Global picoplankton biogeography revealed by metagenomic and climatic data integration

Vinícius W. Salazar[1,2], Vanessa R. Marcelino[1], Heroen Verbruggen[3], Kim-Anh Lê Cao[1]

vinicius.salazar@unimelb.edu.au

[1] Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, [2] Melbourne Bioinformatics, Faculty of Medicine, Dentistry, and Health Sciences, The University of Melbourne, [3] School of Biosciences, The University of Melbourne

Introduction

Background: ocean microbes form the basis of the biogeochemical cycles that sustain all of life on Earth.

Motivation: understanding their distribution and ecology is key to integrating microbial observations into Earth System Models.

Goal: to propose a new model of picoplankton biogeography, integrating multiple blocks of metagenomic data.

Dataset: 1454 metagenomes (NCBI SRA), 15,551 reference genomes (OceanDNA, GTDB), 10 environmental variables (Bio-ORACLE 3).

Methods:

1. **Definition of provinces:** measure *k*-mer level pairwise metagenomic distances and perform hierarchical clustering. Divide into 10 groups.
2. **Projection of province areas:** project province areas across oceans using environmental data layers and a random forest model (Fig. 1-2).
3. **Feature selection:** select most relevant features that characterise each province from the taxonomic (Fig. 3) and functional (Fig. 4) profiles. We used counts tables at different taxonomic ranks and KEGG annotations, and the MixOmics package for data integration.

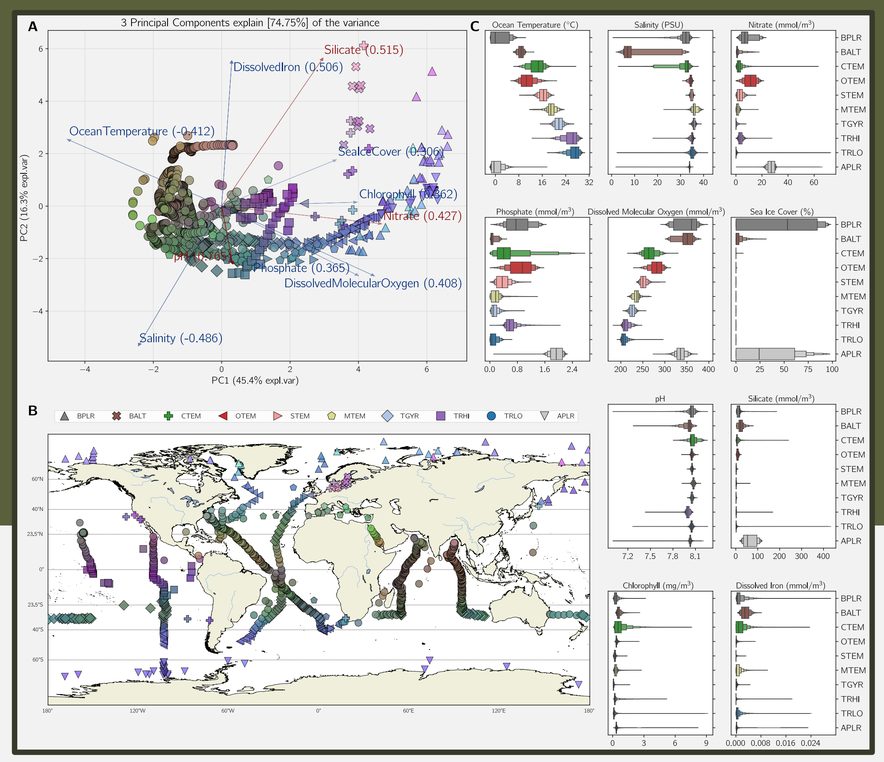

Fig. 2: environmental profiling of provinces. A is a biplot of 3-component PCA ordination, with arrows showing loading vectors of variables. Each component is mapped to a colour channel (blue, red, and green) and sampling stations are coloured according to their position. **B**: world map with sampling stations. **C**: boxenplot of environmental variables associated with each province.

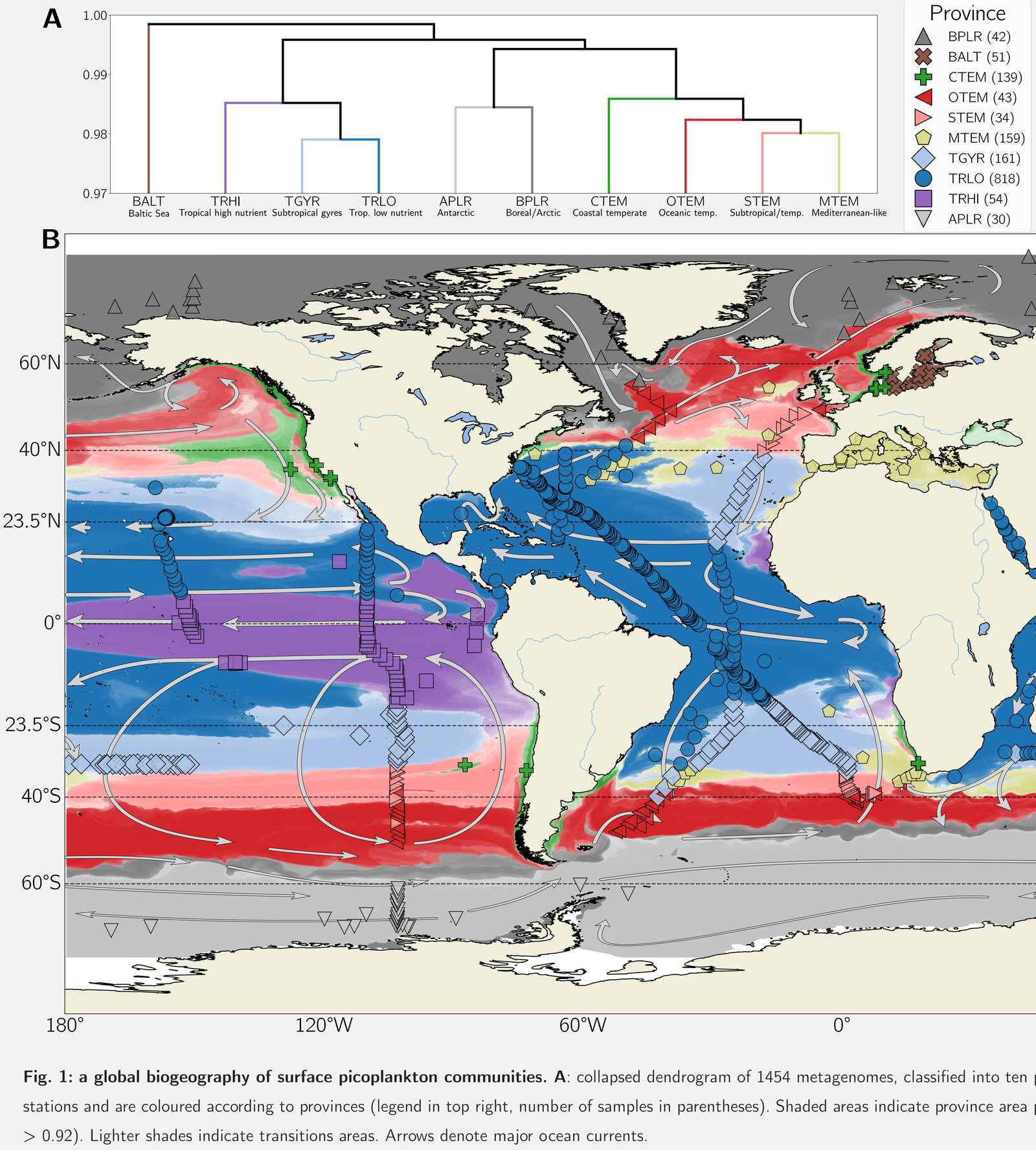

Fig. 1: a global biogeography of surface picoplankton communities. A: collapsed dendrogram of 1454 metagenomes, classified into ten provinces. **B**: world atlas of picoplankton biogeography. Markers indicate sampling stations and are coloured according to provinces (legend in top right, number of samples in parentheses). Shaded areas indicate province area projections based on random forest classifier with 5-fold cross-validation (accuracy > 0.92). Lighter shades indicate transitions areas. Arrows denote major ocean currents.

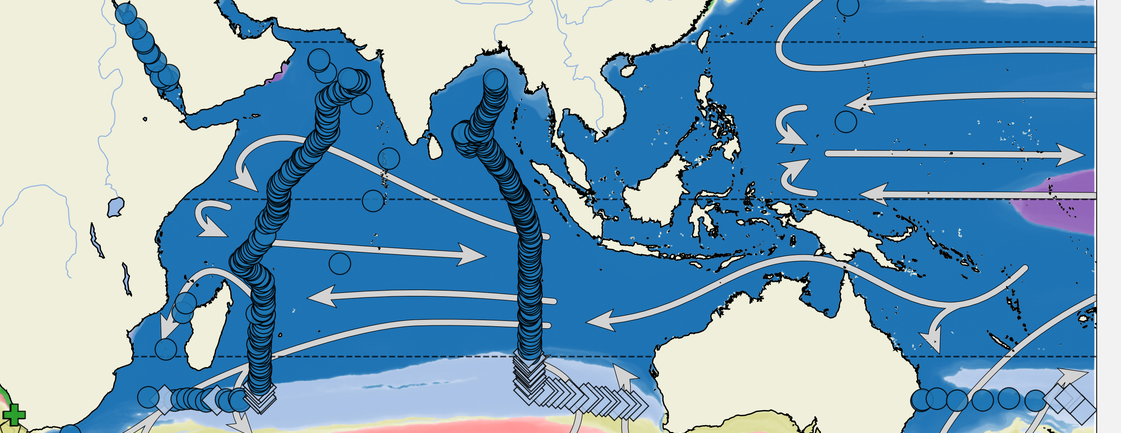

Fig. 3, above: functional profiling reveals metabolic trends across provinces.

Ordination of functional profiles based on KEGG Pathways, generated from an sPLS-DA model with four components. Bars are coloured according to the class which has the maximised value for that given feature.

Fig. 4, left: taxonomic profiles display distinct community structure and dominant taxa.

Relative abundance of the ten most abundant phyla (A) and families (B) across all samples. Panel C: representative communities of provinces, based on most abundant genomes. Sunburst chart shows the relative abundance of taxonomical ranks that each genome belongs to.

Summary

- Characterisation of environmental, taxonomic, functional aspects.
- Community structure can be predicted from environmental data.
- Under described taxa are abundant and play key ecosystem roles.
- AMR and carbon degradation are determinants of biogeography.
- Microbial biogeography as a tool for understanding future global change.

Individual acknowledgements: Saritha Kodikara, Lucie Bittner, Marko Terzin.

Institutional acknowledgements: Melbourne Research Scholarship, UniMelb Research Computing Services, ISME Travel Scholarship.