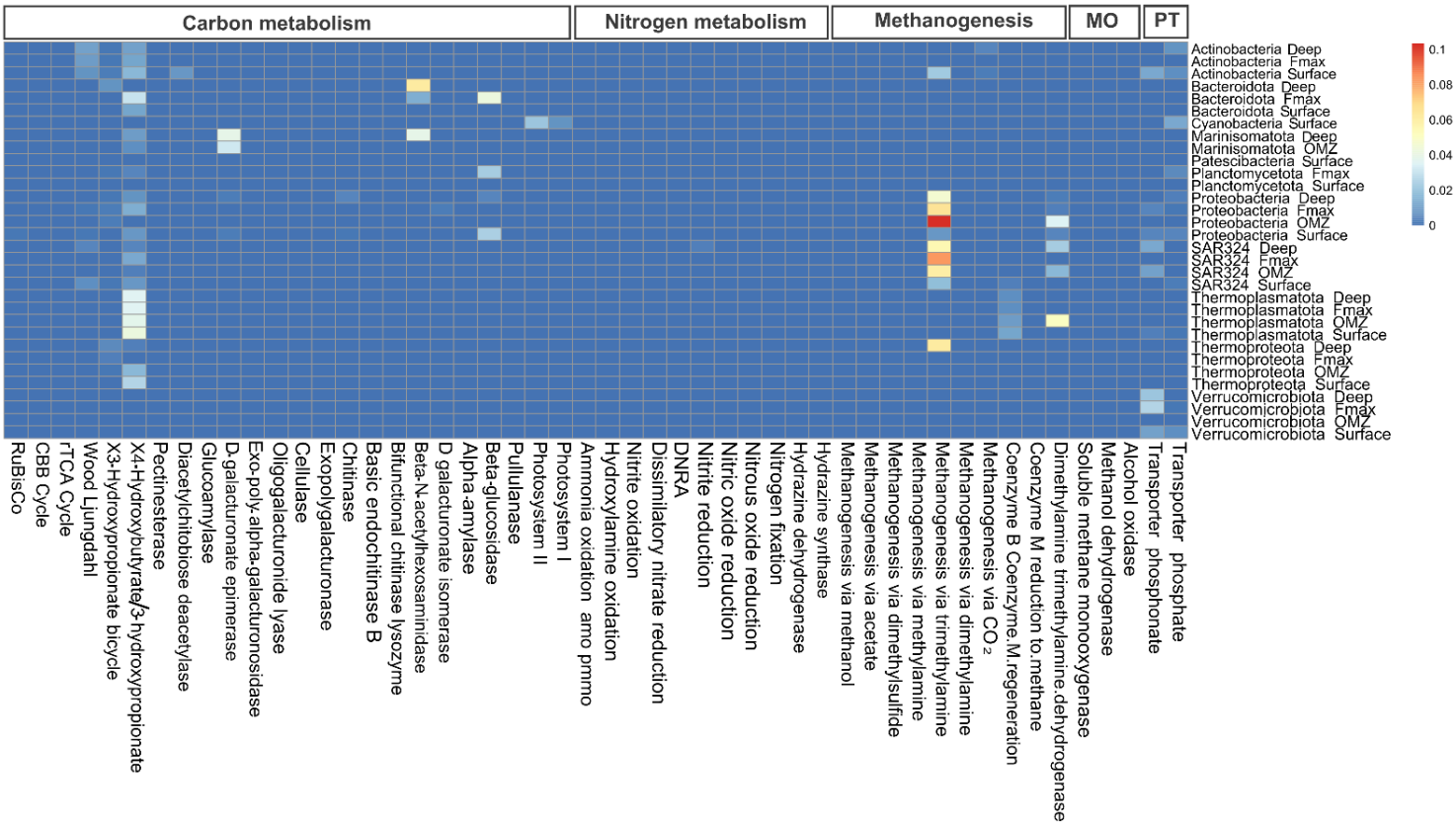

Figure 17: Heatmap showing putative paralog functions of all 228 MAGs for each phylum for each depth. Pathways are separated into carbon, nitrogen and phosphorus metabolism, as well as methanogenesis and methane oxidation. Paralogs responsible for β-glucosidase (BG) in Bacteroidota from deep samples were excluded since they are

outliers. The scale represents the completeness of each pathway from 0.0 being incomplete to 1.0 being complete. ‘MO’ = methane oxidation; ‘PT’ = phosphorus transport.

The relationship between the number of paralogs encoding an extracellular enzyme and the activity of that enzyme at each of the three depths at the stations from the T-transect and X-M-transect was determined by calculating the Spearman’s correlation (Figure S2). The Spearman's rank correlation coefficients for BG, BX, NAG and AP are 0.06, -0.03, -0.05 and 0.12, respectively.

Table 8: Kruskal-Wallis non-parametric test of significance for the number of paralogs for each extracellular enzyme assayed for each station. Significant differences are indicated with a *, with *p-value < 0.05 and ***p-value < 0.001.

Extracellular enzyme	Chi-squared value	Degrees of freedom	P-value	Significance
β-glucosidase (BG)	24	24	0.4616	
β-xylosidase (BX)	24	24	0.4616	
N-acetylglucosaminidase (NAG)	24	24	0.4616	
Alkaline phosphatase (AP)	24	24	0.4616	

3.6. The location of putative biogeochemical cycling genes within the SIO pangenome

The accessory genomes of the 228 MAGs constitutes the largest proportion of their pangenome. None of the predicted genes, including putative paralogs, were found in the core and soft-core genome of the 228 MAGs (Table 9, Figure 18). The predicted genes

and paralogs were found in the accessory genome of the MAGs, with only 132 genes being shared amongst 15% to 95% of the MAGs (Table 9). The rest of the 190 002 predicted genes were strain-specific or shared amongst less than 15% of the MAGs (Table 9).

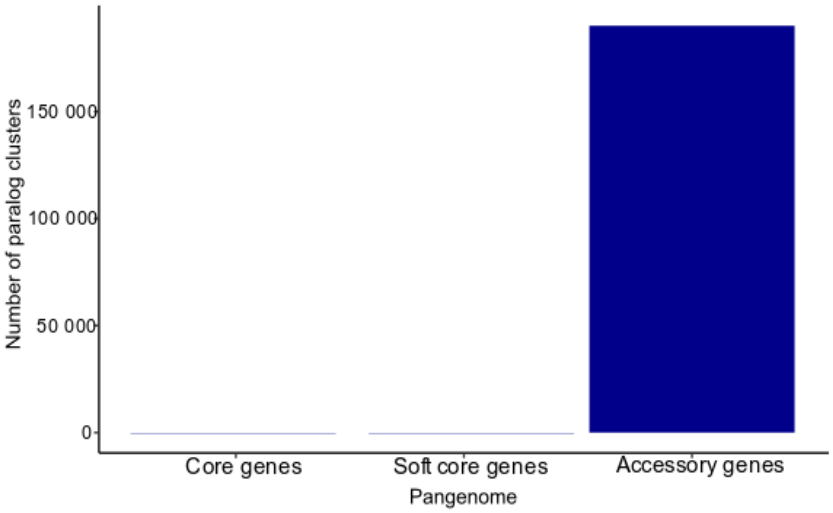

Figure 18: Bar graph of the pangenome of the 228 metagenome-assembled genomes from the Subantarctic Indian Ocean. Core genes = genes found in 99% < = strains < = 100; soft core genes = genes found in 95% < = strains < 99%; accessory genes = genes found in 95% > strains.

Table 9: The pangenome of the 228 metagenome-assembled genomes from the Subantarctic Indian Ocean.

Pangenome	Definition	Number of genes
Core genes	99% < = strains < = 100%	0
Soft core genes	95% < = strains < 99%	0
Shell genes	15% < = strains < 95%	132

Cloud genes	0% < = strains < 15%	190002
Total genes	0% < = strains < = 100%	190134

A kingdom-level pangenomic analysis also revealed no core genes in bacteria and archaeal MAG lineages. Due to the lack of a core genome, all 16 SAR324 members were used in a phylum-level pangenomic analysis. SAR324 was chosen because the taxa possesses paralogs which partially complete pathways in the biogeochemical cycling modules focused on in this study, except methane oxidation. The SAR324 core genome was comprised of 21 genes, which includes seven ribosomal proteins. In addition, we found four proteins involved in transcription, four proteins involved in carbon metabolism (Holness and Sugden 2003; Kim et al. 2004; Wiegand and Remington 1986; Yao et al. 2011), and one protein with the potential for nitrogen and carbon metabolism (Melo et al. 2004; Peng et al. 2011). Genes for NAG, BG and BX were not found in the SAR324 pangenome, while the pangenome possessed multiple copies of genes encoding AP. The pangenome included four genes that encode multifunctional alkaline phosphatase superfamily protein. Two of the genes were shared amongst 12 members, two shared amongst two members and the last gene is specific to C29.

4. Discussion

Microbial communities play profound roles in ecosystems including the world's oceans. However, we lack fundamental insights regarding their diversity and function. Specifically, little is known regarding the extent of gene duplications and their effects on ecosystem services. This study provides the first insights regarding the extent of gene duplications

in microbial communities in the SIO. The data provides insights into the factors which potentially affect the rate of gene duplication events in this region, as well as the potential function of the paralogs observed in these communities. Previous studies on microbial communities in the PEIs and surrounding waters have found that microbes play a role in biogeochemical cycling in the region, specifically the cycling of carbon and nitrogen (Kaehler et al. 2000; Phoma and Makhalanyane 2021). The combination of extracellular enzymatic activities and in silico analysis suggests that paralogs may play a role in these cycles, as well as in the cycling of phosphorous.

4.1. Chemical characteristics of PEI marine microbiomes

Nutrient concentrations were in line with previous studies (Henley et al. 2020), where nutrient concentrations were lowest from open ocean samples taken towards South Africa (Figure 5) (Henley et al. 2020; Jasmine et al. 2009; Sigman et al. 2000). Exceptions to this trend were nitrite (Figure 5b) and ammonium (Figure 5d), which were higher in samples from the open-ocean compared with those from inter-island stations. Higher nutrient concentrations are expected in the inter-island samples due to the effects of land-nutrient supply, where the inter-island stations receive nutrients from the PEIs in the form of run-off including macrofauna colony faecal matter and rivers flowing into the ocean (Boden 1988; Treasure et al. 2015). Seals and penguins may be responsible for local changes in nutrient concentrations in PEI marine communities, such as iron, through faecal matter and run-off from seabird guano (Treasure et al. 2015; Wing et al. 2014). Rivers are also known to be a source of dissolved organic nitrogen (DON) (Seitzinger et al. 2005), which is converted to ammonium through ammonification or remineralisation

(Rush and Sinninghe Damsté 2017). This suggests that higher ammonium concentrations in surface samples (Fmax) are expected, which was observed in this study (Figure 5).

The concentration of nutrients measured showed a depth-dependant trend (Figure 5, Table 5). The concentration of nitrate, nitrate + nitrite, phosphorus and silicate are highest from OMZ samples, while the concentration of nitrite and ammonium is highest from Fmax samples. These depth-dependant trends are in line with a study conducted on nutrients in the SIO by Henley et al. (2020). The trends in the concentration of phosphate and nitrate could be the same since nitrogen-fixing microbial species rely on phosphate, potentially limiting and regulating nitrogen fixation and nitrogen cycling (Berman-Frank et al. 2007). The depth-dependant trends in nutrient concentrations in the Antarctic zone (AZ) are seasonal and driven by upwelling (Kemeny et al. 2016). The ice melts during summer and nutrients that were at the surface due to upwelling in winter are free to become stratified along the water column (Kemeny et al. 2016). Phytoplankton are not light-limited in summer like they are in winter, allowing them to draw nutrients, such as nitrate, down the water column to create higher nutrient concentrations in deeper waters (Altabet and Francois 2001; Kemeny et al. 2016). The concentration of ammonium and nitrite, however, do not follow this same trend.

The concentration of ammonium and nitrite are lowest from OMZ locations and highest from Fmax locations (Figure 5). The depth-dependant trend of nitrogen-containing compounds could be due to the presence or absence of oxygen, since oxygen is required by certain pathways and inhibited by others in the nitrogen metabolism cycle (Voss et al. 2013; Zhang et al. 2020). Denitrification and annamox both occur at depths that are almost absent in oxygen, such as the OMZ (Hawley et al. 2014). Marine microbes convert

dissolved inorganic nitrogen, such as nitrate, nitrite and ammonium, into biologically available dinitrogen gas (Kuypers et al. 2018; Voss et al. 2013; Zehr and Kudela 2011; Zhang et al. 2020). Anammox bacteria carry out anaerobic carbon dioxide fixation while oxidizing ammonium and nitrite to dinitrogen. They also oxidize nitrite to nitrate (Voss et al. 2013). This would explain the low levels of nitrite and ammonium observed in OMZ samples, while the highest concentration of nitrate is observed in OMZ samples (Figure 5). These depth-dependant nutrient trends have been shown to influence marine microbial community structure in the SIO (Phoma et al. 2018).

4.2. Extracellular enzymatic activities of PEI marine microbiomes

Potential extracellular enzyme activity was detected in all samples, however some of the samples were removed as outliers in the AP and NAG assays. The detection of enzyme activities in all samples indicates the presence of functionally heterotrophic microbial communities in the PEIs (Phoma and Makhalanyane 2021), although further analyses are needed to confirm the functional capacity of these microbial communities. Enzyme activity values and depth trends are similar to those found in literature (Misic et al. 2002; Thomson et al. 2019). Extracellular enzymatic activities of marine microbial communities are known to differ significantly by depth and distance from the PEIs (Phoma and Makhalanyane 2021). Highest BG and BX activity could be observed in 5m and Fmax samples, respectively, since it is part of the surface marine environment that is known to be highly productive since this environment harbours the most primary producers (Stock et al. 2019). The lowest AP activities have previously been observed in surface waters as a result of solar radiation (Thomson et al. 2019), which has a decreasing effect on the life-time of cell-free enzymes in marine environments, therefore decreasing enzyme activity

in surface waters (Thomson et al. 2019). This same principle could be applied to the rest of the enzymes assayed in this study, explaining why NAG activities were lowest in surface samples. Increased enzyme activity in surface waters could also be due to the size of dissolved organic matter (DOM) particles, where larger particles act as hotspots of microbial enzymatic activity (Arnosti et al. 2014). DOM is typically found in higher concentrations in the surface waters surrounding the PEIs, due to freshwater run-off from the islands in the IME (Boden 1988; Perissinotto and Rae 1990).

Extracellular enzyme activity assays can be a limited approach in determining the potential functional capacity of microbiomes. This is due to limited access to much of the ocean because of difficult sampling techniques which require extended periods of time on research vessels (Arnosti et al. 2014). Hence, majority of previous studies on extracellular enzymatic activity in marine microbiomes focused on surface, near-surface waters (< 200 m) and coastal waters (Arnosti et al. 2014). A limitation of combining extracellular enzyme assays and MAGs is that only complete MAGs were focused on and some samples did not produce MAGs, therefore we cannot be sure the observed MAGs are responsible for the enzyme activities. Activities could be a result of microbes that were not sequenced in each sample. Despite all of this, the extracellular enzymatic activities in this study are indicative of the cycling of nutrients such as carbon (BX and BG), nitrogen (NAG) and phosphorous (AP).

4.2.1. Carbon acquisition (BX and BG)

Heterotrophic carbon cycling is initiated by the release of extracellular enzymes such as BG and BX (Arnosti et al. 2014). BG and BX hydrolyse monosaccharide beta bonds that

can be used as energy sources from glycol- and xylo-oligosaccharides, respectively, to support microbial life in the ocean (Galbe and Zacchi 2002; Luo et al. 2017). These enzymes hydrolyse carbon-containing substances into smaller compounds that can be taken up by microbial cells and used for respiration and primary production as well (Arnosti et al. 2014). The activity of BG has been linked to specific bacterial lineages (Knight and Dick 2004), such as Bacteroidetes, Proteobacteria and Actinobacteria groups (Tiwari et al. 2016). Each of these bacterial groups have been found across the depths sampled in the PEI marine microbiomes in this study (Supplementary table S5), hence the data suggests that BG is expressed and used to acquire carbon in these microbial communities. The detection of both BG and BX activity in our samples suggests that carbon acquisition occurs in PEIs marine microbial communities using both enzymes.

4.2.2. Phosphorus acquisition (AP)

Alkaline phosphatase is an enzyme possessed by marine phytoplankton and bacteria that allows them to access inorganic phosphate (Pi) for growth (Duhamel et al. 2010; Hoppe 2003). The enzyme hydrolyses phosphomonoesters and phosphodiesters, releasing biologically available Pi for microorganisms to use (Luo et al. 2017). Pi is one of the main limiting macronutrients of primary production in open oceans (Björkman and Karl 1994; Tyrrell 1999). It is the preferred form of phosphorous for marine microbes since microbes possess high affinity uptake pathways, despite Pi being limited in the open ocean (Duhamel et al. 2010; Karl 2000; Thomson et al. 2019). Our study found the highest rates of AP activity (Figure 8) at the depth with the highest concentration of phosphate (OMZ) (Figure 5). This could also be due to the alkaline phosphatase paradox (Chrost 1991; Thomson et al. 2019), where high activity is observed in regions with high concentrations

of Pi (Thomson et al. 2019) due to the long 16 day lifetime of the enzyme (Thomson et al. 2019) or higher concentrations of trace elements needed for AP functioning rather than direct phosphate concentrations.

4.2.3. Nitrogen acquisition (NAG)

N-acetylglucosaminidase (NAG) is essential for nitrogen cycling since it catalyses the terminal degradation of chitin and N-acetylglucosamine substrates (Shi et al. 2012), releasing biologically useful nitrogen and carbon (Herdyastuti et al. 2009; Ikeda et al. 2007). The enzyme allows marine microbes living in nitrogen-limited waters to carry out important functions such as primary production (Björkman and Karl 1994; Thomson et al. 2019; Tyrrell 1999). Thus, NAG has been used as an indicator of nitrogen acquisition (Luo et al. 2017; Phoma and Makhalanyane 2021). Many bacteria are able to use chitin exclusively as a nitrogen and carbon source (Santhi 2016; Zobell and Rittenberg 1938), including Pseudoalteromonas (Wang et al. 2014) and Marinisomatota (Coutinho et al. 2021), which are both present in our dataset (Supplementary data, Table S5).

The activity of NAG shows the same trend as the concentration of nitrate, while showing the opposite trend to the concentration of ammonium and nitrite. Nitrate concentrations in our data are highest from OMZ samples (Figure 5), which could explain the activity of NAG being highest in OMZ samples. The trends in NAG activity do not necessarily align with the trends in the concentration of other nitrogen-containing compounds, such as nitrite and ammonium. Ammonium and nitrite concentrations are highest in Fmax samples and lowest in OMZ samples, showing the opposite trend to that of NAG activity. This data

suggests NAG favours nitrate as a source of nitrogen since highest NAG activity is found at the same depth as the highest nitrate concentration.

4.3. Physical factors of PEI marine microbiomes

This study generated 228 high and medium quality MAGs from depths varying between 5 m to 4077 m below the sea surface (Table S1). One limitation of this study is that physical factors were only recorded for the inter-island stations, at three depths (Fmax, OMZ and Deep). Another limitation of this study is that more samples were taken from the Deep layer (n=14) and the Fmax layer (n=14) than the OMZ layer (n=7). This could distort the spread of the Deep samples in the boxplots for each physical factor. To overcome this, inter-island stations that did not possess an OMZ were removed to make a total of seven samples for Fmax, OMZ and Deep depths.

Environmental factors have been shown to play an important role in shaping microbial communities, up to 35%, and thus their potential functions in marine microbiomes of the PEIs (Phoma and Makhalanyane 2021). Depth acts as an important factor in shaping microbial communities in the SIO (Liu et al. 2018; Lønborg et al. 2016; Phoma and Makhalanyane 2021). This can be extended to suggest that depth-dependant variables, such as salinity, temperature, sunlight and oxygen (Phoma and Makhalanyane 2021), also influence the structure of microbial communities and ecosystem services in this region (Diehl et al. 2002; Easson and Lopez 2019).

Physical factors of the PEIs, including pressure, temperature, salinity, conductivity and oxygen levels, show a significant depth-dependant trend (Figure 3, Table 4). Pressure and salinity increase with depth, while temperature and conductivity decrease with an

increase in depth. Depth acts as a proxy for sunlight penetration (Phoma and Makhalanyane 2021), hence the deeper the sample the less sunlight the water receives and hence the lower the temperature will be (Diehl et al. 2002; Orcutt et al. 2011). Many of the deep samples are found within the AABW mass (Supplementary table S2), which is known to possess low temperatures that range from -0.9 to 1.73°C since the water mass originates in the Antarctic region (Emery 2001). Temperature regulates the kinetic rates of metabolic processes, since enzymes function at an optimal temperature and any temperature above or below the optimal temperature will cause a decrease in enzymatic activity (Arnosti et al. 2014; Price and Sowers 2004). This could explain lower enzyme activity in deeper samples.

Five water masses were identified based on temperature and salinity readings (Figure 3) (Emery 2001), including subantarctic Surface Water (SASW), Antarctic Surface Waters (AASW), Antarctic Intermediate Water (AAIW), Circumpolar Deep Water (CDW) and Antarctic Bottom Water (AABW). Antarctic surface waters are characterised by colder surface temperatures, while subantarctic surface waters are characterised by warmer temperatures and lower salinity measurements (Ansorge and Lutjeharms 2002; Emery 2001). Subantarctic water masses are colder than subtropical water masses, due to the temperatures of the waters the masses originate from (Venkatachalam et al. 2019). Phoma and Makhalanyane (2021) conducted a study in the same region, where the authors found that the significant differences between subantarctic and Antarctic water masses shape the hydrographic conditions of sample stations. Since oceanographic features are shaped by the water masses they originate from, one can conclude that

water masses significantly influence the community structure of marine microbial communities (Phoma and Makhalanyane 2021; Venkatachalam et al. 2019).

4.4. Taxonomy of the PEI marine microbiomes

The PEIs marine microbiomes were dominated by bacteria, with 170 of the 228 MAGs belong to bacterial lineages (Figure 12), and the other 58 MAGs belong to archaeal lineages (Figure 11). This is in line with previous studies that found bacteria dominated each depth in PEIs marine microbiomes in the SIO, followed by archaeal lineages (Phoma et al. 2018). A total of 11 phyla were observed in the dataset (Figure 11, Figure 12). Proteobacteria dominated these microbial communities, making up 86 out of the 228 MAGs in this study. Proteobacteria are known to dominate these waters, followed by Actinobacteria, Bacteroidetes, Firmicutes and Cyanobacteria (Phoma et al. 2018).

Verrucomicrobiota (Mohammadi et al. 2017), SAR324 (Swan et al. 2011), Bacteroidota (DeLong et al. 1993), Cyanobacteria (Garcia et al. 2015; Herrero et al. 2001), Proteobacteria (Hanson and Hanson 1996; Swan et al. 2011) and Patescibacteria (Bayer et al. 2020) are known to contribute to the marine carbon cycle. Proteobacteria include methanotrophs, organisms that use methane as a sole carbon and energy source and contribute to the carbon cycle (Hanson and Hanson 1996). Gammaproteobacteria that are known to play a role in carbon cycling using ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) genes (Swan et al. 2013). Verrucomicrobiota possess genes responsible for methane metabolism, as well as the CBB pathway for CO_2 fixation (Kim et al. 2022). SAR324 members have been implicated carbon fixation pathways in the deep ocean using genes responsible for RuBisCO (Swan et al. 2011), tricarboxylic acid (TCA) cycle, glycolysis/gluconeogenesis and the Calvin–Benson–Bassham (CBB)

pathways (Sheik et al. 2014). Bacteroidota have evolved partly due to gene duplications (Thomas et al. 2011) and are thought to be adapted to use both dissolved and particulate high molecular weight compounds (DeLong et al. 1993) in the marine carbon cycle (Bauer et al. 2006). Patescibacteria are anaerobes that have reduced genomes that possess glycolysis genes, implicating them in carbon cycling (Bayer et al. 2020). Cyanobacteria fix carbon dioxide through the Calvin cycle (Stanier and Cohen-Bazire 1977) coupled to dinitrogen fixation (Herrero et al. 2001), linking the carbon and nitrogen biogeochemical cycles (Garcia et al. 2015). The presence of phyla known to play a role in carbon cycling in our data, suggests that carbon cycling is being carried out in the PEIs. Phyla found in our data have also previously been implicated in nitrogen cycling in marine microbiomes.

Planctomycetota (Jetten et al. 2003; Strous et al. 1999), Marinisomatota (Bertagnolli et al. 2017; Pérez Castro et al. 2021), Verrucomicrobiota (Kim et al. 2022) and SAR324 (Sheik et al. 2014) are known to play a role in the biogeochemical cycling of nitrogen. Planctomycetes mediate the anammox (anaerobic oxidation of ammonia) process, playing a role in nitrogen cycling in marine microbiomes (Jetten et al. 2003; Strous et al. 1999). Marinisomatota are predominant in OMZ layers, where the phylum participates in marine nitrogen cycles through genes responsible for partial denitrification (Bertagnolli et al. 2017; Pérez Castro et al. 2021). Verrucomicrobiota are known as key nitrogen cycling bacteria, since the group possesses genes responsible for nitrogen fixation and assimilation pathways (Kim et al. 2022). SAR324 possess genes that encode a nitrite reductase, implicating SAR324 in nitrogen cycling (Sheik et al. 2014). The presence of phyla known to play a role in the biogeochemical cycling of nitrogen in our data, suggests that carbon cycling is being carried out in the PEIs.

Archaeal lineages, such as Thermoplasmatota and Thermoproteota, were also observed in the PEIs microbial communities (Figure 11). Both these phyla are known to be ubiquitous in these microbiomes (Phoma et al. 2018). Thermoproteota are chemolithoautotrophic ammonia oxidizers that play a key role in biogeochemical cycling in the ocean (Pester et al. 2011), specifically in nitrogen cycling. All known ammonia-oxidising archaea (AOA) belong to this phylum, which carry out the first step of nitrification in the nitrogen cycle (Pester et al. 2011). Members of Thermoproteota are known to live in deep waters (Aylward and Santoro 2020; Orcutt et al. 2011) and drive primary production and nitrogen cycling in these waters (Herndl et al. 2005; Karner et al. 2001). 45 out of the 49 Thermoplasmatota belong to Marine Group II (MGII) (Supplementary table S5). MGII is the dominant archaeal group of open ocean surface waters, however the class is ubiquitous amongst all depths (Adam et al. 2017). Little is known about the functional capacity of MGII since we lack cultured representatives and genome references (Tully et al. 2018). However, it is thought that MGII play a role in the carbon cycle, since the group has displayed abundant transcripts for nitrite oxidoreductase complex subunits and nitrite oxidation electron transfer components, as well as amino-acid transporters involved in the recycling of organic carbon and nitrogen from amino acids (Baker et al. 2013; Zhang et al. 2015). The presence of these AOA Thermoproteota and Thermoplasmatota capable of nitrogen cycling in our data suggests that these processes are being carried out in the marine microbial communities of the SIO.

4.5. Factors affecting gene duplications in the SIO

Metagenomes were not produced for one CR03 sample, one S04 sample, one S03 sample, MT1-7 (Fmax), MT1-1 (Deep), MT1-1 (Fmax), X04 (Deep), T09 (Deep), T08B

(Deep) and T05 (Deep). There is thus no metagenomic and gene duplication data for four deep samples (T05, T08B, T09 and MT1-1) and two Fmax samples (MT1-1 and MT1-7). This could affect the trends in the number of paralogs per depth, position (open ocean and inter-island), phylum and station. Despite this, the 228 MAGs possessed a total of 88 890 significant putative paralogs out of a total of 487 476 predicted genes. The two MAGs with the highest number of paralogs are members of Cyanobacteria and Alphaproteobacteria (Gammaproteobacteria), while the MAG with the lowest relative proportion of paralogs was assigned to Patescibacteria (Supplementary table S5). Interestingly, both MAGs with the highest number of paralogs were sampled from open ocean surface waters. Despite these two outliers being removed from further analyses, the open ocean surface waters still possess the highest number of paralogs (Figure 13).

Despite a lack of a significant difference in the number of paralogs per depth, station and position, there is a significant difference in the number of paralogs per phylum (Table 7, Figure 14) ($p < 0.001$). The Spearman's rank correlation coefficient for the scatterplot of the number of MAGs and total number of paralogs found within a phylum is 0.86 (Figure S1), showing a strong correlation between the two variables. Proteobacteria, Actinobacteria and Verrucomicrobiota show high levels of gene duplications in this study (Figure 14). These phyla are known to share environmental conditions in soil communities (Janssen et al. 2002) and have been clustered together phylogenetically as phyla that share the same lifestyle (Álvarez-Lugo and Becerra 2021). Thus, it would be expected that these phyla possess gene duplications in similar proportions and for similar roles since they share environmental conditions and lifestyles. Planctomycetes were also found to have a high number of paralogous genes in this study (Figure 14). Planctomycetes,

Verrucomicrobiota, Chlamydiae make up the PVC super-phylum (Kamneva et al. 2012; Kim et al. 2016; Wagner and Horn 2006). Studies on the evolution of the PVC have found that gene duplication events play a role in the diversity of these organisms (Kamneva et al. 2012; Kim et al. 2016). Gene duplications were specifically responsible for the divergence of Planctomycetota genomes (Kamneva et al. 2012). Patescibacteria possesses the lowest mean number of paralogs in this study (Figure 14), which is expected since the phylum is known to possess reduced genomes (Bayer et al. 2020; Tian et al. 2020). Reduced genomes could be a result of the loss of gene duplicates through the accumulation of deleterious mutations if duplicates do not confer a benefit to the organism possessing them (Mira et al. 2001).

4.6. The functional capacity of gene duplicates in the SIO

Paralogs with the genomic potential for carbon, nitrogen and phosphorous metabolism, as well as methanogenesis were observed in our data (Figure 17). Although, the use of metagenomic data to quantify gene duplication events is limited since the genomes are incomplete and duplicates contained within a single genome may not be identified (Anderson et al. 2017; Chen et al. 2020). Possibly due to this limitation, the Spearman's correlation coefficient for the number of paralogs assigned to each extracellular enzyme assayed versus the activity of the enzyme was between -0.05 and 0.12 (Figure S2), indicating a very weak correlation.

Paralogs were absent from all three methane oxidation pathways, suggesting they are not responsible for methane oxidation in the SIO. However, paralogs with the genomic potential to carry out nitrogen metabolism were observed in the nitrite reduction pathway (Figure 17). This suggests that paralogs play a role in nitrogen metabolism in the SIO,

albeit at a limited capacity. Phosphorous and carbon metabolism, as well as methanogenesis pathways possessed paralogs with varying levels of completeness (Figure 17). These results suggest that gene duplicates aid in the biogeochemical cycling of phosphorous and carbon in the SIO.

4.6.1. Nitrogen cycling

Paralogs were found in only one nitrogen metabolism pathway, in one phylum at one depth. Paralogs responsible for the nitrite reduction pathway were present in SAR324 from deep samples (Figure 17). This suggests gene duplication events play a role in nitrogen metabolism in the deep ocean in SAR324, however gene duplicates do not play a key role in overall nitrogen cycling in the SIO. This is compounded by the lack of paralogs responsible for NAG activity, with less than half the stations possessing paralogs responsible for NAG (Figure 16). Only Proteobacteria from Deep samples possessed paralogs with the genomic potential for NAG (Figure 17). NAG activity is opposite to the trend in the number of paralogs per depth (Figure 13). The NAG activities were amongst those previously recorded for the area (Supplementary table S4) (Phoma and Makhalanyane 2021), suggesting the enzyme is present however there are very few paralogs of this enzyme.

The lack of paralogs encoding nitrogen cycling genes is compounded by previous studies having linked nitrogen cycling to all 11 phyla found in this study (Bertagnolli et al. 2017; Boeuf et al. 2021; Delmont et al. 2018; Herrero et al. 2001; Phoma et al. 2018; Rinke et al. 2019; Strous et al. 2006; Tian et al. 2020; Zeng et al. 2021b). The lack of paralogs responsible for nitrogen metabolism represented in the heatmap, along with the presence

of NAG activity (Figure 8) and nitrogen cycling phyla suggests that nitrogen cycling in the SIO is not carried out by duplicated genes, but instead by single copies. Although there appears to be a lack of paralogs responsible for direct nitrogen cycling, paralogs have the potential to provide nitrogen sources through methanogenesis in these microbial communities.

Methanogenesis contributes to the marine nitrogen cycle by providing sources of nitrogen (King 1984; Orcutt et al. 2011; Zeng et al. 2021a). Paralogs responsible for methanogenesis are only found in methanogenesis via trimethylamine and CO_2, coenzyme B/M regeneration and dimethylamine/trimethylamine dehydrogenase pathways (Figure 17). The paralogs are mostly found in the methanogenesis via trimethylamine pathway. This pathway is most complete by Proteobacteria and SAR324 at all depths, however the pathway is least complete in 5 m samples. This suggests that paralogs play a role in carrying out methanogenesis in Proteobacteria and SAR324 in deeper communities, contributing to the cycling of both nitrogen and carbon in deeper waters. Trimethylamine is a methylated compound that acts as a source of energy and carbon, acting as an important component of the nitrogen and carbon cycles (Chistoserdova 2011; King 1984; Naqvi et al. 2005; Zhuang et al. 2018).

4.6.2. Carbon cycling

β-xylosidase is typically produced by marine fungi (Raghukumar et al. 2004). Fungi were not identified in the metagenomes produced from the PEIs in this study, however BX activity is still present. This could be due to the high cut-off values for contamination (< 10%) and completeness (> 50%) of MAGs included in this study, thus BX activity could

be a result of unsequenced fungi. Few paralogs were assigned to BX (Figure 16), while BX was observed in all samples from the SIO (Figure 10). This could suggest that paralogs do not play an important role in BX activity in microbial communities in waters surrounding the PEIs. The lack of paralogs responsible for BX, while paralogs responsible for BG are observed at each station might suggest that gene duplicates contribute to the acquisition of carbon primarily through BG.

A significant difference was observed in the number of paralogs responsible for BX and BG activity per station (Table 8). The number of paralogs that were assigned to BG are higher in samples from the T- and CR-transects (Figure 16). This is the same trend in BG activity, suggesting that paralogs play a key role in BG activity in microbial communities in waters surrounding the PEIs. Paralogs responsible for the BG pathway were observed in Bacteroidota and Planctomycetota from Fmax samples, as well as Proteobacteria from Deep and 5 m samples (Figure 17). This suggests paralogs play a role in BG activity in specific phyla from surface and deep waters, but not in OMZ samples. Gene duplication events have been observed in BG before, however it was observed in other systems such as maize (Gómez-Anduro et al. 2011).

Paralogs show great genomic potential for the biogeochemical cycling of carbon, through carbon metabolism and methanogenesis (Figure 17). Paralogs responsible for carbon cycling are found in Wood-Ljungdahl, 3-hydroxypropionate bicycle, 4-hydroxybutyrate/ 3-hydroxypropionate, diacetylchitobiose deacetylase, D-galacturonate epimerase, chitinase, β-N-acetylhexosaminidase, D-galacturonate isomerase, α-amylase and β-glucosidase pathways (Figure 17). 4-hydroxybutyrate/ 3-hydroxypropionate (3HP/4HB) is the most complete pathway amongst the MAGs. This pathway fixes CO_2 in extremely

thermoacidophilic archaea (Berg et al. 2007; Loder et al. 2016). Paralogs responsible for the 3HP/4HB pathway are found in nine of the phyla in this dataset, and is most complete in Thermoplasmatota in all four depths. This result is in line with studies that reveal that 3HP/4HB is an important carbon cycling pathway in archaea (Berg et al. 2007; Loder et al. 2016). This suggests that paralogs play an important role in carbon cycling in this group in the SIO. Paralogs contribute to carbon metabolism through other pathways such as Photosystem I and II, the Calvin Cycle and the Reductive Tricarboxylic Acid Cycle (rTCA).

Carbon metabolism pathways, such as Photosystem I and II, the Calvin Cycle and the rTCA, have previously been implicated in carbon cycling in microbial communities from the SIO (Phoma et al. 2018). Previous studies used RuBisCO genes to identify aerobic carbon fixation via the Calvin Cycle in microbial communities from all four depths in the SIO (Phoma et al. 2018). Paralogs were responsible for the partial completeness of this pathway in our data (Figure 17). Proteobacteria possess paralogs with the genomic potential to carry out the Calvin cycle in this study (Supplementary table S6), however paralogs do not significantly complete the pathway. Proteobacteria and Cyanobacteria are known to possess this pathway for carbon cycling (Herrero et al. 2001; Stanier and Cohen-Bazire 1977), with Cyanobacteria being the primary carbon fixers in these surface waters (Phoma et al. 2018). Both these groups are present in this study, however paralogs are not responsible for carbon cycling using the Calvin Cycle in these groups.

Another pathway used to investigate carbon cycling in literature is Photosystem I and II (Fischer et al. 2016; Phoma et al. 2018). Paralogs with the genomic potential for these pathways were observed in surface Cyanobacteria and Proteobacteria, however the

paralogs do not significantly complete the pathway and thus were not visible in the heatmap. Cyanobacteria are known to possess these pathways in surface waters around the PEIs (Phoma et al. 2018), however from our data it can be suggested that paralogs do not play a role in these pathways in the SIO. Genes responsible for anaerobic carbon fixation via rTCA have also been used to investigate carbon cycling (Acinas et al. 2021; Fischer et al. 2016; Phoma et al. 2018). This pathway was found to be the dominant energy mechanism in Fmax samples from the PEIs marine microbial communities (Phoma et al. 2018). Paralogs responsible for the rTCA pathway were not present at all in our data (Supplementary table S6), hence we can conclude that paralogs do not play a role in this pathway in the carbon cycle in the SIO.

4.6.3. Phosphorus cycling

Paralogs with the genomic potential for phosphorous cycling through AP were observed in this study. Paralogs encoding AP were present in MAGs at all stations except T08B, MT1-1 and CR07 (Figure 16). AP activity (Figure 8) follows the same trend as the concentration of phosphate (Figure 5) and the number of paralogs that encode AP (Figure 16), which could suggest that paralogs play a role in AP activity in marine microbial communities of the SIO. Recent gene duplication events in AP variants in dinoflagellates have previously been observed in this enzyme (Lin et al. 2015; O'Brie and Herschlag 2001). Paralogs responsible for phosphorus cycling pathways, including phosphate transporter, phosphonate transporter and bidirectional polyphosphate were observed in all the phyla in this study, except Marinisomatota, Patescibacteria and Thermoproteota (Figure 17). Bacteroidota paralogs showed the most complete bidirectional polyphosphate pathway amongst the MAGs. Bacteroidetes have been associated with

phosphorous removal in previous studies (Mehlig et al. 2013; Xiao et al. 2016), suggesting that Bacteroidota do indeed play a role in phosphorus cycling. Thus, it can be suggested that paralogs play a role in phosphorous cycling in Bacteroidota in the SIO. Verrucomicrobiota paralogs from 5 m, Fmax and Deep samples show the most complete phosphonate transporter pathway. The last phosphorus cycling pathway, the phosphate transporter pathway, is most complete by paralogs in Cyanobacteria from 5 m. Cyanobacteria are known to possess highly efficient phosphate uptake and retention mechanisms (Hudek et al. 2016; Ritchie et al. 2001), thus it can be suggested that gene duplication events play a role in the efficient uptake and retention of phosphate in this phylum. The presence of paralogs with the genomic potential for phosphorous metabolism in phyla known for phosphorous cycling suggests that gene duplicates do indeed play a role in phosphorous cycling in the SIO.

4.7. Pangenome

The 228 MAGs analyzed in this study did not contain a set of core genes, although core genes were identified at the phylum-level. The SAR324 group was specifically examined due to the presence of paralogs that partially complete biogeochemical cycling pathways used in this study. Among the 16 SAR324 members analyzed using draft genomes, 21 core genes were identified, with 75% sequence identity. These core genes were involved in various functions such as transcription, cell wall synthesis, carbon metabolism, and nitrogen metabolism, and included ribosomal proteins, tRNA-synthetases, ABC-type transporters, translation elongation factors, topoisomerases, polymerases, and ATP/GTPases.

Four carbon metabolism-related proteins, and one protein with potential for both nitrogen and carbon metabolism were among the SAR324 core genes identified in this study (Holness and Sugden 2003; Kim et al. 2004; Melo et al. 2004; Peng et al. 2011; Wiegand and Remington 1986; Yao et al. 2011). SAR324 members occupying surface waters were found to exhibit a higher degree of plasticity in energy-related metabolic pathways, indicating a photoheterotrophic lifestyle (Boeuf et al. 2021). The SAR324 core genome contained several genes related to energy metabolism, including citrate synthase 1 (gltA2_2), the dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex (aceF), putative 3-oxopropanoate dehydrogenase (bauC), 2-hydroxy-3-oxopropionate reductase (glxR), and chloroplastic NAD(P)H-quinone oxidoreductase subunit K (ndhK) (Holness and Sugden 2003; Kim et al. 2004; Melo et al. 2004; Peng et al. 2011; Wiegand and Remington 1986; Yao et al. 2011).

While genes involved in carbon and nitrogen metabolism were present in the SAR324 core genome, other biogeochemical cycling genes were found in the accessory or strain-specific genome. Four genes encoding alkaline phosphatase were identified in the SAR324 pangenome. These four genes encoded multifunctional alkaline phosphatase superfamily proteins, with two shared amongst 12 members, two shared amongst two members, and one specific to C29. The presence of a strain-specific multifunctional alkaline phosphatase superfamily protein gene in C29 suggests that this gene copy has enabled this organism to adapt to conditions in the SIO (Brockhurst et al. 2019; Fraser-Liggett 2005), potentially through the use of alkaline phosphatase in surface waters. Further experimental investigation would be necessary to confirm the role of these strain-

specific genes in adaptation to specific environments (Bezuidt et al. 2016; Nelson et al. 2016).

5. References

Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, Royo-Llonch M, Paoli L, Sunagawa S, Hingamp P. 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun. Biol. 4(1)**:** 1-15. https://doi.org/10.1038/s42003-021-02112-2.

Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. 2017. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. 11(11)**:** 2407-2425. https://doi.org/10.1038/ismej.2017.122.

Altabet MA, Francois R. 2001. Nitrogen isotope biogeochemistry of the Antarctic Polar Frontal Zone at 170 W. Deep Sea Res. Part II Top. Stud. Oceanogr. 48(19-20)**:** 4247-4273. https://doi.org/10.1016/S0967-0645(01)00088-1.

Álvarez-Lugo A, Becerra A. 2021. The role of gene duplication in the divergence of enzyme function: a comparative approach. Front. Genet. 12(https://doi.org/10.3389/fgene.2021.641817.

Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber JA. 2017. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat. Commun. 8(1)**:** 1-11. https://doi.org/10.1038/s41467-017-01228-6.

Ansorge I, Lutjeharms J. 2002. The hydrography and dynamics of the ocean environment of the Prince Edward Islands (Southern Ocean). J. Mar. Syst. 37(1-3)**:** 107-127. https://doi.org/10.1016/S0924-7963(02)00198-7.

Arnosti C, Bell C, Moorhead D, Sinsabaugh R, Steen A, Stromberger M, Wallenstein M, Weintraub M. 2014. Extracellular enzymes in terrestrial, freshwater, and marine environments: perspectives on system variability and common research needs. Biogeochemistry. 117(1)**:** 5-21. https://doi.org/10.1007/s10533-013-9906-5.

Aylward FO, Santoro AE. 2020. Heterotrophic Thaumarchaea with small genomes are widespread in the dark ocean. mSystems. 5(3)**:** e00415-00420. https://doi.org/10.1128/mSystems.00415-20.

Baker BJ, Sheik CS, Taylor CA, Jain S, Bhasi A, Cavalcoli JD, Dick GJ. 2013. Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. ISME J. 7(10)**:** 1962-1973. https://doi.org/10.1038/ismej.2013.85.

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature. 461(7268)**:** 1243-1247.

Bauer M, Kube M, Teeling H, Richter M, Lombardot T, Allers E, Würdemann CA, Quast C, Kuhl H, Knaust F. 2006. Whole genome analysis of the marine Bacteroidetes '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. Environ. Microbiol. 8(12)**:** 2201-2213.

Bayer K, Busch K, Kenchington E, Beazley L, Franzenburg S, Michels J, Hentschel U, Slaby BM. 2020. Microbial strategies for survival in the glass sponge *Vazella*

pourtalesii. mSystems. 5(4): e00473-00420. https://doi.org/10.1111/j.1462-2920.2006.01152.x.
Belkin IM, Gordon AL. 1996. Southern Ocean fronts from the Greenwich meridian to Tasmania. J. Geophys. Res. Oceans. 101(C2): 3675-3696. https://doi.org/10.1029/95JC02750.
Berg IA, Kockelkorn D, Buckel W, Fuchs G. 2007. A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. Science. 318(5857): 1782-1786. https://doi.org/10.1126/science.1149976.
Berman-Frank I, Quigg A, Finkel ZV, Irwin AJ, Haramaty L. 2007. Nitrogen‐fixation strategies and Fe requirements in cyanobacteria. Limnol. Oceanogr. 52(5): 2260-2269. https://doi.org/10.4319/lo.2007.52.5.2260.
Bertagnolli AD, Padilla CC, Glass JB, Thamdrup B, Stewart FJ. 2017. Metabolic potential and in situ activity of marine Marinimicrobia bacteria in an anoxic water column. Environ. Microbiol. 19(11): 4392-4416. https://doi.org/10.1111/1462-2920.13879.
Bezuidt OK, Pierneef R, Gomri AM, Adesioye F, Makhalanyane TP, Kharroub K, Cowan DA. 2016. The *Geobacillus* pan-genome: implications for the evolution of the genus. Front. Microbiol. 7(723): https://doi.org/10.3389/fmicb.2016.00723.
Björkman K, Karl DM. 1994. Bioavailability of inorganic and organic phosphorus compounds to natural assemblages of microorganisms in Hawaiian coastal waters. Mar. Ecol. Prog. Ser. 265-273.
Boden BP. 1988. Observations of the island mass effect in the Prince Edward Archipelago. Polar Biol. 9(1): 61-68. https://doi.org/10.1007/BF00441765.
Boeuf D, Eppley JM, Mende DR, Malmstrom RR, Woyke T, DeLong EF. 2021. Metapangenomics reveals depth-dependent shifts in metabolic potential for the ubiquitous marine bacterial SAR324 lineage. Microbiome. 9(1): 1-18. https://doi.org/10.1186/s40168-021-01119-5.
Brockhurst MA, Harrison E, Hall JP, Richards T, McNally A, MacLean C. 2019. The ecology and evolution of pangenomes. Curr. Biol. 29(20): R1094-R1103. https://doi.org/10.1016/j.cub.2019.08.012.
Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. Genome Res. 30(3): 315-333. http://doi:10.1101/gr.258640.119.
Chirakul S, Somprasong N, Norris MH, Wuthiekanun V, Chantratita N, Tuanyok A, Schweizer HP. 2019. *Burkholderia pseudomallei* acquired ceftazidime resistance due to gene duplication and amplification. Int. J. Antimicrob. Agents. 53(5): 582-588. https://doi.org/10.1016/j.ijantimicag.2019.01.003.
Chistoserdova L. 2011. Modularity of methylotrophy, revisited. Environ. Microbiol. 13(10): 2603-2622. https://doi.org/10.1111/j.1462-2920.2011.02464.x.
Chrost RJ. 1991. Environmental control of the synthesis and activity of aquatic microbial ectoenzymes. In: Microbial enzymes in aquatic environments. Springer. p. 29-59.
Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9(12): 938-950. https://doi.org/10.1038/nrg2482.
Cory RM, Ward CP, Crump BC, Kling GW. 2014. Sunlight controls water column processing of carbon in arctic fresh waters. Science. 345(6199): 925-928. https://doi.org/10.1126/science.1253119.

Coutinho F, von Meijenfeldt F, Walter J, Haro-Moreno J, Lopéz-Pérez M, van Verk M, Thompson C, Cosenza C, Appolinario L, Paranhos R. 2021. Ecogenomics and metabolic potential of the South Atlantic Ocean microbiome. Sci. Total Environ. 765(142758. https://doi.org/10.1016/j.scitotenv.2020.142758.

Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee S, Rappé MS, McLellan SL, Lücker S, Eren AM. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat. Microbiol. 3(7)**:** 804-813. https://doi.org/10.1038/s41564-018-0176-9.

DeLong EF, Franks DG, Alldredge AL. 1993. Phylogenetic diversity of aggregate‐attached vs. free‐living marine bacterial assemblages. Limnol. Oceanogr. 38(5)**:** 924-934. https://doi.org/10.4319/lo.1993.38.5.0924.

Diehl SF, Budd JW, Ullman D, Cayula J-F. 2002. Geographic window sizes applied to remote sensing sea surface temperature front detection. J. Atmos. Ocean. Technol. 19(7)**:** 1105-1113. https://doi.org/10.1175/1520-0426(2002)019%3C1105:GWSATR%3E2.0.CO;2.

Duhamel S, Dyhrman ST, Karl DM. 2010. Alkaline phosphatase activity and regulation in the North Pacific Subtropical Gyre. Limnol. Oceanogr. 55(3)**:** 1414-1425. https://doi.org/10.4319/lo.2010.55.3.1414.

Easson CG, Lopez JV. 2019. Depth-Dependent environmental drivers of microbial plankton community structure in the Northern Gulf of Mexico. Front. Microbiol. 9(3175. https://doi.org/10.3389/fmicb.2018.03175.

Elliott KT, Cuff LE, Neidle EL. 2013. Copy number change: evolving views on gene amplification. Future Microbiol. 8(7)**:** 887-899. https://doi.org/10.2217/fmb.13.53.

Emery WJ. 2001. Water types and water masses. Encyclopedia of ocean sciences. 6(3179-3187.

Farías L, Fernández C, Faúndez J, Cornejo M, Alcaman M. 2009. Chemolithoautotrophic production mediating the cycling of the greenhouse gases N2O and CH4 in an upwelling ecosystem. Biogeosciences. 6(12)**:** 3053-3069.

Fewer DP, Metsä‐Ketelä M. 2020. A pharmaceutical model for the molecular evolution of microbial natural products. FEBS J. 287(7)**:** 1429-1449. https://doi.org/10.1111/febs.15129.

Fischer WW, Hemp J, Johnson JE. 2016. Evolution of oxygenic photosynthesis. Annu. Rev. Earth Planet. Sci. 44(647-683. https://doi.org/10.1146/annurev-earth-060313-054810.

Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res. 15(12)**:** 1603-1610. https://doi.org/10.1101/gr.3724205.

Galbe M, Zacchi G. 2002. A review of the production of ethanol from softwood. Appl. Microbiol. Biotechnol. 59(6)**:** 618-628. https://doi.org/10.1007/s00253-002-1058-9.

Garcia NS, Fu F, Sedwick PN, Hutchins DA. 2015. Iron deficiency increases growth and nitrogen-fixation rates of phosphorus-deficient marine cyanobacteria. ISME J. 9(1)**:** 238-245. https://doi.org/10.1038/ismej.2014.104.

Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12(4)**:** 148-154. https://doi.org/10.1016/j.tim.2004.02.007.

Glöckner FO, Gasol JM, McDonough N, Calewaert J-B. 2012. Marine microbial diversity and its role in ecosystem functioning and environmental change. European Science Foundation.
Gómez-Anduro G, Ceniceros-Ojeda EA, Casados-Vázquez LE, Bencivenni C, Sierra-Beltrán A, Murillo-Amador B, Tiessen A. 2011. Genome-wide analysis of the beta-glucosidase gene family in maize (*Zea mays* L. *var B73*). Plant Mol. Biol. 77(1)**:** 159-183. https://doi.org/10.1007/s11103-011-9800-2.
Grasshof K. 1976. Methods of seawater analysis. Verlag Chemie, Weinheim and New York.
GÜnter J, Zubkov MV, Yakushev E, Labrenz M, Jürgens K. 2008. High abundance and dark CO2 fixation of chemolithoautotrophic prokaryotes in anoxic waters of the Baltic Sea. Limnol. Oceanogr. 53(1)**:** 14-22. https://doi.org/10.4319/lo.2008.53.1.0014.
Hanson RS, Hanson TE. 1996. Methanotrophic bacteria. Microbiol. Rev. 60(2)**:** 439-471. https://doi.org/10.1128/mr.60.2.439-471.1996.
Hawley AK, Brewer HM, Norbeck AD, Paša-Tolić L, Hallam SJ. 2014. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. Proc. Natl. Acad. Sci. 111(31)**:** 11395-11400. https://doi.org/10.1073/pnas.1322132111.
Henley SF, Cavan EL, Fawcett SE, Kerr R, Monteiro T, Sherrell RM, Bowie AR, Boyd PW, Barnes DK, Schloss IR. 2020. Changing biogeochemistry of the Southern Ocean and its ecosystem implications. Front. Mar. Sci. 7(581. https://doi.org/10.3389/fmars.2020.00581.
Herdyastuti N, Raharjo TJ, Mudasir M, Matsjeh S. 2009. Chitinase and chitinolytic microorganism: Isolation, characterization and potential. Indian J. Chem. 9(1)**:** 37-47. https://doi.org/10.22146/ijc.21580.
Hermes JC, Masumoto Y, Beal L, Roxy MK, Vialard J, Andres M, Annamalai H, Behera S, D'Adamo N, Feng M. 2019. A sustained ocean observing system in the Indian Ocean for climate related scientific knowledge and societal needs. Front. Mar. Sci. 355. https://doi.org/10.3389/fmars.2019.00355.
Herndl GJ, Reinthaler T, Teira E, van Aken H, Veth C, Pernthaler A, Pernthaler J. 2005. Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. Appl. Environ. Microbiol. 71(5)**:** 2303-2309. https://doi.org/10.1128/AEM.71.5.2303-2309.2005.
Herrero A, Muro-Pastor AM, Flores E. 2001. Nitrogen control in cyanobacteria. J. Bacteriol. 183(2)**:** 411-425. https://doi.org/10.1128/JB.183.2.411-425.2001.
Holness M, Sugden M. 2003. Regulation of pyruvate dehydrogenase complex activity by reversible phosphorylation. Biochem. Soc. Trans. 31(6)**:** 1143-1151. https://doi.org/10.1042/bst0311143.
Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. 20(6)**:** 945-954.
Hoppe H-G. 2003. Phosphatase activity in the sea. Hydrobiologia. 493(https://doi.org/10.1023/A:1025453918247.
Hudek L, Premachandra D, Webster W, Bräu L. 2016. Role of phosphate transport system component PstB1 in phosphate internalization by *Nostoc punctiforme*.

Appl. Environ. Microbiol. 82(21): 6344-6356. https://doi.org/10.1128/AEM.01336-16.
Ikeda S, Ytow N, Ezura H, Minamisawa K, Miyashita K, Fujimura T. 2007. Analysis of molecular diversity of bacterial chitinase genes in the maize rhizosphere using culture-independent methods. Microbes Environ. 22(1): 71-77. https://doi.org/10.1264/jsme2.22.71.
Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M. 2002. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. Appl. Environ. Microbiol. 68(5): 2391-2396. https://doi.org/10.1128/AEM.68.5.2391-2396.2002.
Jasmine P, Muraleedharan K, Madhu N, Devi CA, Alagarsamy R, Achuthankutty C, Jayan Z, Sanjeevan V, Sahayak S. 2009. Hydrographic and productivity characteristics along 45 E longitude in the southwestern Indian Ocean and Southern Ocean during austral summer 2004. Mar. Ecol. Prog. Ser. 389(97-116.
Jetten MS, Sliekers O, Kuypers M, Dalsgaard T, van Niftrik L, Cirpus I, van de Pas-Schoonen K, Lavik G, Thamdrup B, Le Paslier D. 2003. Anaerobic ammonium oxidation by marine and freshwater planctomycete-like bacteria. Appl. Microbiol. Biotechnol. 63(2): 107-114. https://doi.org/10.1007/s00253-003-1422-4.
Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. 11(4): 555-565.
Kaehler S, Pakhomov E, McQuaid C. 2000. Trophic structure of the marine food web at the Prince Edward Islands (Southern Ocean) determined by δ13C and δ15N analysis. Mar. Ecol. Prog. Ser. 208(13-20. http://doi.org/10.3354/meps208013.
Kamneva OK, Knight SJ, Liberles DA, Ward NL. 2012. Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol. Evol. 4(12): 1375-1390. https://doi.org/10.1093/gbe/evs113.
Karl DM. 2000. Phosphorus, the staff of life. Nature. 406(6791): 31-33. https://doi.org/10.1038/35017683.
Karner MB, DeLong EF, Karl DM. 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature. 409(6819): 507-510. https://doi.org/10.1038/35054051.
Kemeny PC, Weigand M, Zhang R, Carter BR, Karsh KL, Fawcett SE, Sigman DM. 2016. Enzyme‐level interconversion of nitrate and nitrite in the fall mixed layer of the Antarctic Ocean. Global Biogeochem. Cy. 30(7): 1069-1085. https://doi.org/10.1002/2015GB005350.
Kim HJ, Kim TH, Kim Y, Lee HS. 2004. Identification and characterization of *glxR*, a gene involved in regulation of glyoxylate bypass in *Corynebacterium glutamicum*. J. Bacteriol. 186(11): 3453-3460. https://doi.org/10.1128/jb.186.11.3453-3460.2004.
Kim HW, Kim NK, Phillips AP, Parker DA, Liu P, Whitaker RJ, Rao CV, Mackie RI. 2022. Genome sequence of a thermoacidophilic methanotroph belonging to the Verrucomicrobiota phylum from geothermal hot springs in Yellowstone National Park: A metagenomic assembly and reconstruction. Microorganisms. 10(1): 142. https://doi.org/10.3390/microorganisms10010142.
Kim JW, Brawley SH, Prochnik S, Chovatia M, Grimwood J, Jenkins J, LaButti K, Mavromatis K, Nolan M, Zane M, Schmutz J, Stiller JW, Grossman AR. 2016.

Genome analysis of Planctomycetes inhabiting blades of the red alga *Porphyra umbilicalis*. PLoS One. 11(3)**:** https://doi.org/10.1371/journal.pone.0151883.
King GM. 1984. Metabolism of trimethylamine, choline, and glycine betaine by sulfate-reducing and methanogenic bacteria in marine sediments. Appl. Environ. Microbiol. 48(4)**:** 719-725. https://doi.org/10.1128/aem.48.4.719-725.1984.
Kirkman S, Mann B, Sink K, Adams R, Livingstone T, Mann-Lang J, Pfaff M, Samaai T, van der Bank M, Williams L. 2021. Evaluating the evidence for ecological effectiveness of South Africa's marine protected areas. Afr. J. Mar. Sci. 43(3)**:** 389-412.
Kirkman SP, Yemane DG, Lamont T, Meyer MA, Pistorius PA. 2016. Foraging behavior of subantarctic fur seals supports efficiency of a marine reserve's design. PLoS One. 11(5)**:** e0152370.
Knight TR, Dick RP. 2004. Differentiating microbial and stabilized β-glucosidase activity relative to soil quality. Soil Biol. Biochem. 36(12)**:** 2089-2096. https://doi.org/10.1016/j.soilbio.2004.06.007.
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B: Biol. Sci. 279(1749)**:** 5048-5057. https://doi.org/10.1098/rspb.2012.1108.
Kuypers MM, Marchant HK, Kartal B. 2018. The microbial nitrogen-cycling network. Nat. Rev. Microbiol. 16(5)**:** 263-276. https://doi.org/10.1038/nrmicro.2018.9.
Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am. Nat. 138(6)**:** 1315-1341.
Lenski RE, Travisano M. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. 91(15)**:** 6808-6814.
Lin X, Wang L, Shi X, Lin S. 2015. Rapidly diverging evolution of an atypical alkaline phosphatase (PhoAaty) in marine phytoplankton: insights from dinoflagellate alkaline phosphatases. Front. Microbiol. 6(868. https://doi.org/10.3389/fmicb.2015.00868.
Liu R, Wang L, Liu Q, Wang Z, Li Z, Fang J, Zhang L, Luo M. 2018. Depth-resolved distribution of particle-attached and free-living bacterial communities in the water column of the New Britain Trench. Front. Microbiol. 9(625. https://doi.org/10.3389/fmicb.2018.00625.
Loder AJ, Han Y, Hawkins AB, Lian H, Lipscomb GL, Schut GJ, Keller MW, Adams MW, Kelly RM. 2016. Reaction kinetic analysis of the 3-hydroxypropionate/4-hydroxybutyrate CO2 fixation cycle in extremely thermoacidophilic archaea. Metab. Eng. 38(446-463. https://doi.org/10.1016/j.ymben.2016.10.009.
Lombard A, Reyers B, Schonegevel L, Cooper J, Smith-Adao L, Nel D, Froneman PW, Ansorge IJ, Bester M, Tosh C. 2007. Conserving pattern and process in the Southern Ocean: designing a marine protected area for the Prince Edward Islands. Antarctic Science. 19(1)**:** 39-54. https://doi.org/10.1017/S0954102007000077.
Lønborg C, Cuevas LA, Reinthaler T, Herndl GJ, Gasol JM, Morán XAG, Bates NR, Álvarez-Salgado XA. 2016. Depth dependent relationships between temperature and ocean heterotrophic prokaryotic production. Front. Mar. Sci. 3(90. https://doi.org/10.3389/fmars.2016.00090.

Luo L, Meng H, Gu J-D. 2017. Microbial extracellular enzymes in biogeochemical cycling of ecosystems. J. Environ. Manag. 197(539-549. https://doi.org/10.1016/j.jenvman.2017.04.023.
Lutjeharms JR. 2006. The coastal oceans of south-eastern Africa. The sea. 14(783-834.
Mehlig L, Petzold M, Heder C, Günther S, Müller S, Eschenhagen M, Röske I, Röske K. 2013. Biodiversity of polyphosphate accumulating bacteria in eight WWTPs with different modes of operation. J. Environ. Eng. 139(1089-1098. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000711.
Melo AM, Bandeiras TM, Teixeira M. 2004. New insights into type II NAD(P)H: quinone oxidoreductases. Microbiol. Mol. Biol. Rev. 68(4)**:** 603-616. https://doi.org/10.1128/MMBR.68.4.603-616.2004.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17(10)**:** 589-596. https://doi.org/10.1016/S0168-9525(01)02447-7.
Misic C, Povero P, Fabiano M. 2002. Ectoenzymatic ratios in relation to particulate organic matter distribution (Ross Sea, Antarctica). Microb. Ecol. 44(3)**:** 224-234. https://doi.org/10.1007/s00248-002-2017-9.
Moghe GD, Last RL. 2015. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol. 169(3)**:** 1512-1523. https://doi.org/10.1104/pp.15.00994.
Mohammadi S, Pol A, van Alen TA, Jetten MS, Op den Camp HJ. 2017. *Methylacidiphilum fumariolicum* SolV, a thermoacidophilic 'Knallgas' methanotroph with both an oxygen-sensitive and-insensitive hydrogenase. ISME J. 11(4)**:** 945-958. https://doi.org/10.1038/ismej.2016.171.
Morrissey EM, Gillespie JL, Morina JC, Franklin RB. 2014. Salinity affects microbial activity and soil organic matter content in tidal wetlands. Global Change Biol. 20(4)**:** 1351-1362. https://doi.org/10.1111/gcb.12431.
Naqvi SWA, Bange HW, Gibb SW, Goyet C, Hatton AD, Upstill-Goddard RC. 2005. Biogeochemical ocean-atmosphere transfers in the Arabian Sea. Prog. Oceanogr. 65(2-4)**:** 116-144. https://doi.org/10.1016/j.pocean.2005.03.005.
Nelson DD, Duprau JL, Wolff PL, Evermann JF. 2016. Persistent *Bovine Viral Diarrhea* virus infection in domestic and wild small ruminants and camelids including the mountain goat (*Oreamnos americanus*). Front. Microbiol. 6(https://doi.org/10.3389/fmicb.2015.01415.
O'Brie PJ, Herschlag D. 2001. Functional interrelationships in the alkaline phosphatase superfamily: phosphodiesterase activity of *Escherichia coli* alkaline phosphatase. Biochemistry. 40(19)**:** 5691-5699. https://doi.org/10.1021/bi0028892.
Ohno S. 1970. Evolution by gene duplication. . London-New York: Allen & Unwin.
Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. 2011. Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiol. Mol. Biol. Rev. 75(2)**:** 361-422. https://doi.org/10.1128/MMBR.00039-10.
Pedler BE, Aluwihare LI, Azam F. 2014. Single bacterial strain capable of significant contribution to carbon cycling in the surface ocean. Proc. Natl. Acad. Sci. 111(20)**:** 7202-7207. https://doi.org/10.1073/pnas.1401887111.

Peng L, Yamamoto H, Shikanai T. 2011. Structure and biogenesis of the chloroplast NAD(P)H dehydrogenase complex. Biochim. Biophys. Acta - Bioenerg. 1807(8)**:** 945-953. https://doi.org/10.1016/j.bbabio.2010.10.015.

Pérez Castro S, Borton MA, Regan K, Hrabe de Angelis I, Wrighton KC, Teske AP, Strous M, Ruff SE. 2021. Degradation of biological macromolecules supports uncultured microbial populations in Guaymas Basin hydrothermal sediments. ISME J. 15(12)**:** 3480-3497. https://doi.org/10.1038/s41396-021-01026-5.

Perissinotto R, Rae CD. 1990. Occurrence of anticyclonic eddies on the Prince Edward Plateau (Southern Ocean): effects on phytoplankton biomass and production. Deep Sea Res. Part I Oceanogr. Res. Pap. 37(5)**:** 777-793. https://doi.org/10.1016/0198-0149(90)90006-H.

Pester M, Schleper C, Wagner M. 2011. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. Curr. Opin. Microbiol. 14(3)**:** 300-306. https://doi.org/10.1016/j.mib.2011.04.007.

Phoma BS, Makhalanyane TP. 2021. Depth-dependent variables shape community structure and functionality in the Prince Edward Islands. Microb. Ecol. 81(2)**:** 396-409. https://doi.org/10.1007/s00248-020-01589-4.

Phoma S, Vikram S, Jansson JK, Ansorge IJ, Cowan DA, Van de Peer Y, Makhalanyane TP. 2018. Agulhas Current properties shape microbial community diversity and potential functionality. Sci. Rep. 8(1)**:** 1-12. https://doi.org/10.1038/s41598-018-28939-0.

Price PB, Sowers T. 2004. Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. Proc. Natl. Acad. Sci. 101(13)**:** 4631-4636. https://doi.org/10.1073/pnas.0400522101.

Raes J, Van de Peer Y. 2003. Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates *in silico*. Appl. Bioinform. 2(2)**:** 91-101.

Raghukumar C, Muraleedharan U, Gaud V, Mishra R. 2004. Xylanases of marine fungi of potential use for biobleaching of paper pulp. J. Ind. Microbiol. Biotechnol. 31(9)**:** 433-441. https://doi.org/10.1007/s10295-004-0165-2.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. Curr. Opin. Plant Biol. 17(43-48. https://doi.org/10.1016/j.pbi.2013.11.002.

Riehle MM, Bennett AF, Long AD. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. Proc. Natl. Acad. Sci. 98(2)**:** 525-530. https://doi.org/10.1073/pnas.98.2.525.

Rinke C, Rubino F, Messer LF, Youssef N, Parks DH, Chuvochina M, Brown M, Jeffries T, Tyson GW, Seymour JR. 2019. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (*Ca.* Poseidoniales ord. nov.). ISME J. 13(3)**:** 663-675. https://doi.org/10.1038/s41396-018-0282-y.

Ritchie RJ, Trautman DA, Larkum AW. 2001. Phosphate limited cultures of the cyanobacterium *Synechococcus* are capable of very rapid, opportunistic uptake of phosphate. New Phytol. 152(2)**:** 189-201. https://doi.org/10.1046/j.0028-646X.2001.00264.x.

Rush D, Sinninghe Damsté JS. 2017. Lipids as paleomarkers to constrain the marine nitrogen cycle. Environ. Microbiol. 19(6)**:** 2119-2132. https://doi.org/10.1111/1462-2920.13682.

Sánchez-Baracaldo P, Bianchini G, Wilson JD, Knoll AH. 2022. Cyanobacteria and biogeochemical cycles through Earth history. Trends in Microbiol. 30(2)**:** 143-157. https://doi.org/10.1016/j.tim.2021.05.008.

Santhi R. 2016. Isolation of chitinase producing *Streptomyces albus* FS12, production and optimization of extracellular chitinase. Int. J. Adv. Res. Biol. Sci. 3(4)**:** 229-237.

Schunck H, Lavik G, Desai DK, Großkopf T, Kalvelage T, Löscher CR, Paulmier A, Contreras S, Siegel H, Holtappels M. 2013. Giant hydrogen sulfide plume in the oxygen minimum zone off Peru supports chemolithoautotrophy. PLoS One. 8(8)**:** e68661. https://doi.org/10.1371/journal.pone.0068661.

Seitzinger S, Harrison J, Dumont E, Beusen AH, Bouwman A. 2005. Sources and delivery of carbon, nitrogen, and phosphorus to the coastal zone: An overview of Global Nutrient Export from Watersheds (NEWS) models and their application. Global Biogeochem. Cy. 19(4)**:** https://doi.org/10.1029/2005GB002606.

Serres MH, Kerr AR, McCormack TJ, Riley M. 2009. Evolution by leaps: gene duplication in bacteria. Biol. Direct. 4(1)**:** 1-17. https://doi.org/10.1186/1745-6150-4-46.

Sheik CS, Jain S, Dick GJ. 2014. Metabolic flexibility of enigmatic SAR 324 revealed through metagenomics and metatranscriptomics. Environ. Microbiol. 16(1)**:** 304-317. https://doi.org/10.1111/1462-2920.12165.

Sheridan PO, Raguideau S, Quince C, Holden J, Zhang L, Williams TA, Gubry-Rangin C. 2020. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat. Commun. 11(1)**:** 1-12. https://doi.org/10.1038/s41467-020-19132-x.

Shi W, Bowman D, Rufty T. 2012. Microbial control of soil carbon accumulation in turfgrass systems. In: Carbon Sequestration in Urban Ecosystems. Dordrecht: Springer. p. 215-231.

Sigman DM, Altabet M, McCorkle D, Francois R, Fischer G. 2000. The δ15N of nitrate in the Southern Ocean: Nitrogen cycling and circulation in the ocean interior. J. Geophys. Res. Oceans. 105(C8)**:** 19599-19614.

Stanier R, Cohen-Bazire G. 1977. Phototrophic prokaryotes: the cyanobacteria. Annu. Rev. Microbiol. 31(1)**:** 225-274. https://doi.org/10.1146/annurev.mi.31.100177.001301.

Stock CA, Cheung WW, Sarmiento JL, Sunderland EM. 2019. Changing ocean systems: a short synthesis. In: Predicting Future Oceans. Elsevier. p. 19-34.

Strous M, Fuerst JA, Kramer EH, Logemann S, Muyzer G, van de Pas-Schoonen KT, Webb R, Kuenen JG, Jetten MS. 1999. Missing lithotroph identified as new planctomycete. Nature. 400(6743)**:** 446-449. https://doi.org/10.1038/22749.

Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P. 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature. 440(7085)**:** 790-794. https://doi.org/10.1038/nature04647.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A. 2015. Structure and function of the global ocean microbiome. Science. 348(6237)**:** 1261359. https://doi.org/10.1126/science.1261359.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML. 2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science. 333(6047)**:** 1296-1300.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonźalez JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc. Natl. Acad. Sci. U.S.A. 110(28)**:** 11463-11468. 10.1073/pnas.1304246110.

Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G. 2011. Environmental and gut bacteroidetes: the food connection. Front. Microbiol. 2(93. https://doi.org/10.3389/fmicb.2011.00093.

Thomson B, Wenley J, Currie K, Hepburn C, Herndl GJ, Baltar F. 2019. Resolving the paradox: continuous cell-free alkaline phosphatase activity despite high phosphate concentrations. Mar. Chem. 214(103671. https://doi.org/10.1016/j.marchem.2019.103671.

Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, Shi W, Wu L, Zhang Y, Yang Y. 2020. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. Microbiome. 8(1)**:** 1-15. https://doi.org/10.1186/s40168-020-00825-w.

Tiwari R, Kumar K, Singh S, Nain L, Shukla P. 2016. Molecular detection and environment-specific diversity of glycosyl hydrolase family 1 β-glucosidase in different habitats. Front. Microbiol. 7(1597. https://doi.org/10.3389/fmicb.2016.01597.

Treasure AM, Ruzicka JJ, Moloney CL, Gurney LJ, Ansorge IJ. 2015. Land–sea interactions and consequences for sub-antarctic marine food webs. Ecosyst. 18(5)**:** 752-768. https://doi.org/10.1007/s10021-015-9860-2.

Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data. 5(1)**:** 1-8. https://doi.org/10.1038/sdata.2017.203.

Tyrrell T. 1999. The relative influences of nitrogen and phosphorus on oceanic primary production. Nature. 400(6744)**:** 525-531. https://doi.org/10.1038/22941.

Venkatachalam S, Matcher GF, Lamont T, van den Berg M, Ansorge IJ, Dorrington RA. 2019. Influence of oceanographic variability on near‐shore microbial communities of the sub‐Antarctic Prince Edward Islands. Limnol. Oceanogr. 64(1)**:** 258-271. https://doi.org/10.1002/lno.11035.

Virdi JS, Sachdeva P. 2005. Genetic diversity of pathogenic microorganisms: Basic insights, public health implications and the Indian initiatives. Curr. Sci. 113-123.

Voss M, Bange HW, Dippner JW, Middelburg JJ, Montoya JP, Ward B. 2013. The marine nitrogen cycle: recent discoveries, uncertainties and the potential relevance of climate change. Philos. Trans. R. Soc. B: Biol. Sci. 368(1621)**:** 20130121. https://doi.org/10.1098/rstb.2013.0121.

Wagner M, Horn M. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr. Opin. Biotechnol. 17(3)**:** 241-249. https://doi.org/10.1016/j.copbio.2006.05.005.

Wang X, Zhao Y, Tan H, Chi N, Zhang Q, Du Y, Yin H. 2014. Characterisation of a chitinase from *Pseudoalteromonas sp.* DL-6, a marine psychrophilic bacterium. Int. J. Biol. Macromol. 70(455-462. https://doi.org/10.1016/j.ijbiomac.2014.07.033.

Wiegand G, Remington SJ. 1986. Citrate synthase: structure, control, and mechanism. Annu. Rev. Biophys. Biophys. Chem. 15(1)**:** 97-117. https://doi.org/10.1146/annurev.bb.15.060186.000525.

Wing S, Jack L, Shatova O, Leichter J, Barr D, Frew R, Gault-Ringold M. 2014. Seabirds and marine mammals redistribute bioavailable iron in the Southern Ocean. Mar. Ecol. Prog. Ser. 510(1-13. https://doi.org/10.3354/meps10923.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. Science. 342(6164)**:** 1364-1367.

Xiao K, Zhou L, He B, Qian L, Wan S, Qu L. 2016. Nitrogen and phosphorus removal using fluidized-carriers in a full-scale A2O biofilm system. Biochem. Eng. J. 115(47-55. https://doi.org/10.1016/j.bej.2016.08.004.

Yao X, He W, Lu C-D. 2011. Functional characterization of seven γ-glutamylpolyamine synthetase genes and the *bauRABCD* locus for polyamine and β-alanine utilization in *Pseudomonas aeruginosa* PAO1. J. Bacteriol. 193(15)**:** 3923-3930. https://doi.org/10.1128/JB.05105-11.

Zehr JP, Kudela RM. 2011. Nitrogen cycle of the open ocean: from genes to ecosystems. Annu. Rev. Mar. Sci. 3(197-225. https://doi.org/10.1146/annurev-marine-120709-142819.

Zeng X, Alain K, Shao Z. 2021a. Microorganisms from deep-sea hydrothermal vents. Mar. Life sci. Technol. 3(204-230. https://doi.org/10.1007/s42995-020-00086-4.

Zeng YX, Luo W, Li HR, Yu Y. 2021b. High diversity of planktonic prokaryotes in Arctic Kongsfjorden seawaters in summer 2015. Polar Biol. 44(1)**:** 195-208. https://doi.org/10.1007/s00300-020-02791-3.

Zhang CL, Xie W, Martin-Cuadrado A-B, Rodriguez-Valera F. 2015. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. Front. Microbiol. 6(1108. https://doi.org/10.3389/fmicb.2015.01108.

Zhang X, Ward BB, Sigman DM. 2020. Global nitrogen cycle: critical enzymes, organisms, and processes for nitrogen budgets and dynamics. Chem. Rev. 120(12)**:** 5308-5351. https://doi.org/10.1021/acs.chemrev.9b00613.

Zhao Y, Liu P, Rui J, Cheng L, Wang Q, Liu X, Yuan Q. 2020. Dark carbon fixation and chemolithotrophic microbial community in surface sediments of the cascade reservoirs, Southwest China. Sci. Total Environ. 698(134316. https://doi.org/10.1016/j.scitotenv.2019.134316.

Zhuang GC, Peña‐Montenegro TD, Montgomery A, Hunter KS, Joye SB. 2018. Microbial metabolism of methanol and methylamine in the Gulf of Mexico: insight into marine carbon and nitrogen cycling. Environ. Microbiol. 20(12)**:** 4543-4554. https://doi.org/10.1111/1462-2920.14406.

Zobell CE, Rittenberg SC. 1938. The occurrence and characteristics of chitinoclastic bacteria in the sea. J. Bacteriol. 35(3)**:** 275-287. https://doi.org/10.1128%2Fjb.35.3.275-287.1938.

Chapter 4

Synthesis

Microbial communities are known to provide important ecosystem services. However, the extent to which intra- and inter-community mechanisms shape the provision of these services remains unclear. An example of these unclear mechanisms includes the extent to which evolutionary processes, such as gene duplications, shape and influence ecosystem services.

Previous studies on gene duplications have focused on individual species rather than trends in duplication events in microbial communities at different depths or distances from island masses. This study used shotgun metagenomic sequencing to determine the trends and extent of gene duplications in marine microbial communities of the SIO. The extent of gene duplications found within a genome appear to be linked to the phylum to which the organism belongs. In this study, Proteobacteria, Actinobacteria and Verrucomicrobiota showed the highest levels of gene duplications. These phyla occupy similar niches in soils, and have been clustered together phylogenetically as taxa that share the same niche. Consequently, it is reasonable to predict that these phyla may possess gene duplications in similar proportions, with similar ecological roles. Patescibacteria possessed the lowest mean number of paralogs in this study, which was expected because the phylum possesses reduced genomes. This may be a result of the loss of gene duplicates through the accumulation of deleterious mutations if the duplicates do not confer a benefit to the organism possessing them. The findings highlight the factors influencing the number of putative paralogs in microbial genomes of marine microbes in the PEIs. Importantly, the results show that some of these factors may be linked, since

these phyla are found at different depths in an environment with depth-dependant physical factors and nutrient concentrations.

Marine microbes play an important role in biogeochemical cycling, but little is known about the role of gene duplication events in these processes. This knowledge deficit is especially true for the understudied PEIs and surrounding waters. Our data suggests that paralogs do not play a large role in nitrogen cycling in PEI marine microbial communities. There was a lack of paralogs responsible for NAG activity, with less than half the stations possessing paralogs responsible for this extracellular enzymatic activity. We concluded that nitrogen is not a limited resource in PEIs marine microbial communities, a finding supported by nitrate, ammonium and nitrite concentrations which were in line with previous studies. Despite the lack of paralogs responsible for nitrogen cycling, the results showed several paralogs with the potential for carbon and phosphorous cycling.

Paralogs may play an important role in carbon metabolism in the PEIs marine microbial communities as multiple pathways are present in carbon metabolism and methanogenesis. The presence of paralogs encoding different carbon cycling pathways could be due to the SIO representing a globally important carbon sink, where it is responsible for a large portion of global carbon dioxide sequestration. The number of paralogs assigned to BG is highest in samples that have the highest BG activity, suggesting that paralogs play a key role in BG activity in microbial communities in waters surrounding the PEIs. Few paralogs were assigned to BX, while BX was observed in all samples from the SIO, indicating that the presence of gene paralogs may not play an important role in BX activity in microbial communities in waters surrounding the PEIs. The lack of paralogs responsible for BX while paralogs responsible for BG are observed at

each station might suggest that gene duplicates contribute to the acquisition of carbon primarily through BG. This can be confirmed using the heatmap, where the BG pathway, as well as other carbon metabolism pathways, are partially completed by paralogs.

Paralogs responsible for phosphorus cycling pathways were observed in most of the phyla in this study, with the pathways being most complete by paralogs in Cyanobacteria, Verrucomicrobiota, and Bacteroidota. The presence of paralogs with the genomic potential for phosphorus metabolism in phyla known for phosphorus cycling suggests that gene duplicates play a role in phosphorus cycling in this environment.

The presence of paralogs encoding cell wall synthesis proteins in the SAR324 core genome suggests that gene duplications have played a role in the evolution of the phylum's cell wall. Furthermore, the presence of paralogs encoding ribosomal proteins and transcription factors suggests that gene duplications have also played a role in the adaptation of SAR324 to the SIO environment by enabling it to efficiently produce proteins and respond to changing environmental conditions.

This study sheds light on the extent and trends of gene duplications in marine microbial communities of the PEIs and surrounding waters. The findings suggest that the extent of gene duplications in a genome is linked to the phylum to which the organism belongs, with some phyla showing higher levels of gene duplications than others. Paralogs appear to play an important role in carbon and phosphorus cycling, but not in nitrogen cycling, in these microbial communities. The location of paralogs within the pangenome of SAR324 provides insights into the potential functions of gene duplicates in the adaptation and evolution of a phylum. Overall, this study contributes to our understanding of the role of

gene duplications in shaping the genomic diversity and functions of marine microbial communities.

In summary, this study substantially advances current insights into the ecology and evolution of microbial communities in the SIO, highlighting the importance of microbial communities in biogeochemical cycling and carbon sequestration in this critical ecosystem. Additionally, the identification of key genes and functional pathways involved in biogeochemical cycling in the SIO could have significant implications for understanding global carbon cycling and climate change. Our study provides new evidence that the microbial communities in the SIO play a critical role in the sequestration of atmospheric CO_2 through the biogeochemical cycling of carbon, phosphorous and nitrogen.

Overall, our findings highlight the need for further research to fully understand the diversity and function of microbial communities in the SIO and other understudied marine environments. By gaining a better understanding of the microbial communities in these environments, we can better predict how these communities will respond to environmental changes and how they will influence global biogeochemical cycles. This knowledge will be essential for developing effective strategies to mitigate the impact of climate change on our planet.

Supplementary data

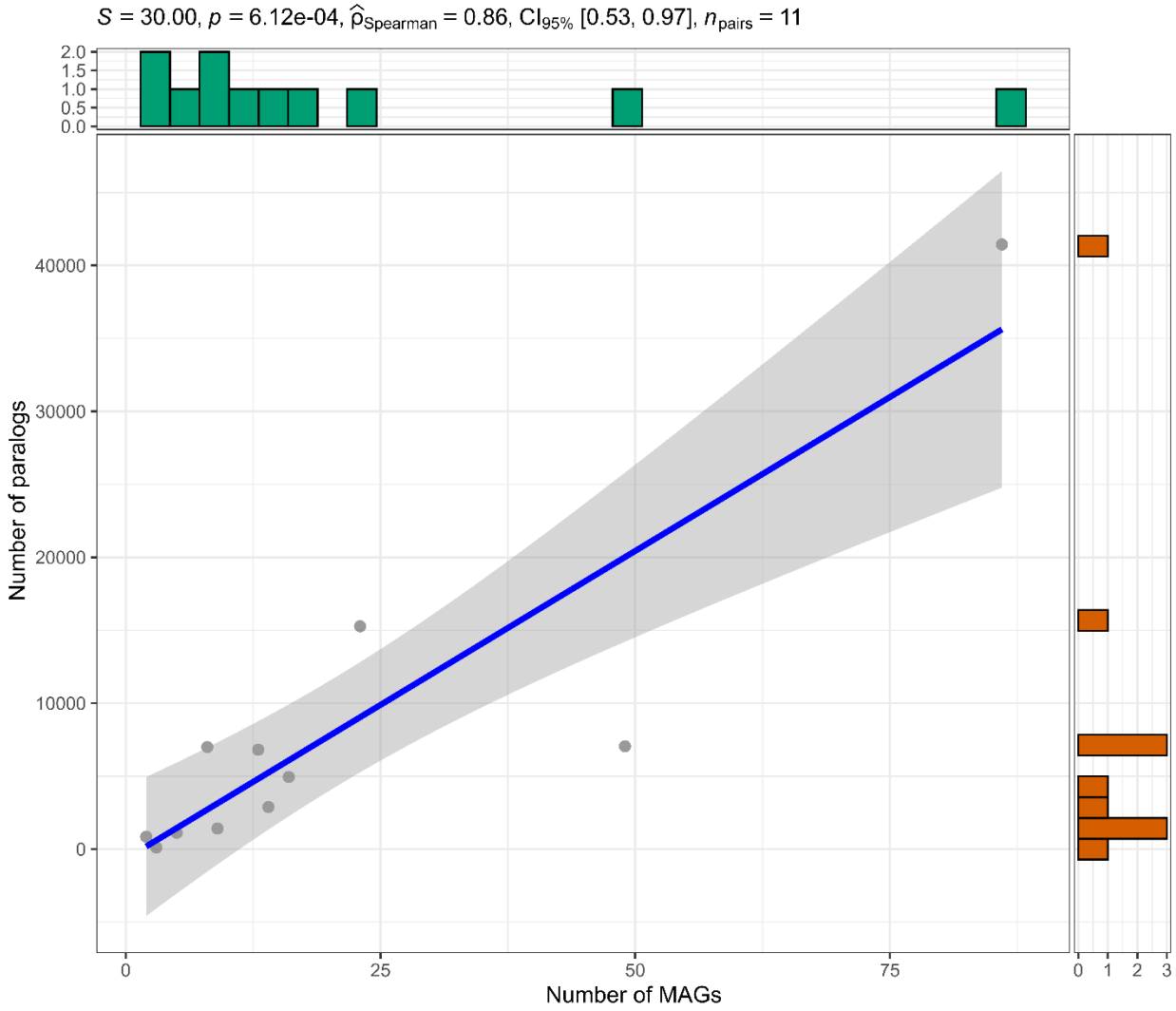

Figure S1: A scatterplot of the number of MAGs belonging to each of the 11 phyla against the total number of paralogs observed in each phyla, including the correlation statistics of the two non-parametric variables. *S* = the observed value of S statistics; *p* = significance; $\hat{\rho}_{Spearman}$ = Spearman's rank correlation coefficient; $CI_{95\%}$ = 95% confidence interval; n_{pairs} = number of observations.

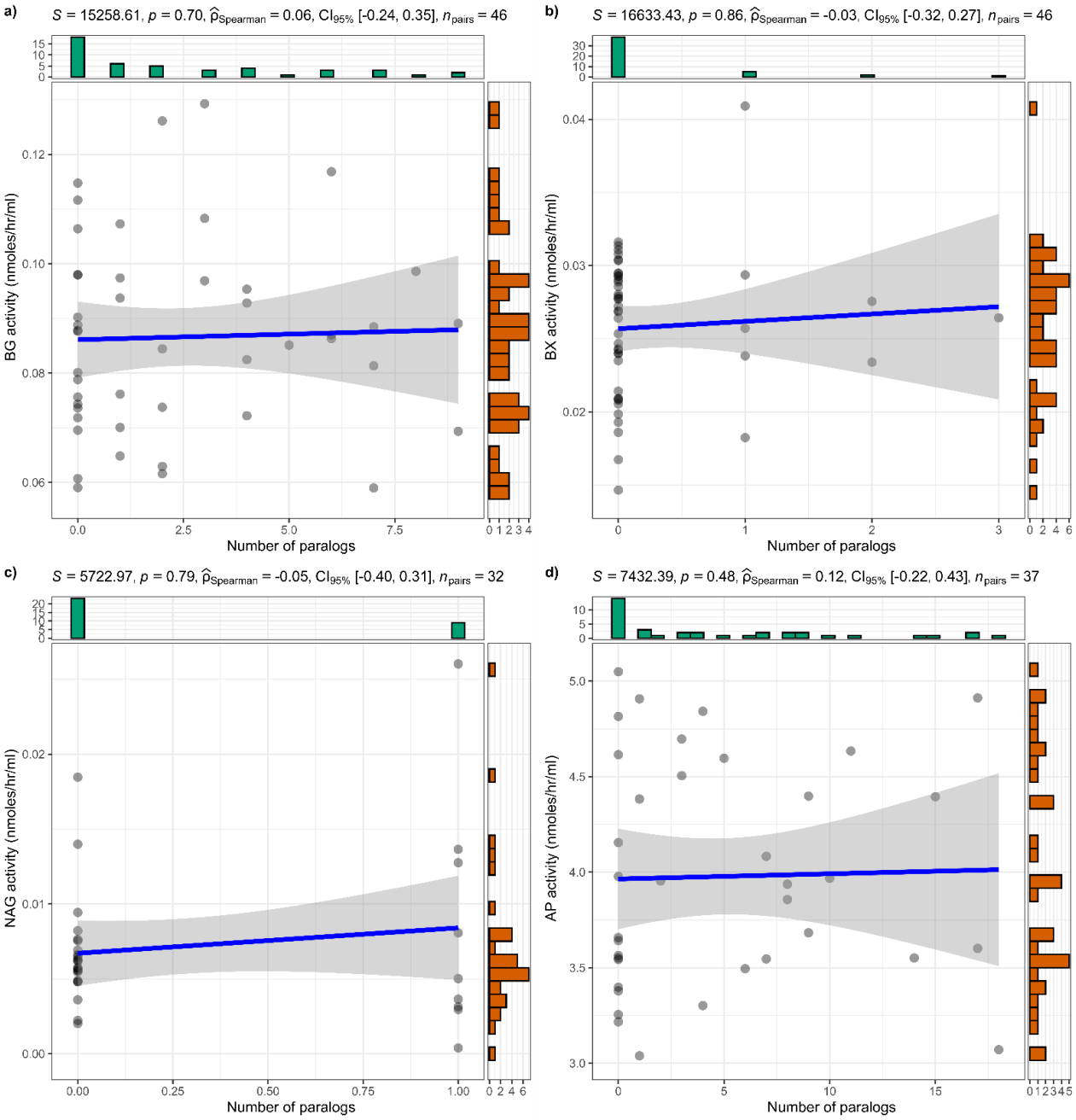

Figure S2: Scatterplots of the number paralogs assigned to a) BG, b) BX, c) NAG and d) AP against their respective activity (nmoles/hr/ml), including the correlation statistics of the two non-parametric variables. *S* = the observed value of S statistics; *p* = significance; $\hat{\rho}_{Spearman}$ = Spearman's rank correlation coefficient; $CI_{95\%}$ = 95% confidence interval; n_{pairs} = number of observations.