# Microbial Ecology and Biogeography

OF THE

# Southern Ocean

David Wilkins

October 7, 2012

# Contents

# List of Figures

# List of Tables

# Acronyms

GAAS  Genome relative Abundance and Average Size.

GLPSOL  GLPK ʟɪɴᴇᴀʀ ᴘʀᴏɢʀᴀᴍᴍɪɴɢ/MIP ꜱᴏʟᴠᴇʀ.

**ANOSIM**  Analysis of SIMilarities.

**AZ**  Antarctic Zone.

**CEAMARC/CASO**  Collaborative East Antarctic Marine Census/Climate of Southern Ocean.

**CTD**  Conductivity, Temperature and Depth.

**GLPK**  GNU Linear Programming Toolkit.

**KEGG**  Kyoto Encyclopedia of Genes and Genomes.

**NZ**  North Zone.

**OTU**  Operational Taxonomic Unit.

**PF**  Polar Front.

**PFZ**  Polar Frontal Zone.

**SIMPER**  SIMilarity PERcentages.

**SZ**  South Zone.

**UFO**  Unidentified Flying Object.

# Acknowledgements

x

# Abstract

# Introduction

This is a test of the acronyms: I saw a Unidentified Flying Object (UFO). It was not the first UFO I'd ever seen. In fact, I've seen 100 UFOs.

Here is some greek: µg.

## Microbial ecology of the Southern Ocean

## Oceanography of the Southern Ocean

### Water masses and fronts

### Effect of climate change

## Role of the Polar Front in biogeography

## Project questions and hypotheses

# The Polar Front as a major biogeographic boundary in the Southern Ocean

Sections of this chapter have been previously published in Wilkins D., Lauro F. M., Williams T. J., DeMaere M. Z., Brown M. V., Hoffman J. M., Andrews-Pfannkoch C., Mcquaid J. B., Riddle M. J., Rintoul S. R., and Cavicchioli R. (2012). Biogeographic partitioning of Southern Ocean picoplankton revealed by metagenomics. *Molecular Ecology*.

## Summary

## Introduction

## Methods

### Sampling and metagenomic sequencing

Sampling[1] was conducted on board the RSV *Aurora Australis* during cruise V3 Collaborative East Antarctic Marine Census/Climate of Southern Ocean (CEAMARC/CASO) from 13 December 2007 – 26 January 2008. This cruise occupied the SR3 latitudinal transect from Hobart, Australia (44° S) to the Mertz Glacier, Antarctica (67° S) within a longitudinal range of 140–150° E. Nineteen samples (16 surface, 3 deep) were obtained along almost the entire latitudinal range (Figure 1).

A range of data were recorded by integrated instruments on the RSV *Aurora Australis* including location, water column depth, water temperature, salinity, fluorescence and meterological data (Table 1).

---

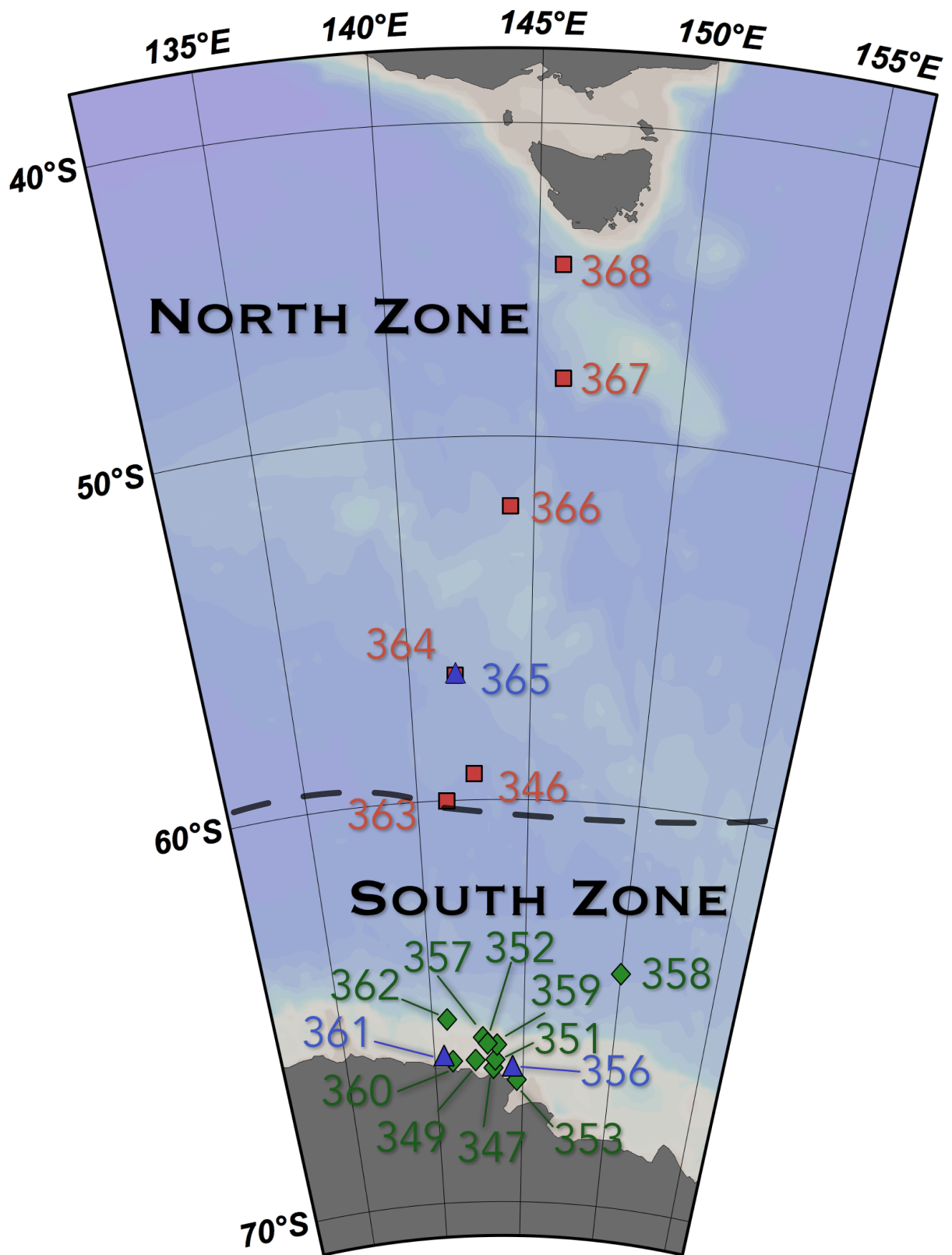[1]Sampling was performed by Jeffrey M. Hoffman and Jeffrey B. Mcquaid

Figure 1: Sites of seawater samples used in this study. Red squares indicate surface samples from the North Zone; green diamonds samples from the South Zone; and blue triangles indicate deep samples. The dashed line gives the approximate location of the Polar Front.

These data were used to locate the Polar Frontal Zone (PFZ) based on a surface temperature gradient of $\sim 1.35$ °C across a distance of 45–65 km, placing the Polar Front (PF) at approximately $-59.70°$ of latitude, consistant with previous descriptions (Moore *et al.*, 1999; Sokolov and Rintoul, 2002). Samples were accordingly grouped into "North" and "South" zones, while the three deep samples composed a "Deep" zone (Table 1). The North Zone (NZ) represents waters from the Subtropical, Subantarctic and PFZ regions, while the South Zone (SZ) represents the Antarctic Zone (AZ).

At each station, $\sim 250$–560 L of seawater was pumped from $\sim 1.5$–2.5 m below the sea surface into drums stored at ambient temperature on deck. In the case of deep samples, $\sim 225$–230 L of seawater was collected from Niskin bottles attached to a CTD (SeaBird, Bellevue, USA). Seawater samples were prefiltered through a 20 µmplankton net, then filtrate was captured on sequential 3.0 µm0.8 µmand 0.1 µm293 mm polyethersulfone membrane filters (Port Washington, USA), and immediately stored at $-20$ °C (Rusch *et al.*, 2007; Ng *et al.*, 2010).

DNA extraction[2] was performed at the J. Craig Venter Institute (Rockville, USA) as described in Rusch *et al.* (2007). Pyrosequencing was performed on a GS20 FLX Titanium instrument (Roche, Branford, USA) also at the J. Craig Venter Institute as described in Lauro *et al.* (2011). Duplicate reads and reads with many pyrosequencing errors were removed as described in Lauro *et al.* (2011).

## Phylogenetic analysis of metagenomic data

### BLAST comparison to RefSeq database

A subset of the RefSeq microbial (bacterial and archaeal) genome database (release 41, retrieved May 31 2012 from `ftp://ftp.ncbi.nih.gov/refseq/release/`) was prepared by excluding sequences with the words "shotgun", "contig", "partial", "end" or "part" in their headers (Angly *et al.*, 2009). Because this database was not expected to contain representative genomes for every species present, OTUs in this study are defined by the best species match to this database, and may for example represent congeners.

The metagenomic reads from each sample were compared against this database using TBLASTX, with default parameters except for: E-value threshold $1.0 \times 10^{-3}$, cost to open gap 11, cost to extend gap 1, masking of query sequence by SEG masking with lookup table only.

### Identification of minimal species sets with MINSPEC

A computational method to minimise false OTU identifications and increase the accuracy of OTU abundance estimates (MINSPEC) was developed and implemented in PERL. Following the approach of

---

[2]DNA extraction was performed by Cynthia Andrews-Pfannkoch and others at the J. Craig Venter Institute

Table 1: Sampling time, location and physiochemical properties of samples used in this study. All data were retrieved from underway instruments aboard the RSV *Aurora Australis*, with the exception of temperature, salinity and fluorescence data for the three deep samples, which was obtained from the CTD (SeaBird, Bellevue, USA) instrument used to collect the samples.

| Sample | Zone | Date | Latitude | Longitude | Water Column Depth (m) | Sample Depth (m) | Temperature (°C) | Salinity (PSU) | Fluorescence ($\mu gL^{-1}$) | Volume filtered (L) |
|---|---|---|---|---|---|---|---|---|---|---|
| 346 | North | 20/12/07 | −59.3120 | 142.5949 | 4294 | 2 | 2.9 | 33.75 | 0.3 | 500 |
| 347 | South | 23/12/07 | −66.0213 | 142.7380 | 450 | 2 | 0.6 | 34.20 | 4.0 | 250 |
| 349 | South | 27/12/07 | −66.5662 | 142.3169 | 370 | 1.5 | −1.3 | 34.40 | 2.3 | 250 |
| 351 | South | 28/12/07 | −66.5587 | 143.4303 | 823 | 1.5 | −0.6 | 34.30 | 1.3 | 500 |
| 352 | South | 29/12/07 | −66.7650 | 143.3240 | 164 | 2.5 | −0.8 | 34.30 | 3.1 | 500 |
| 353 | South | 30/12/07 | −67.0521 | 144.6786 | 180 | 2 | −1.8 | 34.40 | 0.3 | 500 |
| 356 | Deep | 03/01/08 | −66.7617 | 144.4138 | 920 | 920 | −1.9 | 34.69 | 0.1 | 230 |
| 357 | South | 05/01/08 | −66.1719 | 144.0193 | 580 | 2 | −0.4 | 34.15 | 2.5 | 500 |
| 358 | South | 09/01/08 | −64.3001 | 150.0306 | 3550 | 2 | 0 | 33.55 | 0.5 | 500 |
| 359 | South | 12/01/08 | −66.1903 | 143.5292 | 540 | 2 | −0.2 | 34.21 | 2.5 | 500 |
| 360 | South | 13/01/08 | −66.5817 | 141.0211 | 316 | 2 | −0.7 | 34.04 | 6.2 | 500 |
| 361 | Deep | 14/01/08 | −66.4727 | 140.5572 | 1203 | 1170 | −1.8 | 34.56 | 0.1 | 225 |
| 362 | South | 19/01/08 | −65.5367 | 140.8287 | 1064 | 2 | 0.7 | 32.20 | 0.5 | 500 |
| 363 | North | 22/01/08 | −60.0001 | 141.3094 | 4473 | 2 | 3.3 | 33.77 | 0.1 | 500 |
| 364 | North | 23/01/08 | −56.6953 | 141.8780 | 3693 | 2 | 4 | 33.70 | 0.1 | 500 |
| 365 | Deep | 23/01/08 | −56.6967 | 141.9125 | 3693 | 3693 | 0.5 | 34.69 | 0.5 | 230 |
| 366 | North | 24/01/08 | −52.0233 | 144.1362 | 3180 | 2 | 7.6 | 33.84 | 0.3 | 500 |
| 367 | North | 25/01/08 | −48.2487 | 145.9025 | 3490 | 2 | 11 | 34.43 | 0.2 | 500 |
| 368 | North | 26/01/08 | −44.7180 | 145.7775 | 3201 | 2 | 14.8 | 34.96 | 1.3 | 560 |

Ye and Doak (2009) to the parsimonious reconstruction of biochemical pathways (MinPath), minspec computes the smallest set of OTUs sufficient to explain a set of observed high-quality hits against RefSeq (or any other sequence database). The minimal set computation is framed as a linear programming problem and solved with the GNU Linear Programming Toolkit (GLPK) tool GLPK linear programming/MIP solver (GLPSOL) (Free Software Foundation, Boston). This approach eliminates many of the spurious OTU identifications which result from reads with strong identity to more than one OTU. The "minimal species set" is liable to exclude some low-abundance OTUs, but gives more faithful abundance estimates and eliminates many false positives.

To validate this approach and estimate error rates, an assemblage of hypothetical taxa was simulated with varying degrees of overlapping genomic identity and a logarithmic rank-abundance curve. A simulated metagenomic sampling and blast search was performed on this set, and the results processed with minspec. The outputs of all tblastx searches against RefSeq were processed by minspec, and hits not belonging to the minimal sets were removed.

**OTU abundances and variance between zones**

The relative OTU abundances for each sample were determined using the perl script Genome relative Abundance and Average Size (GAAS) (Angly *et al.*, 2009). Briefly, GAAS estimates the relative abundance of OTUs from the number and quality of BLAST hits to each species, taking into account differences in genome size. GAAS was run with the default settings. To normalise for reads which did not yield acceptable hits, the relative abundances for each sample were scaled by that sample's effective blast hit rate. An OTU profile was generated for each sample by encoding the scaled relative abundance of each OTU from each size fraction as a separate variable.

To test the hypothesis that the oceanic zones harbour significantly different communities, Analysis of SIMilarities (ANOSIM) with 999 permutations was performed on a standardised, log-transformed Bray-Curtis resemblance matrix of OTU profiles with PRIMER 6. SIMilarity PERcentages (SIMPER) analysis was performed with PRIMER 6 to identify the contribution of individual OTUs to differences between the zones. All statistical procedures using PRIMER 6 were performed as described by Clarke and Warwick (2001).

**Functional analysis of metagenomic data**

blast **comparison to Kyoto Encyclopedia of Genes and Genomes (KEGG) database**

In order to identify functional differences between the zones, the set of metagenomic reads from each sample was compared against the KEGG GENES database (retrieved July 2 2010 from `ftp:`

`//ftp.genome.jp/pub/kegg/genes/fasta/genes.pep`) with BLASTX, with default parameters except for: maximum number of database sequence alignments 10; E-value threshold $1.0 \times 10^{-3}$; gap opening penalty 11; gap extension penalty 1; masking of query sequence by SEG masking for lookup table only.

**Analysis of functional potential**

Genes identified by BLASTX were aggregated to KEGG ortholog groups according to the KEGG Orthology schema (`ftp://ftp.genome.jp/pub/kegg/genes/ko`, retrieved Mar 29 2011), and ortholog group abundances calculated for each sample. Following Coleman and Chisholm (2010), a read was considered a hit to a given ortholog group if the top three hits for that read (or all hits if fewer than three total hits) were to genes from the same ortholog group, and had bit scores > 40. If the bit score difference between any two top hits was greater than 30, only the hits above this difference were considered.

Ortholog group counts were then used to calculate the abundance of KEGG modules. Because many ortholog groups are members of more than one module, the abundance $a_m$ of each module $m$ was calculated as

$$a_m = \sum_{K=1}^{n} \frac{C_K}{M_K}$$

where $n$ is the number of ortholog groups $K$ belonging to module $m$, $C_K$ is the number of hits to ortholog group $K$, and $M_K$ is the total number of modules to which $K$ belongs. To account for differences in sequencing depth between samples, module abundances were scaled to 500,000 reads per sample. To test the hypothesis that the NZ and SZ harbour significantly different functional potential, one-way ANOSIM with 999 permutations was performed as above on a standardised, log-transformed Bray-Curtis distance resemblance matrix of the module and ortholog group profiles. A functional profile was generated for each sample by summing the scaled abundances of each module from all size fractions, and SIMPER performed as above to identify modules which contributed highly to the variation in functional potential between the two zones. Modules with a high contribution to variance or otherwise of interest were then linked to taxonomy ("taxonomic decomposition") by noting the genus of the organism associated with each gene in the KEGG GENES database and thus calculating the relative contribution of each genus to each module's abundance. This allowed functional contributions to be putatively assigned to genera which were not identified in our taxonomic analysis, as the database included gene sequences for organisms for which a full genome was not available.

# Results

## Metagenomic sequencing

6.6 Gbp of 454 sequence data representing picoplankton in the size range 0.1 – 3.0 µmwas obtained from 16 samples. After removal of low-quality reads, 454 sequencing yielded 157,507 – 597,689 reads per sample (mean 354,399) of lengths ranging from 100 to 606 bp (mean 378).

## Phylogenetic analysis of metagenomic data

TODO WORKING ON THIS SECTION

The proportion of reads in each sample which yielded matches to RefSeq ranged from 25% to 85% (mean 62%).

The most abundant OTUs in each sample are given in Table 2 and a full list of OTU abundances in the supplementary material (`PF-all-OTUs.csv`).

TODO RANK ABUNDANCE (Figure 2)

TODO ADD RANK ABUNDANCE INTO METHODS

ANOSIM analysis showed that the zones harbor significantly different microbial communities (R = 0.451, p < 0.004). SIMPER was performed in order to identify the contribution of individual OTUs to the differences between the NZ and SZ. TODO TOP SIMPER TABLE a full list in the supplementary material (`PF-OTUs-SIMPER.csv`).

The SIMPER analysis found that no single OTU contributed more than 2.9% of variance TODO REF SIMPER TABLE - SUPP? and 74% of variance was contributed by OTUs with a contribution less than 1%. There was also a large difference in the contribution to variance of the three size fractions, with approximately 52% of all variance contributed by OTUs from the 3.0 µmfraction, 37% by the 0.8 µmfraction, and 9% by the 0.1 µmfraction TODO TABLE (Table S3, supporting information). Notably, OTUs within several taxonomic groups that had high contribution to variance covaried in their relative representation in the NZ and SZ. For example, Bacteroidetes and GSO-EOSA-1 representatives were on average more abundant in the SZ; while *Prochlorococcus* and *Synechococcus* spp., SAR11 and SAR116 were on average more abundant in the NZ (**??**). Some groups, such as the Alteromonadales, had variable relative representation depending on size fraction.

TODO nMDS

TODO HERE

Table 2: Relative abundances (as percentages) of the twenty most abundant OTUs identified in this study, in each zone and size fraction.

| | North | | | South | | | Deep | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 μm | 0.8 μm | 3.0 μm | 0.1 μm | 0.8 μm | 3.0 μm | 0.1 μm | 0.8 μm | 3.0 μm |
| *Pelagibacter ubique* HTCC1062 | 61.76 | 25.00 | 23.87 | 58.85 | 22.40 | 17.61 | 37.05 | 24.56 | 17.66 |
| *Nitrosopumilus maritimus* SCM1 | 0.01996 | 0.01438 | 0.009508 | 1.076 | 1.309 | 1.210 | 19.09 | 9.463 | 17.77 |
| *Ruthia magnifica* str. Cm (*Calyptogena magnifica*) | 0.6699 | 0.6458 | 0.5484 | 2.987 | 2.616 | 1.025 | 3.945 | 4.601 | 2.264 |
| *Roseobacter* sp. OCh114 | 0.3125 | 2.932 | 1.588 | 0.4477 | 3.994 | 2.657 | 0.1259 | 1.228 | 0.6792 |
| *Synechococcus* sp. CC9902 | 0.1081 | 9.837 | 4.973 | 0.0007484 | 0.004156 | 0.09733 | 0.002846 | 0.01502 | 0.01058 |
| *Silicibacter pomeroyi* DSS-3 | 0.2578 | 2.286 | 1.154 | 0.3070 | 2.505 | 1.576 | 0.1224 | 0.9417 | 0.4988 |
| *Gramella forsetii* KT0803 | 0.2412 | 1.210 | 1.755 | 0.4993 | 2.347 | 1.890 | 0.2078 | 0.6179 | 0.5173 |
| *Candidatus* Vesicomyosocius okutanii strain HA | 0.4634 | 0.4642 | 0.2078 | 1.970 | 1.807 | 0.2174 | 2.480 | 2.662 | 1.167 |
| *Robiginitalea biformata* strain HTCC2501 | 0.2751 | 1.099 | 1.297 | 0.4722 | 1.878 | 1.405 | 0.2265 | 0.6188 | 0.6946 |
| *Flavobacterium psychrophilum* strain JIP02/86 | 0.1718 | 0.8409 | 1.224 | 0.4316 | 1.960 | 1.598 | 0.1599 | 0.4744 | 0.6001 |
| *Synechococcus* sp. CC9311 | 0.03014 | 4.624 | 4.409 | 0.0007221 | 0.002778 | 0.02764 | 0.001580 | 0.002863 | 0.009241 |
| *Candidatus* Puniceispirillum marinum IMCC1322 | 0.6444 | 2.077 | 1.267 | 0.3586 | 1.377 | 0.7109 | 0.3425 | 1.062 | 0.5345 |
| *Silicibacter* sp. TM1040 | 0.2274 | 1.652 | 0.8738 | 0.2709 | 1.803 | 1.233 | 0.07665 | 0.5890 | 0.2957 |
| *Jannaschia* sp. DFL-12 | 0.1776 | 1.378 | 0.7350 | 0.2443 | 1.692 | 0.8009 | 0.07338 | 0.6515 | 0.3078 |
| *Zunongwangia profunda* strain SM-A87 | 0.1522 | 0.7487 | 1.059 | 0.2968 | 1.410 | 1.204 | 0.1353 | 0.3478 | 0.4971 |
| *Colwellia* sp. 34H | 0.02345 | 0.3636 | 2.736 | 0.05207 | 0.5140 | 1.041 | 0.05137 | 0.4687 | 0.8013 |
| *Coraliomargarita akajimensis* strain DSM 45221 | 0.03698 | 0.07573 | 0.1197 | 0.1154 | 1.543 | 1.680 | 0.02614 | 0.3040 | 0.2740 |
| *Jannaschia* sp. CCS1 | 0.1173 | 0.9344 | 0.4784 | 0.1711 | 1.230 | 0.8239 | 0.05865 | 0.4462 | 0.2118 |
| *Pseudoalteromonas atlantica* strain T6c | 0.01251 | 0.4772 | 1.993 | 0.02270 | 0.4089 | 1.132 | 0.02634 | 0.2143 | 0.7459 |
| *Saccharophagus degradans* strain 2-40 | 0.06532 | 0.4325 | 0.5429 | 0.1289 | 1.072 | 0.8663 | 0.07798 | 0.2844 | 0.3165 |
| *Flavobacterium johnsoniae* strain UW101 | 0.08822 | 0.4220 | 0.6141 | 0.2034 | 0.9389 | 0.8578 | 0.07545 | 0.2255 | 0.3300 |
| *Capnocytophaga ochracea* strain DSM 7271 | 0.1143 | 0.4830 | 0.5399 | 0.2314 | 0.8815 | 0.6814 | 0.08964 | 0.2840 | 0.5043 |
| *Marinomonas* sp. MWYL1 | 0.03777 | 0.2529 | 0.3026 | 0.1514 | 1.300 | 0.7006 | 0.07393 | 0.2439 | 0.2155 |
| *Cellvibrio japonicus* strain Ueda107 | 0.05884 | 0.3080 | 0.3231 | 0.1155 | 0.9917 | 0.4713 | 0.06774 | 0.2981 | 0.2549 |
| *Marinobacter hydrocarbonoclasticus* VT8 | 0.04093 | 0.2889 | 0.3883 | 0.08418 | 0.7195 | 0.3848 | 0.1250 | 0.6667 | 1.066 |
| *Pseudoalteromonas haloplanktis* strain TAC125 | 0.01389 | 0.2505 | 0.8896 | 0.03427 | 0.3561 | 0.6530 | 0.1092 | 1.203 | 0.1503 |
| *Teredinibacter turnerae* strain T7901 | 0.05665 | 0.3051 | 0.3081 | 0.1138 | 0.9174 | 0.5127 | 0.06558 | 0.2649 | 0.1885 |
| *Acinetobacter baumannii* strain SDF | 0.004886 | 0.007187 | 0.4073 | 0.006260 | 0.04218 | 1.459 | 0.004285 | 0.01229 | 0.3155 |
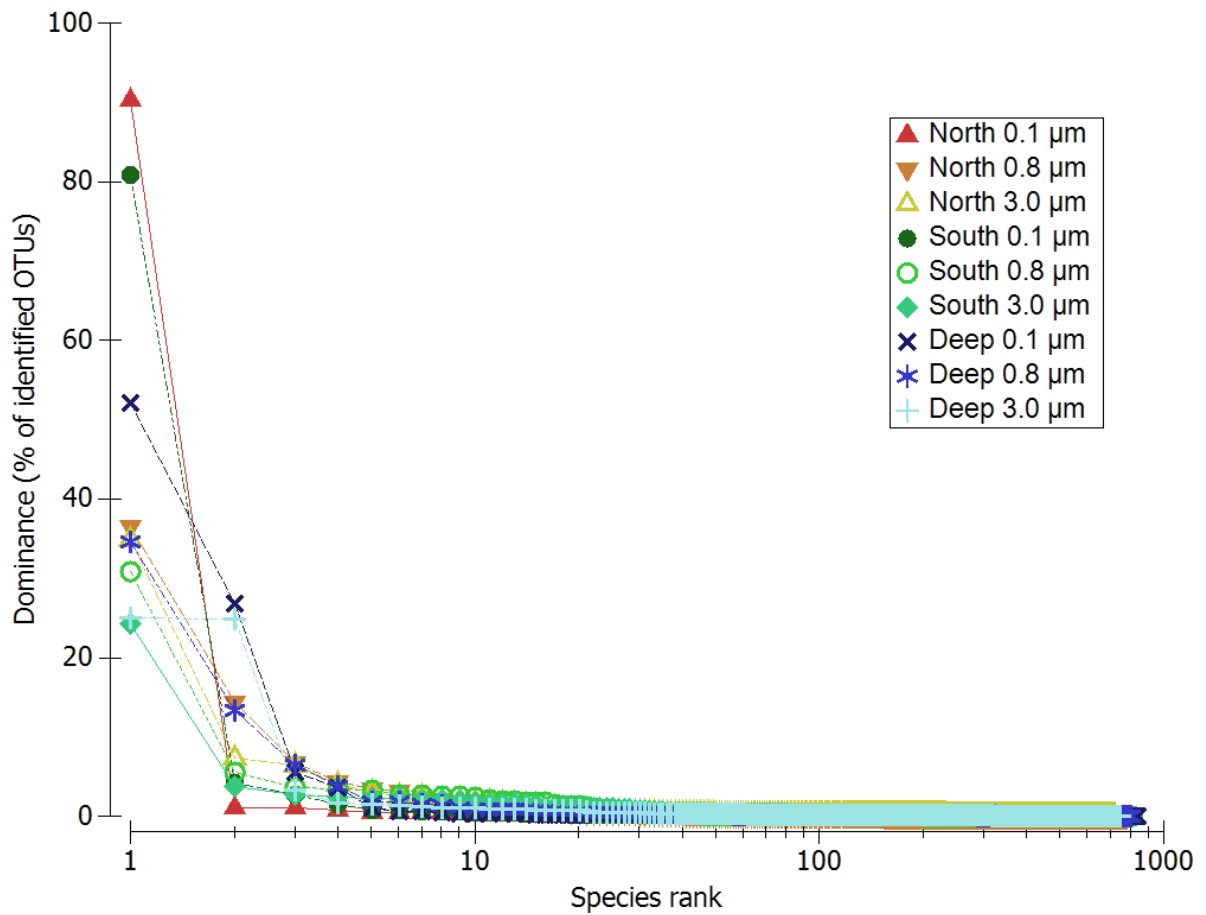
Figure 2: Rank-abundance curves for OTUs identified in each zone and size fraction. The dominance of a given OTU is its relative abundance as a percentage of all identified OTUs. The x-axis is scaled logarithmically.

Table 3: TODO Abundances are zonal averages and have been standardises and log-transformed.

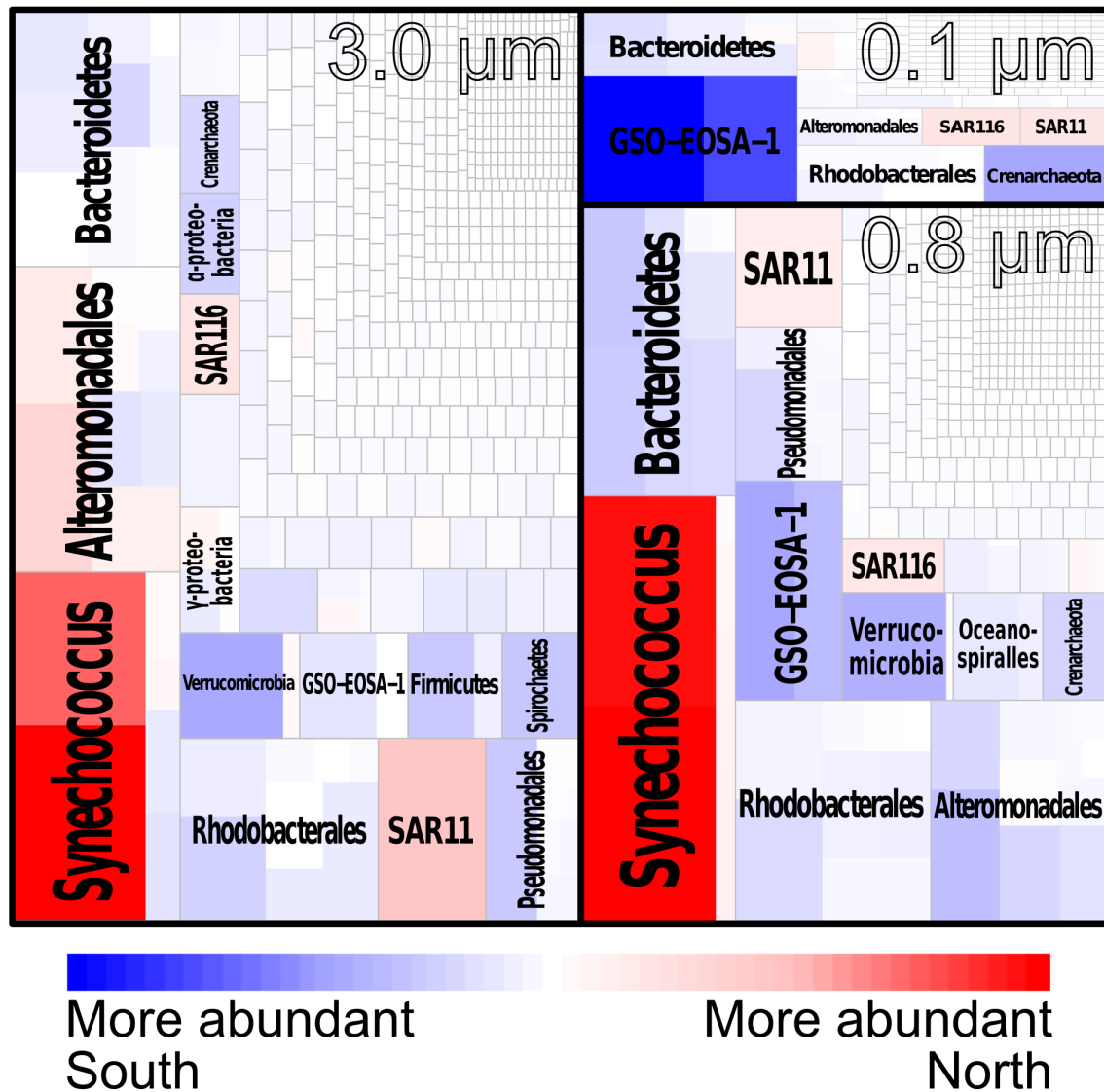| OTU | South | North | Contribution to variance (%) |
|---|---|---|---|
| Synechococcus sp. CC9311-08 | 0 | 1.08 | 2.88 |
| Synechococcus sp. CC9902-08 | 0 | 1.04 | 2.81 |
| Synechococcus sp. CC9311-30 | 0.01 | 0.98 | 2.59 |
| Synechococcus sp. CC9902-30 | 0.04 | 0.76 | 2.03 |
| Pelagibacter ubique HTCC1062-30 | 1.97 | 2.4 | 1.97 |
| Ruthia magnifica str. Cm (Calyptogena magnifica)-01 | 0.82 | 0.25 | 1.57 |
| Colwellia sp. 34H-30 | 0.34 | 0.66 | 1.32 |
| Ruthia magnifica str. Cm (Calyptogena magnifica)-08 | 0.74 | 0.25 | 1.32 |
| Pelagibacter ubique HTCC1062-08 | 2.32 | 2.48 | 1.32 |
| Candidatus Vesicomyosocius okutanii strain HA-01 | 0.62 | 0.18 | 1.2 |
| Coraliomargarita akajimensis strain DSM 45221-08 | 0.48 | 0.04 | 1.13 |
| Coraliomargarita akajimensis strain DSM 45221-30 | 0.49 | 0.06 | 1.1 |
| Roseobacter sp. OCh114-08 | 1.01 | 0.81 | 1.08 |
| Pseudoalteromonas atlantica strain T6c-30 | 0.38 | 0.54 | 1.08 |
| Candidatus Vesicomyosocius okutanii strain HA-08 | 0.57 | 0.19 | 1.04 |
| Acinetobacter baumannii strain SDF-30 | 0.45 | 0.18 | 0.95 |
| Gramella forsetii strain KT0803-08 | 0.72 | 0.43 | 0.94 |
| Marinomonas sp. MWYL1-08 | 0.46 | 0.11 | 0.92 |
| Roseobacter sp. OCh114-30 | 0.76 | 0.54 | 0.91 |
| Flavobacterium psychrophilum strain JIP02/86-08 | 0.63 | 0.32 | 0.89 |
| Silicibacter pomeroyi DSS-3-08 | 0.75 | 0.69 | 0.86 |
| Brachyspira hyodysenteriae strain WA1-30 | 0.47 | 0.19 | 0.84 |
| Ruthia magnifica str. Cm (Calyptogena magnifica)-30 | 0.34 | 0.21 | 0.82 |
| Pseudoalteromonas haloplanktis strain TAC125-30 | 0.22 | 0.33 | 0.77 |
| Robiginitalea biformata strain HTCC2501-08 | 0.61 | 0.4 | 0.74 |
| Seattle Aquarium strain SCM1-01 | 0.27 | 0.01 | 0.72 |
| Gramella forsetii strain KT0803-30 | 0.59 | 0.59 | 0.71 |
| Lysinibacillus sphaericus strain C3-41-30 | 0.29 | 0.02 | 0.71 |
| Seattle Aquarium strain SCM1-08 | 0.25 | 0.01 | 0.7 |
| Silicibacter sp. TM1040-08 | 0.59 | 0.55 | 0.69 |

Figure 3: Contribution of OTUs to variance between the North and South zones, and differential abundance of OTUs from each size fraction between the two zones. Each coloured (red or blue) rectangle represents an OTU identified through analysis of BLAST matches between SO metagenome data and the RefSeq database. The area of each rectangle as a proportion of the total plot area corresponds to that OTU's contribution to the total variance between the two zones. The colour of each rectangle corresponds to difference in relative abundance of that OTU between the zones, with blue indicating a higher relative abundance south of the PF, and red a higher abundance north of the PF. OTUs from clades or taxonomic ranks of interest have been grouped, with labels in bold and groups separated by gray lines. Groups and OTUs with a low contribution to variance which were not grouped are unlabeled. OTUs from each size fraction have also been grouped, with labels in black outline and size fractions separated by thick black lines. The total contribution to variance of each size fraction is given as a percentage.

13

**Validation of** MINSPEC

**Functional analysis of metagenomic data**

# Discussion

# Conclusions

# Meso-scale biogeographic drivers of planktonic diversity

# Conclusions

# References

Angly F. E., Willner D., Prieto-Davó A., Edwards R. A., Schmieder R., Vega-Thurber R., Antonopoulos D. A., Barott K., Cottrell M. T., Desnues C., Dinsdale E. A., Furlan M., Haynes M., Henn M. R., Hu Y., Kirchman D. L., McDole T., McPherson J. D., Meyer F., Miller R. M., Mundt E., Naviaux R. K., Rodriguez-Mueller B., Stevens R., Wegley L., Zhang L., Zhu B., and Rohwer F. (2009). The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. *PLoS Computational Biology*, 5(12):e1000593.

Clarke K. R. and Warwick R. M. *Changes in marine communities: an approach to statistical analysis and interpretation* . PRIMER-E, Plymoth, 2nd edition, 2001.

Coleman M. L. M. and Chisholm S. W. S. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Audio, Transactions of the IRE Professional Group on*, 107 (43):18634–18639.

Lauro F. M., DeMaere M. Z., Yau S., Brown M. V., Ng C., Wilkins D., Raftery M. J., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME journal*, 5(5):879–895.

Moore J. K., Abbott M. R., and Richman J. G. (1999). Location and dynamics of the Antarctic Polar Front from satellite sea surface temperature data. *Journal of Geophysical Research*, 104:3052–3073.

Ng C., DeMaere M. Z., Williams T. J., Lauro F. M., Raftery M., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2010). Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME journal*, 4(8):1002–1019.

Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter J. E., Li K., Kravitz S., Heidelberg J. F., Utterback T., Rogers Y.-H., Falcón L. I., Souza V., Bonilla-Rosso G., Eguiarte L. E., Karl D. M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M. R., Strausberg R. L., Nealson K., Friedman R., Frazier M., and Venter J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77–e77.

Sokolov S. and Rintoul S. R. (2002). Structure of Southern Ocean fronts at 140° E. *Journal of Marine Systems*, 37(1):151–184.

Wilkins D., Lauro F. M., Williams T. J., DeMaere M. Z., Brown M. V., Hoffman J. M., Andrews-Pfannkoch C., Mcquaid J. B., Riddle M. J., Rintoul S. R., and Cavicchioli R. (2012). Biogeographic partitioning of Southern Ocean picoplankton revealed by metagenomics. *Molecular Ecology*.

Ye Y. and Doak T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Computational Biology*, 5(8):e1000465.