Microbial Ecology and Biogeography

OF THE

Southern Ocean

David Wilkins

October 2, 2012

Contents

List of Figures

List of Tables

Acronyms

GAAS Genome relative Abundance and Average Size.

GLPSOL GLPK LINEAR PROGRAMMING/MIP SOLVER.

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

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. Biogeographic partitioning of Southern Ocean picoplankton revealed by metagenomics. *Molecular Ecology*, 2012.

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

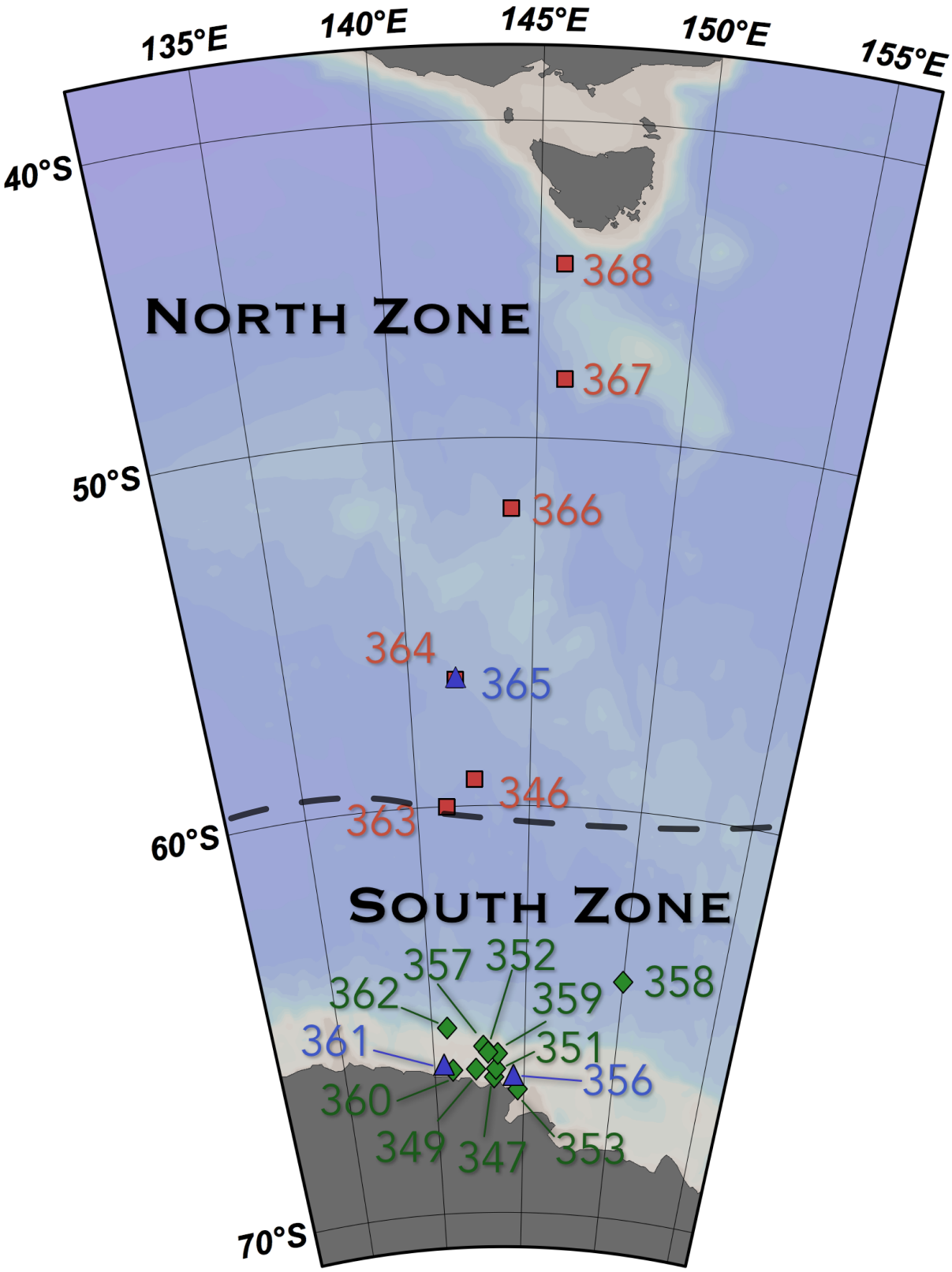

Figure 1: Sites of seawater samples used in this study. Red squares indicate surface samples from the North Zone; green diamonds samples from the South Zone; and blue triangles indicate deep samples. The dashed line gives the approximate location of the Polar Front.

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 µm plankton net, then filtrate was captured on sequential 3.0 µm, 0.8 µm and 0.1 µm 293 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, Operational Taxonomic Units (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 1.0×10^{-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 (μgL^{-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	143.0193	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	0.5	500
365	Deep	23/01/08	−56.6967	141.9125	3693	3693	0.5	34.69	0.1	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 1.0×10^{-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 (âĂIJtaxonomic decompositionâĂİ) 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 moduleâĂŹs 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

TODO HERE

Metagenomic sequencing

Phylogenetic analysis of metagenomic data

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. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. *PLoS Computational Biology*, 5(12):e1000593, 2009.

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

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. An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME journal*, 5(5):879–895, 2011.

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

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. Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME journal*, 4(8):1002–1019, 2010.

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. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77–e77, 2007.

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

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. Biogeographic partitioning of Southern Ocean picoplankton revealed by metagenomics. *Molecular Ecology*, 2012.

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