Characterising and Monitoring Great Barrier Reef Seawater Microbial Communities in a Changing Climate

Submitted by

Marko Terzin, BSc, MSc

ORCID_ID: 0009-0003-0896-1666

A thesis submitted for the degree of Doctor of Philosophy at

James Cook University in 2025

College of Science and Engineering

Dedication

In loving memory of my father, with gratitude for his love, support, and unwavering belief in me.

У сећање на мог оца. Хвала ти за љубав, подршку, и веру у мене.

1.1 Acknowledgements

My PhD journey brought me to Australia, a land of unique beauty and the world's oldest living cultures. I wish to acknowledge and pay my deepest respects to the Traditional Owners, the Bindal, Wulgurukaba, and Manbarra peoples, of the land and sea country where this research was conducted. I honour Elders past, present, and emerging and acknowledge their ongoing struggle to protect their cultural heritage and Country.

First and foremost, I wish to thank my supervisors Patrick Laffy, Nicole Webster, David Bourne, Yun Kit Yeoh, and Steven Robbins for their endless support, encouragement, and for providing the opportunity to work on this project. You all formed an amazing team that I have been privileged to learn from.

Patrick and Nicole, I owe you both more than I can possibly say - you introduced me to the world of coral reefs (a dream for someone from a landlocked country). This entire journey started with our first Skype call in February 2018 (yes, people still used Skype back then) to discuss a Master thesis project; I was incredibly nervous, but the interview went well and I will never forget how Nicole said "I have a good feeling about this." – and here we are. Nicole, your passion for science, your kindness, and your optimism are a true inspiration. I loved working with you during my MSc, and though your role changed during my PhD, I always valued our meetings and missed your daily guidance.

Patrick, you introduced me to bioinformatics. I remember feeling very impressed when you first showed me how to code in Bash (as you basically wrote the code from scratch), and your patience was fundamental in guiding me from there (feeling intimidated looking at the bash terminal) to writing ~10,000 lines of code in this thesis. Thank you for your generosity with your time in walking me through complex problems, and all the debugging sessions I hijacked for impromptu meetings.

To David, your guidance was invaluable in helping me focus on what truly mattered in my PhD, and not get distracted. I will always be amazed by your super-fast and detailed responses to my writing, which were instrumental in keeping my PhD on track from start to finish. Also, over the course of my candidature I came to realise that you are somehow always right (Patrick says that too). I wish I had listened more!

To Yun Kit, thank you for your kind, patient, and calm attitude. Although you were not my supervisor from the beginning, your guidance was so impactful that it feels as if you were here all along. Thank you for your support throughout my candidature and especially for our invaluable writing sessions. My thanks also goes to Steve and others at the Australian Centre for Ecogenomics (ACE) for their collaborative support, in particular Katherine Dougan, Julian Zaugg, and Phil Hugenholtz.

To Kim-Anh, I cannot thank you enough. While you were never formally my supervisor, you contributed so much of your time to this thesis, and your statistical methods were instrumental in

identifying patterns that are simply not observable using the more traditional analysis approaches, which completely changed the way we think about this data. Thank you for welcoming me into your lab, twice! I spent a total of 11 months with you, and I will forever be impressed by the research being done at the Melbourne Integrative Genomics center. Apart from the super hard statistics (that still hurt my brain!), I also learned from you the importance of having a balanced life, dedicating myself to hobbies like climbing and yoga, and remembering to switch off. For all of this, you are a true role model and inspiration to me.

I am deeply grateful to the many people whose monumental efforts in the field and lab made this research possible, especially while I was stuck far away in cold, landlocked Serbia during the pandemic. A huge thank you to Sara Bell, who undertook the vast majority of the fieldwork and all of the complex metagenomics lab work; without her dedication, this thesis truly would not exist. My sincere thanks also to the entire AIMS Water Quality Lab and LTMP team for their skill in generating such high-quality environmental data. A special thanks to Renee Gruber for all our meetings spent deciphering biological patterns; your phrase 'microbes ARE the nutrients' was a game-changer for me! I also wish to thank Mike Emslie and Daniela Ceccarelli for their vital support in interpreting biological patterns related to benthic cover and fish abundance data.

My deepest gratitude goes to the AIMS@JCU program, a prestigious partnership between James Cook University and the Australian Institute of Marine Science, which provides an unparalleled environment for tropical marine research. Thank you for selecting me as a recipient of the AIMS@JCU PhD Scholarship; this thesis would not have been possible without your financial support. I am also profoundly grateful for the additional AIMS@JCU funding provided through the program's science communication and professional development awards, which were instrumental in enabling my participation in key international conferences, specialised training workshops, and transformative research visits. I also wish to express my immense gratitude to the AIMS@JCU coordination team, and in particular Lauren Gregory, Libby Evans-Illidge, and Cherie Motti, for their enormous effort in running the program. I have fond memories of our student seminar days, which were always both fun and insightful, and the restorative writing retreats on Magnetic Island, complete with yoga and mindfulness sessions.

To my incredible friends in Townsville, you have been my family away from home. My journey here started with Dani and Sara, my dear Spanish friends who became my first friends and housemates back in 2019—gracias por todos los recuerdos que hemos creado juntos en estos años. Dani, nunca voy a olvidar cuando fuimos a buscar muebles en la calle (¡qué risa!), y Sara, cuando fui contigo al "turtle rodeo"—qué experiencia tan maravillosa. Os considero amigos para toda la vida. Thank you to Emma, Gabi, Fer, and Cris; one of my first parties was at your house, where I was immediately welcomed by your Mediterranean and Latino warmth. Emma, I'm so glad we also worked in the same office and for all our chats when our PhDs were rough, it felt so sad when you finished and moved on! To Lau, you became one of my closest friends in Townsville; I loved living with you, our deep (and stupid) conversations, and those moments smoking a cigarette in the garden. Un jour, je promets que j'apprendrai le français! A huge thank

you to my climbing buddies Michael, Chloe, Josh, and the entire Urban Climb community. Those bouldering sessions were my perfect escape, and I loved spending hours in the gym. And to all my other friends, Emily, Jared, Lú, Iván, Kevin, Ketki, Ky, Matt, Danny, and Rish (my food guru)—thank you for all the unforgettable parties, weekend escapes to Maggie to switch off, camping and hiking trips, and for making Townsville feel like a true tropical home. I have been so fortunate to share this journey with you all.

Chris, you effortlessly made me feel more carefree during the most stressful times, and you are a huge reason why I fell in love with Melbourne and decided to make it my long-term home. Thank you for every adventure: our van trip through Far North Queensland, seeing the ancient Daintree, cassowaries at Mission Beach, and crocodiles in the wild; exploring remote coral reefs and volcanic islands in Fiji and Thailand; swimming in the crystal clear waters of Mljet; and finding peace at the river house in Kendjija. I look forward to us discovering more.

И за крај, желим да се захвалим породици на подршци: мами, тати, Реи, Мици, и Биси. Знам да ова даљина некад изгледа као да је превише, али уз вашу подршку и љубав сам успео да приведем овај рад крају.

1.2 Statement of the Contribution by Others to this Thesis

The dedicated efforts and support of many colleagues and collaborators were essential to this research. The following section outlines the specifics of their invaluable contributions.

Nature of Assistance	Contribution	Names and Affiliations of Co-contributors
Data Collection & Processing	Fieldwork (seawater collections for metagenomics and water chemistry)	**Sara C. Bell** (AIMS);
	Laboratory work (microbial metagenomics)	**Sara C. Bell** (AIMS);
	Laboratory work (physicochemical data analysis)	AIMS Water Quality and Analytical Lab: **Renee K. Gruber** (AIMS), Ulysse Bove (AIMS), Keeley Glasson (AIMS), Daniel Moran (AIMS);
	Laboratory work (benthic cover and fish abundance data)	AIMS-LTMP team: **Michael J. Emslie** (AIMS), Daniela M. Ceccarelli (AIMS), Emmanuelle Botté (AIMS), Johnston Davidson (AIMS), Veronique Mocellin (AIMS), Josephine Nielsen (AIMS); RV Solander & RV Cape Ferguson crew
	Sequencing support	Microba Life Sciences Ltd. (Gene Tyson, NovaSeq facility); Australian Centre for Ecogenomics Sequencing Facility;
Financial Support	Sequencing costs	Queensland Research Infrastructure Co-investment Fund (RICF) by the Department of Environment and Science, Queensland;
	Research funding (science communication, professional development, research visits, conferences)	**Internal**: AIMS@JCU Quantitative Marine Science Program ($5,000 AUD), Science Communication ($8,250 AUD) and Professional Development ($3,000 AUD) awards **External**: EMBL Australia Travel Grant ($2,000 AUD – LS2N Nantes & EMBL Heidelberg); FEMS Congress Attendance Grant ($1,300 AUD – FEMS MICRO 2025); ACRS Student Grant ($350 AUD – ACRS 2022)
	Presentation awards (used towards science communication)	AIMS@JCU Student Seminar Awards (Total $4,100 AUD: Research Subject Photography, Poster Prize, Seminar Presentations, 3MT)
	Stipend	AIMS@JCU doctoral scholarship
Intellectual	Editorial Assistance (manuscript	**Patrick W. Laffy** (AIMS, JCU), David G. Bourne

Nature of Assistance	Contribution	Names and Affiliations of Co-contributors
Support	review, feedback, editing)	(AIMS, JCU), Yun Kit Yeoh (AIMS, JCU), Steven J. Robbins (UQ), Nicole S. Webster (AIMS, UQ, UTAS), Kim-Anh Lê Cao (UoM), Renee K. Gruber (AIMS), Michael J. Emslie (AIMS), Sara C. Bell (AIMS), Daniela M. Ceccarelli (AIMS), Samuel Chaffron (LS2N, Tara Oceans), Philip Hugenholtz (UQ), Julian Zaugg (UQ), Pedro R. Frade (NHMW), Bettina Glasl (DOME);
		AI tools (ChatGPT and DeepSeek) were used exclusively for proofreading (grammar, syntax, and clarity checks).
	Project coordination	**Patrick W. Laffy** (AIMS, AIMS@JCU), David G. Bourne (AIMS, JCU), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), Steven J. Robbins (UQ), Nicole S. Webster (AIMS, UQ, UTAS), Sara C. Bell (AIMS);
	Read-based metagenomics (guidance)	**Patrick W. Laffy** (AIMS, AIMS@JCU), Steven J. Robbins (UQ), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), David G. Bourne (AIMS, JCU);
	Genome-centric metagenomics (analysis)	Terzin M did not conduct metagenome hybrid assembly, binning, taxonomic annotation, and abundance estimation. This analysis was done by Steven J. Robbins (UQ), Katherine E. Dougan (UQ), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), and Julian Zaugg (UQ), under the guidance of Philip Hugenholtz (UQ), Patrick W. Laffy (AIMS, AIMS@JCU), David G. Bourne (AIMS, JCU).
		Functional annotation of prokaryotic metagenomes (guidance): Patrick W. Laffy (AIMS, AIMS@JCU), Steven J. Robbins (UQ), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), David G. Bourne (AIMS, JCU);
	Statistical data integration (guidance)	**Patrick W. Laffy** (AIMS, AIMS@JCU), **Kim-Anh Lê Cao** (UoM), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), Steven J. Robbins (UQ), David G. Bourne (AIMS, JCU), Pedro R. Frade (NHMW), Murray Logan (AIMS);
	Bioinformatics logistics (software install and set-up, debugging)	The AIMS HPC system: Patrick W. Laffy (AIMS, AIMS@JCU), Geoff Millar (AIMS).
		The ACE HPC system: Steven J. Robbins (ACE), Brian Kemish (ACE), Julian Zaugg (ACE).

1.3 AUTHORSHIP DECLARATION

Chapter No.	Details of publication(s) on which the chapter is based	Nature and extend of the intellectual input of each author, including the candidate	I confirm the candidate's contribution to this paper and consent to the inclusion of the paper in this thesis
1	published as: **Terzin, M.**, Laffy, P.W., Robbins, S., Yeoh, Y.K., Frade, P.R., Glasl, B., Webster, N.S., Bourne, D.G., 2024. The road forward to incorporate seawater microbes in predictive reef monitoring. *Environ. Microbiome* 19, 5. https://doi.org/10.1186/s40793-023-00543-4	**Terzin M** wrote the first draft of the manuscript which was revised with the editorial input from all other co-authors. **Terzin M** and Glasl B developed the figures.	Name: Patrick W. Laffy Name: David G. Bourne Name: Yun Kit Yeoh Name: Steven J. Robbins Name Nicole S. Webster
2	published as: **Terzin, M.**, Robbins, S.J., Bell, S.C., Lê Cao, K.-A., Gruber, R.K., Frade, P.R., Webster, N.S., Yeoh, Y.K., Bourne, D.G., Laffy, P.W., 2025. Gene content of seawater microbes is a strong predictor of water chemistry across the Great Barrier Reef. *Microbiome* 13, 11. https://doi.org/10.1186/s40168-024-01972-0	Nicole S. Webster obtained funding for the project. Nicole S. Webster, David G. Bourne, Patrick W. Laffy, Steven J. Robbins, and Renee K. Gruber conceived the sampling design. Samples were collected in the field and processed in the laboratory by Sara C. Bell (microbial metagenomics), AIMS Water Quality team (physicochemical variables). **Terzin M** analysed and prepared the manuscript (including figures and tables), with assistance of all co-authors. All authors revised the final manuscript.	Name: Patrick W. Laffy Name: David G. Bourne Name: Yun Kit Yeoh Name: Steven J. Robbins Name Nicole S. Webster

3	submitted to Nature Microbiology as: **Terzin, M**., Robbins, S.J., Lê Cao, K.-A., Bell, S.C., Dougan, K.E., Zaugg, J., Gruber, R.K., Emslie, M.J., Ceccarelli, D.M., Chaffron, S., Hugenholtz, P., Webster, N.S., Bourne, D.G., Yeoh, Y.K., Laffy, P.W. 2025. No-take marine reserves promote oligotrophic reef bacterioplankton communities across the Great Barrier Reef.	Nicole S. Webster obtained funding for the project. Nicole S. Webster, David G. Bourne, Patrick W. Laffy, Steven J. Robbins, and Renee K. Gruber conceived the sampling design. Samples were collected in the field and processed in the laboratory by Sara C. Bell (microbial metagenomics), AIMS Water Quality team (physicochemical variables), and AIMS LTMP team (benthic cover and fish abundance variables). Metagenome hybrid assembly, binning, taxonomic annotation, and abundance estimation was conducted by Steven J. Robbins (UQ), Katherine E. Dougan (UQ), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), and Julian Zaugg (UQ), under the guidance of Philip Hugenholtz (UQ), Patrick W. Laffy (AIMS, AIMS@JCU), David G. Bourne (AIMS, JCU). **Terzin M** performed the functional annotation of metagenomes and carried out all subsequent analyses, including dataset integration, and generation of figures and tables, with assistance of all co-authors. All authors revised the final manuscript.	Name: Patrick W. Laffy Name: David G. Bourne Name: Yun Kit Yeoh Name: Steven J. Robbins Name Nicole S. Webster
4	Manuscript in preparation.	Nicole S. Webster obtained funding for the project. Nicole S. Webster, David G. Bourne, Patrick W. Laffy, Steven J. Robbins, and Renee K. Gruber conceived the sampling design. Samples were collected in the field and processed in the laboratory by Sara C. Bell (microbial metagenomics). Viral contig assembly, taxonomic annotation, and abundance estimation was conducted by Katherine E. Dougan (UQ), with guidance	Name: Patrick W. Laffy Name: David G. Bourne Name: Yun Kit Yeoh Name: Steven J. Robbins

	from Steven J. Robbins (UQ), Yun Kit Yeoh (AIMS, JCU, AIMS@JCU), Julian Zaugg (UQ), Philip Hugenholtz (UQ), Patrick W. Laffy (AIMS, AIMS@JCU), and David G. Bourne (AIMS, JCU). **Terzin M** carried out all subsequent analyses, including omics data integration, and generation of figures and tables, with assistance of all co-authors. All authors revised the final manuscript.	Name Nicole S. Webster

I have made every reasonable effort to secure permission and accurately acknowledge all copyright owners. Please contact me if any omission or incorrect attribution is identified.

1.4 OTHER PUBLICATIONS DURING CANDIDATURE

Terzin, M., Paletta, M.G., Matterson, K., Coppari, M., Bavestrello, G., Abbiati, M., Bo, M., Costantini, F., 2021. Population genomic structure of the black coral *Antipathella subpinnata* in Mediterranean Vulnerable Marine Ecosystems. *Coral Reefs* 40, 751–766. https://doi.org/10.1007/s00338-021-02078-x

Poliseno, A., **Terzin, M.**, Costantini, F., Trainito, E., Mačić, V., Boavida, J., Perez, T., Abbiati, M., Cerrano, C., Reimer, J.D., 2022. Genome-wide SNPs data provides new insights into the population structure of the Atlantic-Mediterranean gold coral *Savalia savaglia* (Zoantharia: Parazoanthidae). *Ecol. Genet. Genomics* 25, 100135. https://doi.org/10.1016/j.egg.2022.100135

Terzin, M., Villamor, A., Marincich, L., Matterson, K., Paletta, M.G., Bertuccio, V., Bavestrello, G., Benedetti Cecchi, L., Boscari, E., Cerrano, C., Chimienti, G., Congiu, L., Fraschetti, S., Mastrototaro, F., Ponti, M., Sandulli, R., Turicchia, E., Zane, L., Abbiati, M., Costantini, F., 2024. 2bRAD reveals fine-scale genetic structuring among populations within the Mediterranean zoanthid *Parazoanthus axinellae* (Schmidt, 1862). *Coral Reefs* 43, 357–370. https://doi.org/10.1007/s00338-023-02456-7

Marangon, E., Rädecker, N., Li, J.Y.Q., **Terzin, M.**, Buerger, P., Webster, N.S., Bourne, D.G., Laffy, P.W., 2025. Destabilization of mutualistic interactions shapes the early heat stress response of the coral holobiont. *Microbiome* 13, 31. https://doi.org/10.1186/s40168-024-02006-5

Robbins, S.J., Dougan, K., **Terzin, M.**, Zaugg, J., Bell, S.C., Laffy, P.W., Engelberts, J.P., Webster, N.S., Hugenholtz, P., Bourne, D.G., Yeoh, Y.K., 2025. The planktonic microbiome of the Great Barrier Reef. Under review in *Nature*. https://doi.org/10.1101/2025.05.13.653689

1.5 General Abstract

Coral reefs are experiencing alarming declines due to a combination of local disturbances, such as eutrophication and overfishing, and global pressures from climate change. As these ecosystems continue to degrade, the early identification of environmental stressors and declining reef health becomes critical for implementing effective mitigation and management strategies. Marine microorganisms, including bacteria, archaea, and viruses, are fundamental components of coral reef ecosystems playing critical roles in biogeochemical cycling, food web dynamics, and ecosystem stability. Owing to their short generation times and rapid responses to environmental change, microorganisms have been proposed as early-warning indicators of reef condition, with recent studies suggesting that seawater microbes outperform sediment and host-associated microbiomes (such as those of corals, sponges, and macroalgae) in predicting fluctuations in temperature and nutrient levels in the surrounding reef. Specifically, opportunistic microbes in seawater have shown strong associations with indicators of reef degradation, including poor water quality and reduced coral-to-macroalgae abundance ratios. However, despite growing interest in microbial indicators, their integration into routine reef monitoring remains limited, and their potential as predictive tools for assessing reef health and future trajectories (such as anticipating microbial-driven hypoxic events caused by elevated heterotrophic activity under nutrient enrichment and warming) is largely unexplored.

This thesis explores the potential of seawater microorganisms as predictive (rather than purely descriptive) tools for monitoring coral reef health. To address this objective, this research combines an extensive synthesis of the existing literature (Chapter 1) with findings from large-scale field surveys integrating metagenomic sequencing and environmental data streams (including water chemistry and *in situ* surveys) using statistical omics-integration and machine learning approaches applied across the Great Barrier Reef (GBR) (Chapter 2-4). Finally, Chapter 5 provides a general discussion, summarising the key findings and offering future perspectives for translating microbial observations into operational monitoring tools.

The thesis begins with a critical review of the literature evaluating the current state of microbial-based reef monitoring and the barriers preventing the implementation of microbial biomarkers within reef management frameworks (Chapter 1). While many field-based studies over the last two decades successfully identified microbial taxa linked to various metrics of degraded reef health, progress has been hindered by a lack of unified frameworks, limited experimental validation, and a strong historical focus on taxonomic structure rather than microbial function. In response, Chapter 1 outlines a five-step conceptual framework to catalyse a transition from microbial indicator discovery toward predictive monitoring and applied management, highlighting the need for microbial functional characterisation, index development, and most importantly, scalable (i.e., rapid and cost-effective) microbial-based assays suitable for near real-time deployment.

To address key data limitations in the GBR, where previous microbial indicator analyses have focused primarily on taxonomic structure via 16S rRNA gene amplicon sequencing (notable exception: Glasl et al. 2020[1]), this thesis directly contributed to the generation of metagenomic datasets through Australia's Integrated Marine Observing System (IMOS), forming a core component of the Great Barrier Reef Microbial Genomics Database (GBR-MGD). Unlike amplicon-based approaches, metagenomics enables the simultaneous characterisation of microbial taxonomic composition and functional potential, which is essential to advance our understanding about how environmental changes influence seawater microbial functional processes (such as photosynthesis, sulfate reduction, nitrification, virulence, methanogenesis, and ammonia oxidation) and their feedbacks on reef ecosystem dynamics when integrated with ongoing reef monitoring programs. Specifically, seawater microbial metagenomes collected across 48 offshore GBR-MGD reef sites were integrated with a total of 54 environmental metrics generated by the Australian Institute of Marine Science (AIMS) Long Term Monitoring Program (LTMP) including physico-chemical variables, benthic cover assessments (e.g., coral types, marine algae, abiotic components), and fish counts and biomass categorised by feeding guilds, enabling the largest integrative assessment to date of relationships between microbial communities, reef condition, and environmental context across the GBR.

Using this integrated dataset, the thesis demonstrates that the functional gene content of seawater microbial communities is a strong and consistent predictor of water chemistry across the GBR (Chapter 2). By adopting a gene-centric, read-based metagenomic approach applied to a representative subset of the GBR-MGD Illumina short-read shotgun sequencing data (~80% cheaper than the full dataset), this work shows that microbial functional profiles are associated with environmental gradients twice as stably compared to taxonomic composition alone. In contrast to genome-centric (MAG-based) metagenomics approaches, which typically require months of computational processing for assembly, binning, and annotation of microbial genomes, read-based metagenomics enables the separation of taxonomic and functional signals directly from raw sequencing data within days to weeks, substantially reducing analytical time and cost while retaining strong predictive power. Together, these findings reflect functional redundancy across microbial taxa that enhances the robustness of functional genes as indicators for reef monitoring, consistent with patterns reported across plant, human, and marine microbiomes.

Building on the read-based functional analyses, Chapter 3 adopts a genome-resolved metagenomic approach to examine metagenome-assembled genomes (MAGs), enabling a more direct linkage between microbial taxonomy and functional potential. Using MAG-based analysis, the thesis investigates the influence of fisheries management on seawater microbial communities across the GBR by comparing reefs open to varying levels of fishing vs protected reefs within No-Take Marine Reserves (NTMRs). While rezoning of the GBR Marine Park in 2004 has shown significant benefits in protected areas for reef macro-organisms, including increased fish biomass, lower coral disease, reduced crown-of-

thorns starfish outbreaks, and faster coral recovery from disturbances, little is known about the impact of reef zoning on microbial communities. Using supervised machine learning, species co-occurrence networks, and microbial niche analysis, we identified 350 species-resolved microbes that predict reef zoning with ~71% accuracy. Microbial communities in NTMRs were enriched in oligotrophic taxa such as *Pelagibacter* and SAR86, while copiotrophic microbes like Flavobacteriaceae and Rhodobacteraceae, characterised by larger genomes, higher GC content, and increased metabolic independency measured via increased pathway completeness, were more abundant in the more degraded reefs open to fishing. Microbial indicators of NTMRs were associated with metrics of healthier reefs including higher crustose coralline algae cover, hard coral cover, and increased herbivorous and detritivorous fish, whereas reefs open to fishing were characterised by elevated turf to coral abundance ratios and dissolved nitrogen, which likely resulted in the enrichment of opportunistic microbes reflecting more degraded reef states. These findings mark the first evidence that reef zoning measures in the GBR affect the surrounding reef bacterioplankton, and could support future monitoring efforts and help build a knowledge base for more targeted microbial-based reef health assessments.

With the predictive power of seawater prokaryotes established in previous chapters, Chapter 4 expands the reef monitoring framework to include the most abundant biological entities in the ocean: seawater viruses, which remain largely unexplored in the GBR. Virioplankton are key players in reef health, regulating microbial communities and nutrient cycling via the viral shunt, yet their potential to erase microbial indicator signals through lysis prior to sampling is a crucial and understudied aspect that must be accounted for to achieve robust monitoring. By integrating viral and microbial communities across GBR management zones, Chapter 4 provides the first assessment of whether single-omics or multi-omics models offer greater accuracy in predicting reef zoning status. Independent single-omics models show that viral communities (dominated by Caudoviricetes) were just as accurate in predicting reef zoning status as microbial communities (~72% vs ~71% accuracy, respectively), likely because viruses closely track the dynamics of their zoning-responsive prokaryotic hosts. However, a counterintuitive finding emerged when we integrated the viral and microbial (pMAG) datasets: the multi-omics model's classification accuracy dropped to 59%, potentially because the integration captured the complex biological reality of virus-host interactions, such as a single virus infecting multiple hosts and potential temporal decoupling between viral lysis events and host abundance at the time of sampling. Despite the lower predictive accuracy, this integration provided invaluable ecological insight, revealing that viral indicators of No-Take Marine Reserves (NTMRs) were associated with oligotrophic NTMR-enriched microbes like Pelagibacteraceae, while viral indicators of fished reefs were linked to copiotrophic fished-reef enriched microbial taxa like Rhodobacteraceae and Flavobacteriaceae. Counterintuitive to expectations that more data will facilitate better model predictions, this chapter demonstrates that the decision to use a single-omics or multi-omics approach depends on the monitoring objective: for pure classification accuracy, viral indicators are equally effective as microbial ones, whereas for a mechanistic

understanding of the underlying virus-host interactions that drive ecosystem change, an integrated approach is essential (though comes at the expense of reduced model predictive performance).

A key innovation of this thesis is the application of integrative omics approaches, specifically P- and NP-integration, sensu Lê Cao and Welham (2021)[2], to distinguish stable seawater microbial biomarkers from transient signals in the dynamic ocean environment. P-integration, which combines data from the same omics layer across different samples, populations, or conditions to identify broader, more generalizable patterns[3], is applied here for the first time in environmental microbiology (Chapters 2 and 3), enabling the identification of stable seawater microbial markers consistently linked to specific environmental metrics across different sectors and seasons in offshore GBR surface waters. Since microbial and viral genomes from the GBR-MGD data originate from the same samples, we were able to integrate these two omics layers using NP-integration[2,4] to explore virus-microbe multi-omics correlations and better understand the effects of reef zoning on multiple components of seawater microbial communities. Given the emerging meta-omics data from recent large-scale reef surveys, such as the Tara Pacific Expedition or multi-omics surveys of Florida's coral reefs, integrative omics approaches like those applied in this thesis could be relevant beyond the GBR.

Lastly, to enable (near) real time reef health assessments in the field, ideally within hours or days, we must move beyond sequencing, advancing the objectives of Phases 4 (development of rapid microbial-based assays) and 5 (assay implementation in reef management) from our framework proposed in Chapter 1. In Chapter 5, we outline key priorities, opportunities, and challenges for integrating seawater microorganisms into existing reef monitoring programs, with a primary focus on the potential of microbial-based rapid assays that go beyond sequencing.

Conclusions

In summary, this thesis proposes a framework for integrating seawater microbial observations into coral reef management, highlighting the potential of reef bacterioplankton as predictive biomarkers for reef health. Continued research is needed to refine these microbial biomarkers, develop scalable monitoring solutions, and advance methods beyond sequencing to create rapid diagnostic tools for proactive, real-time reef management: a key research challenge for the future.

Table of Contents

1.6 List of Figures

Figure 1.4. The proposed five-step framework of research and innovation to move from descriptive to predictive reef microbial monitoring, and from reactive to proactive reef management. Functional meta-omics datasets are critical to discover microbial indicators of poor reef health in the field (Phase 1), however high costs (see 'Assay price') and long bioinformatics processing times (see 'Timeframes') of microbial meta-omics datasets suggest their limited utility for rapid decision-making in reef management. We highlight that this milestone has been largely achieved through various localised studies, though in the years to come, the integration of recently generated datasets obtained in large-scale surveys (most notably the Tara Pacific Expedition) will be crucial to understand the ubiquity of identified microbial indicators at global scales. Once microbial indicators of poor ecosystem health are identified based on functional meta-omics datasets, experimental validation (Phase 2) is needed to confirm the same patterns occur in laboratory conditions, as well as to identify the causality of microbiome-environment associations from the field, which we predict still remains a distant goal and will require years of research. Once experimentally validated, microbial indices can be formulated (Phase 3) and applied research can commence to develop rapid (within weeks, days or minutes, see 'Timeframes') and cost-effective (see 'Assay price') assays to quickly assess reef health in the field (Phase 4), which can be used in proactive reef management and rapid decision-making (Phase 5).

Figure 2.1: Field sampling design for the GBR-MGD (Great Barrier Reef Microbial Genomics Database) dataset by Australia's Integrated Marine Observing System (IMOS). Seawater was collected from 48 offshore GBR reef sites for microbial community metagenomic sequencing and analysis of 17 physico-chemical variables over 4 trips between November 2019 and July 2020. Reef sites are colored in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. Trip-specific maps with individual reef names are provided in Supplementary Material (Appendix A, Figures S1–S4).

Figure 2.2: Summarising water chemistry data and identifying drivers of seawater microbial communities. (A) Principal Components Analysis (PCA) shows the main clustering patterns of reef sites based on physico-chemical variables. Reef sites use specific shapes and are coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. (B) The heatmap shows changes in physico-chemical variables (y axis) across the reef sites (x axis). Physico-chemical variables were centered (median = 0) and scaled (standard deviation (SD) = 1) across reef sites, and values that deviate from the median (0) were shown in red (> median) and blue (< median). (C) A total of 34 partial Mantel tests (corrected for geographic distance) were conducted for each of the 17 physico-chemical variables, and for both microbial datasets on taxonomy and GO terms. Non-significant results (p value > 0.05, Bonferroni correction) are shown as white cells, while coloured cells denote statistically significant trends (p value < 0.05, Bonferroni correction), indicating positive (red) or negative (blue) associations (Spearman's rank correlation coefficients ρ shown as the numeric value) between microbial and environmental distance matrices, while corrected for geographic distance between reefs (expressed in km).

Figure 2.3: Main clustering patterns of seawater microbial communities. Principal Components Analysis (PCA) plots show the main clustering patterns of reef sites based on microbial community structure, both for microbial taxonomy (A) and microbial GO terms (B). Reef sites are coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. Stacked bar plots illustrate microbial relative abundances (y-axis) for each sample (x-axis), with reef sites grouped by their corresponding sampling trip. These plots represent: (C) the top 20 most abundant microbial genera, (D) all 29 identified microbial phyla, and (E) all microbial genera within the phylum Bacteroidetes. The top three most abundant genera (C) and phyla (D) are highlighted in bold and the legend for genera within Bacteroidetes (E) was excluded due to the large number of taxa. (F) Boxplots illustrate microbial diversity (Shannon Index) for genera within phylum Bacteroidetes, across sampling trips. The symbols *, **, ***, and **** denote levels of statistical significance in pairwise Wilcoxon rank sum tests when testing variation of Bacteroidetes Shannon diversity scores across the four sampling trips: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$, indicating increasing levels of significance. 'ns' indicates non-significant results, where $p \geq 0.05$.

Figure 2.4: MINT sPLS - Associations between microbial taxa and physico-chemical variables. (A) The heatmap shows similarity values (partial correlations) between 17 continuous physico-chemical variables (predictor dataset) and 100 microbial taxa (response dataset) selected across the first two MINT sPLS dimensions. Heatmap cells are coloured to indicate either positive (red) or negative (blue) correlation. Heatmap rows and columns were clustered with a complete Euclidean distance method, with three clusters highlighted with a dashed line and numbered as they were discussed in the text. (B) Indicator stability barplots as determined by Leave-One-Group-Out Cross-Validation - LOGOCV. Microbial indicator taxa are colored in green if they are shared across sampling trips, or in grey if they are trip-specific. (C) Taxonomic breakdown of indicator microbes, with indicator taxa shared across different sampling trips (as inferred by LOGOCV) further highlighted in bold. (D) Explanation of LOGOCV stability scores through 15 possible scenarios. Indicator microbes were assigned colours if indicative in a particular trip (with colouring scheme for trips corresponding to Fig. 2.1), while non-indicator taxa are colored in grey (D, left). The lowest LOGOCV stability score of 0.25 indicates trip-specific microbial indicators (selected in 1/4 LOGOCV iterations, with four possible scenarios), which were therefore considered unstable as these indicators are not reproducible across sampling trips (D, middle). Stable microbial indicators (shared across trips) were assigned LOGOCV stability scores of either 0.5 (selected in 2/4 of the LOGOCV iterations, with six possible scenarios), 0.75 (selected in 3/4 of the LOGOCV iterations, with four possible scenarios), or 1, which indicated the highest indicator stability score (selected in each of the four LOGOCV iterations) (D, right). Only shared microbial indicator taxa (with LOGOCV stability scores of 0.5, 0.75, and 1) were considered in downstream interpretation and discussion of results.

Figure 2.5: MINT sPLS - Associations between microbial genes/functions (GO terms) and physico-chemical variables. (A) The heatmap shows similarity values (partial correlations) between 17 continuous physico-chemical variables (predictor dataset) and 100 microbial GO terms (response dataset) selected across the first two MINT sPLS dimensions. Heatmap cells are coloured to indicate either positive (red) or negative (blue) correlation. Heatmap rows and columns were clustered with a complete Euclidean distance method, with three clusters highlighted with a dashed line and numbered as they were discussed in the text. (B) Indicator stability barplots as determined by Leave-One-Group-Out Cross-Validation - LOGOCV. Microbial indicator genes are colored in green if they are shared across sampling trips, or in grey if they are trip-specific. (C) GO functional annotation of indicator genes/functions, with indicator GO terms shared across different sampling trips (as inferred by LOGOCV) further highlighted in bold. (D) Explanation of LOGOCV stability scores through 15 possible scenarios. Indicator genes were assigned colours if indicative in a particular trip (with colouring scheme for trips corresponding to Fig. 2.1), while non-indicator genes are colored in grey (D, left). The lowest LOGOCV stability score of 0.25 indicates trip-specific microbial indicators (selected in 1/4 LOGOCV iterations, with four possible scenarios), which were therefore considered unstable as these indicators are not reproducible across sampling trips (D, middle). Stable microbial indicators (shared across trips) were assigned LOGOCV stability scores of either 0.5 (selected in 2/4 of the LOGOCV iterations, with six possible scenarios), 0.75 (selected in 3/4 of the LOGOCV iterations, with four possible scenarios), or 1, which indicated the highest indicator stability score (selected in each of the four LOGOCV iterations) (D, right). Only shared microbial indicator genes (with LOGOCV stability scores of 0.5, 0.75, and 1) were considered in downstream interpretation and discussion of results.

Figure 2.6: Differing diagnostic potential of microbial taxonomy and function to inform changes in continuous physico-chemical variables in the surrounding reef. (A) Bray-Curtis Similarity Index shows within-reef community similarity (0 = dissimilar; 1 = identical) for microbial taxonomy (at genus, family, and order-level classifications) and microbial functions (GO terms collapsed at Ranks 5, 4, and 3). (B) Comparison of how frequently indicator microbes and indicator genes (left and right boxplots, respectively) are re-selected across 200 independent sPLS cross validation runs (4-fold CV x 50 repeats), across all four sampling trips (Trips 1-4). Higher stability scores are a proxy of robustness of the indicator signal for a corresponding microbe/gene (i.e. the stability score of 1 would mean that the indicator microbe/gene was re-selected in sPLS on component 1 in each of the 200 CV runs). The symbols *, **, ***, and **** denote levels of statistical significance in pairwise Wilcoxon rank sum tests when testing variation between stability scores from indicator taxa and GO terms within each of the four sampling trips: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$, indicating increasing levels of significance. ‘ns’ indicates non-significant results, where $p \geq 0.05$.

Figure 2.7: Conceptual overview summarising the roles of seawater microbiomes in nutrient cycling in offshore Great Barrier Reef (GBR) surface waters. (A) Planktonic picocyanobacteria Synechococcus and Prochlorococcus play key roles in nutrient cycling in the outer GBR: they uptake dissolved inorganic nutrients (DIN) such as nitrogen (ammonium - NH4, nitrite - NO2, and nitrate - NO3) and phosphorus (phosphate - PO4), reducing DIN concentrations (1A). In the presence of light and carbon dioxide (CO2) the uptaken DIN will be used during photosynthesis to produce particulate organic matter (POM) including organic carbon (POC), nitrogen (PN), and phosphorus (PP), overall resulting in elevated POM concentrations and higher biomass of these picocyanobacteria, indicated via elevated chlorophyll a (Chl-a) (2A). During summer, elevated photosynthesis rates primarily by Synechococcus result in up to a 3-fold increase in POM production, whereas during winter, nutrient concentrations are lower and we also observe notable contributions of Prochlorococcus to POM production (2A). A fraction of POM deriving from Synechococcus and Prochlorococcus will (B) enter the microbial loop. Here, a diverse consortium of seawater heterotrophic microbes, notably Rhodobacteraceae, Rhodospirillaceae, Oceanospirillaceae, Burkholderiaceae and Flavobacteriaceae, will benefit from nutrient-rich conditions by encoding genes for (1B) nutrient uptake and (2B) cellular respiration to generate energy, which can be directed towards (2C) synthesis of complex compounds and (2D) microbial growth. As a result of microbial activity on phytoplankton-derived POM, organic molecules originally present in particulate form are remineralised into DIN (NH_4, NO2, NO_3, PO_4), and dissolved organic carbon (DOC). These dissolved nutrients are then available for uptake by other organisms, including Synechococcus and Prochlorococcus which can photosynthesise again, ultimately recycling POM in the outer GBR and making it available to higher trophic levels (C). POM from these picocyanobacteria may enter marine food webs via two pathways: (1C) an indirect pathway, where heterotrophic seawater microbes that successfully integrated phytoplankton-derived POM into their biomass will be grazed by flagellates and microzooplankton, which in turn will support larger macroorganisms; or (2C) through direct uptake of POM that escapes immediate metabolism by heterotrophic seawater microbes, thus bypassing the microbial loop.

Figure 3.1: Field sampling design for the Great Barrier Reef Microbial Genomics Database (GBR-MGD) dataset. Seawater was collected from 48 offshore GBR reefs for microbial community metagenomic sequencing and analysis of physico-chemical variables, concurrently with AIMS-LTMP in situ estimates of benthic cover and fish abundance and biomass. Sampling occurred in four trips between November 2019 and July 2020, with red tones indicating Austral summer/wet season (trips 1-3, Nov 2019–Feb 2020) and blue indicating winter/dry season (trip 4, July 2020). Samples were taken across seven GBR sectors, denoted on the maps with different symbols. Trip-specific map insets show that reefs in No-Take Marine Reserves (NTMRs, green zones) and fished reefs (dark-blue and yellow zones) were sampled in pairs to minimise confounding effects of geography.

Figure 3.2: Visualisation of seawater microbes selected to distinguish NTMRs from fished reefs across seven GBR sectors. (A) Sample plot from MINT sPLS-DA showing clustering of reefs by reef zoning (fished vs NTMRs) based on indicator seawater prokaryotic metagenome-assembled genomes (pMAGs). Samples (48 reefs x four replicates) are projected in the first two components of the MINT sPLS-DA space, with ellipses indicating 95% confidence level. (B) Heatmap shows differential abundances for the pMAGs (columns) discriminating between NTMRs and fished reefs (rows). Euclidean clustering of microbes by abundance across reef sites splits the heatmap into two groups: 236 pMAGs indicating NTMRs (left), and 114 microbial indicators of fished reefs (right). (C) Heatmap indicating whether the pMAGs were more relatively abundant in NTMRs (green) or fished reefs (blue) in each GBR sector. (D) Alluvial diagrams summarise the taxonomy of indicator microbes for NTMRs (left) and fished (right) reefs, ordered by the most common taxa for each zone.

Figure 3.3: MINT sPLS integrates abundances of seawater microbes with continuous environmental data, while accounting for GBR sector-specific variation. (A) The biplot from MINT sPLS shows both samples (four replicates per reef; hollow circles represent site centroids connecting the four site replicates) and variables (environmental only, in black; microbial variables omitted for clarity) on the same plot. (B) A heatmap shows the MINT sPLS partial correlations between 350 pMAGs identified in MINT sPLS-DA as reef zoning indicators (rows; colored based on indicator status) and 25 most influential environmental variables (columns).

Figure 3.4: Genomic, functional, and network characteristics of indicator microbes associated with NTMRs (green) and fished reefs (blue). Genomic features (A–D): Barplots show genome size (A) and GC content (C) for indicator pMAGs, with boxplots comparing their distributions between NTMRs and fished reefs (B, D). Lower values are considered signatures of genome streamlining. Functional potential and interaction structure (E–I): (E) Barplots show average KEGG module completeness per pMAG, with (F) corresponding group-level comparisons. Lower values indicate gene loss and incomplete pathways. (G) Microbe-specific ratios of positive to negative co-occurrence edges represent cooperative versus antagonistic interactions, and (H) shows group-level comparisons of these ratios. (I) Sample-level ratios of positive to negative cohesion illustrate co-occurrence within overall bacterioplankton communities. Higher positive to negative ratios (both for connectedness and cohesion; G-I) indicate a prevalence of positive interactions, while lower values are a proxy of negative interactions. Diversity and community structure (J–K): (J) Network modularity scores for microbial networks between reef zones, with higher and lower modularity values indicating increased connectivity 'within' and 'between' modules, respectively. (K) Alpha diversity (Shannon index) is shown for seawater microbiomes from NTMR and fished reef sites. For boxplots (B, D, F, H, I, K), significance levels from Wilcoxon rank sum tests are indicated as: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; "ns" = not significant.

Figure 3.5: Differential metabolic potential of microbial indicator taxa in fished reefs and No-Take Marine Reserves (NTMRs). Heatmaps show the top 45 KEGG modules with the greatest differences in completeness scores between microbial indicator pMAGs from fished reefs and NTMRs. Rows represent KEGG modules (grouped by metabolic category), and columns represent indicator pMAGs clustered by reef zoning status. Module completeness scores (ranging from 0 to 1) are shown by color scale. Columns are annotated by the taxonomic order of each pMAG; only the four most abundant orders in each group are distinctly colored (green for NTMR-enriched, and blue for fished reef-enriched microbes; grey for others). Statistical differences in KEGG module completeness (mean ± SD) between groups were assessed using Wilcoxon rank-sum tests, with adjusted p-values indicated as: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; "ns" = not significant.

Figure 3.6: Microbial predictors of reef environmental variables, derived from random forest (RF) modeling and microbial niche analysis. (A) Boxplots show random forest model performance across environmental variables, based on R-squared (R^2) values from 50 stratified permutation tests per variable. Cross-validation used an 80/20 train/test data split stratified by the GBR sector. Gray points represent individual permutations, and pink diamonds indicate median R^2. Variables are ordered by decreasing median R^2, with boxplots colored by performance category. The dashed line marks null performance ($R^2 = 0$). (B-C) Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the top 50 microbial predictors—(B) per pMAG and (C) combined across all 50 pMAGs—for three environmental variables: seawater temperature (left), fish biomass (middle), and turf algae cover (right). Niche bound values are visualised using distinct point shapes. In (B), microbial predictors are additionally colored by random forest importance (%IncMSE).

Figure 3.7: Conceptual model of the seawater microbial dynamics within No-Take Marine Reserves (NTMRs) and fished reefs in the Great Barrier Reef, in the context of physico-chemical, fish abundance, and benthic cover variables. (A) In some NTMR reefs, the removal of fishing pressure results in higher fish biomass, including herbivorous fish (1), which enhances grazing pressure on algae and promotes increased hard coral cover (e.g., Acropora spp., Millepora) and crustose coralline algae (CCA) abundance (2). Efficient nutrient cycling in such healthier reefs with elevated fish biomass and coral cover leads to lower nutrient availability (3). This reduction in nutrients imposes strong selective pressures on seawater microbial communities (4), enriching NTMRs with oligotrophic taxa (e.g., Pelagibacterales, SAR86, Marinisomatota) that exhibit streamlined genomes and incomplete biochemical pathways, as predicted by the microbial streamlining theory. These microbes form positively connected communities, likely reflecting metabolically co-dependent relationships where streamlined taxa rely on other bacterioplankton members to exchange missing metabolites, consistent with the Black Queen Hypothesis. (B) In fished reefs, lower fish biomass and reduced grazing pressure (1) result in higher turf-to-coral ratios (2) and higher nutrient concentrations (3), likely due to inefficient nutrient cycling and DOM release from turf algae, as predicted by the DDAM model. These conditions favor opportunistic microbial taxa (4), ultimately leading to distinct microbial community dynamics compared to NTMRs.

Figure 4.1: Multi-omics signatures (seawater microbes and viruses) that distinguish No-Take Marine Reserves (NTMRs) and fished reefs across Great Barrier Reef (GBR) sectors. Consensus sample plot from MINT BLOCK sPLS-DA on seawater pMAGs and virus data, used to discriminate reef zoning status (fished reefs in blue; NTMRs in green) across GBR sectors (shapes), and with ellipses denoting 95% confidence intervals. The asterisk for each sample represents its consensus centroid between both omics data blocks, with vectors from this centroid pointing to the sample's projection from each individual block (pMAGs in purple; Virus in orange). The length of a vector indicates the degree of disagreement between the blocks for that sample, with longer arrows representing a stronger block-specific signal. (B) Heatmap of variables selected by MINT BLOCK sPLS-DA (component 1), showing the differential abundance of discriminant seawater pMAGs (purple) and viruses (orange) across reef samples colored by reef zoning status. (C) Distribution of high-confidence iPHoP prediction scores (90-100%, x-axis) for viral contigs identified as indicators of NTMRs (left) or fished reefs (right), showing their predicted prokaryotic host pMAG95%ANI collapsed at family level (y-axis). Points represent individual virus-host predictions, colored by the indicator zoning status (NTMR, fished, non-discriminative) of the host pMAG95%ANI. Alluvial diagrams were used to show the taxonomy of indicator viral contigs.

Figure 4.2: Comparative performance of single- and multi-omics models. Boxplots show the difference in model accuracy scores in discriminating between NTMRs and fished reefs, when integrating pMAGs and viral data to identify multi-omics signatures of NTMRs and fished reefs (MINT BLOCK sPLS-DA; A), and when performed on single-omics data (MINT sPLS-DA from pMAGs - B; and viruses - C). Dashed lines (A-C) show model classification accuracy averaged across sectors. Venn diagrams show agreement between zoning biomarkers selected in single-omics (MINT sPLS-DA, Component 1) and multi-omics (MINT BLOCK sPLS-DA, Component 1) approaches for (D) pMAGs and (E) viral contigs.

Figure 5.1: Data integration will be critical to identify novel microbial indicators of reef health from globally emerging data on reef bacterioplankton. (A) Global coral reef distribution and recent large-scale seawater microbial surveys. (B) Omics data can be represented as matrices of samples (rows) and features e.g genes, transcripts, proteins, metabolomes) in columns. Integration can occur (C) across different omics types measured on the same samples (N-integration) to for example connect genetic potential with molecular activity, or (D) within a single omics type to identify universal indicators via meta-analysis (P-integration). (E) Viral, prokaryotic and eukaryotic microbial 'layers' can potentially also be partitioned from a single-omics study (e.g., metagenomics) to investigate how environmental change affects microbe-to-microbe interactions.

1.7 List of Tables

Chapter 1

General Introduction | The road forward to incorporate seawater microbes in predictive reef monitoring

This chapter is published in *Environmental Microbiome* as:
Terzin, M., Laffy, P.W., Robbins, S., Yeoh, Y.K., Frade, P.R., Glasl, B., Webster, N.S., Bourne, D.G., 2024. The road forward to incorporate seawater microbes in predictive reef monitoring. Environ. Microbiome 19, 5. https://doi.org/10.1186/s40793-023-00543-4

1.8 Abstract

Marine bacterioplankton underpin the health and function of coral reefs and respond in a rapid and sensitive manner to environmental changes that affect reef ecosystem stability. Numerous meta-omics surveys over recent years have documented persistent associations of opportunistic seawater microbial taxa, and their associated functions, with metrics of environmental stress and poor reef health (e.g. elevated temperature, nutrient loads and macroalgae cover). Through positive feedback mechanisms, disturbance-triggered heterotrophic activity of seawater microbes is hypothesised to drive keystone benthic organisms towards the limit of their resilience and translate into shifts in biogeochemical cycles which influence marine food webs, ultimately affecting entire reef ecosystems. However, despite nearly two decades of work in this space, a major limitation to using seawater microbes in reef monitoring is a lack of a unified and focused approach that would move beyond the indicator discovery phase and towards the development of rapid microbial indicator assays for (near) real-time reef management and decision-making. By reviewing the current state of knowledge, we provide a comprehensive framework (defined as five phases of research and innovation) to catalyse a shift from fundamental to applied research, allowing us to move from descriptive to predictive reef monitoring, and from reactive to proactive reef management.

1.9 Introduction

Coral reefs are some of the most biodiverse and productive aquatic environments on the planet, providing shelter, nutrition, and habitat for many marine species, and offering valuable ecosystem services to humans, including protection of coastal areas, tourism, and fisheries[5,6]. Despite their ecological significance and economic value, coral reefs have suffered major declines in recent decades due to the synergistic effects of local chronic impacts and global climate change[7,8] with recent estimates indicating that half the world's coral cover has been lost since the 1950s[9]. To preserve coral reefs, an improved understanding is needed of the mechanisms involved in coral resilience to local and global environmental stressors.

Marine microorganisms account for ~65-90% of the marine biomass[10,11] and therefore constitute the life support system of the biosphere, being central to planetary marine food webs and biogeochemical cycles, and responsible for approximately 50% of the world's primary production[11–18]. Marine plankton also play a vital role in the stability and function of coral reefs by providing crucial ecosystem services. Heterotrophic microbes in reef seawater rapidly capture and recycle nutrients from the water column, for example, dissolving coral derived mucus before it sinks to the sediment[19]. By rapidly taking up nutrients from the water column, seawater microbes have a critical role in making these nutrients available to higher trophic levels[19–22]. This efficient recycling of nutrients ultimately allows corals to thrive in oligotrophic and nutrient-deplete environments, often referred to as 'marine deserts'[22].

Host-associated microbes also provide various functions to their metazoan hosts, including nutrition, removal of waste products (e.g. ammonia), protection from invading pathogens, and stimulation of developmental processes and morphogenesis[23–28]. However, environmental stressors such as eutrophication and elevated temperatures may shift host-associated microbial communities from mutualistic to pathogenic states once critical thresholds are reached[29–31]. The emergence of copiotrophic and potentially pathogenic microbes (e.g. Flavobacteriaceae, Cryomorphaceae, Rhodobacteraceae, Rhodospirillaceae, *Vibrio*) along with their associated functions (e.g. virulence factors and toxin production) has been associated with increases in coral diseases leading to tissue necrosis, and ultimately partial or whole colony mortality[30,32,33].

This sensitivity of reef microbes to environmental perturbations potentially allows microbes to be used as indicators of environmental change in the surrounding reef[34,35]. Importantly, reef microorganisms may represent early warning indicators of environmental disturbances since microbial communities change in their composition and function before the development of visual signs of stress, such as coral disease, bleaching, and tissue necrosis[29,34–43]. These traditional visual signs of reef disturbance often become evident only after prolonged periods and potentially once ecosystem tipping points are reached[44]. Current monitoring efforts are therefore often reactive, reporting the outcomes of impact with limited potential to mitigate future reef decline. Incorporating microbial processes within reef monitoring frameworks could represent a powerful way to observe early signs of stress, providing more time to implement reef management strategies and mitigate the impacts of environmental disturbances on reefs[34,35].

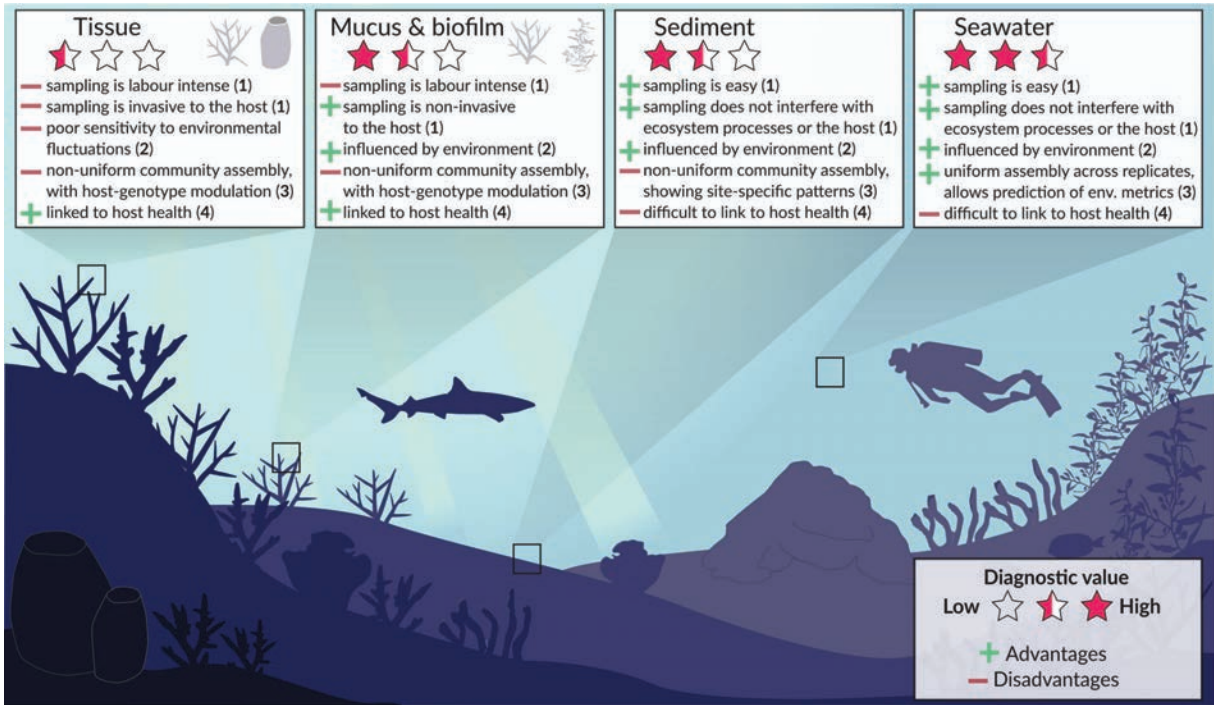

Figure 1.1: Overview of the diagnostic value of various coral reef microbiomes. The diagnostic value (indicated as stars) is based on the sum of advantages (+) and disadvantages (-) for key characteristics of optimal microbial indicators: (1) ease of sampling, (2) sensitivity towards environmental fluctuations, (3) uniformity of community assembly, as well as (4) our ability to link microbiome shifts to host health. Based on these criteria, seawater microbial communities collectively have the highest diagnostic potential to be used as microbial indicators of reef health, followed by sediment-associated and host-associated microbial communities, respectively. Free-living microbial communities (seawater and sediment) can be easily collected, without interfering with ecosystem processes and/or the health of reef organisms, consistent with desirable characteristics for environmental monitoring programs. In contrast, the collection of host-associated microbiomes is labour intensive and potentially poses a certain risk for host health when collecting tissue, although collections of the host-biofilm are non-invasive for the host. Seawater also revealed the highest sensitivity to changes in the surrounding environment (e.g., temperature and eutrophication) due to uniform community assembly patterns of the seawater microbiome across replicates, while sediments were primarily influenced by site-specific patterns (e.g. grain size) and host-associated microbiomes predominantly showed a host-genotype modulation. While the diagnostic value is highest for most criteria in the seawater microbiome, it is challenging to link disturbance-induced shifts in marine bacterioplankton to host health. Given the importance of host-associated microbes to the health of reef holobionts, the establishment of microbial baselines for host-associated microbiomes and the search for host health microbial indicators are still warranted.

A framework to implement microbial observations within reef monitoring programs was recently proposed[40]. Using indicator value analysis and machine learning approaches, seawater microbial communities (inferred from 16S rRNA gene amplicon sequencing data) were documented to provide accurate predictions of water temperature and eutrophication states of reefs. In contrast, macroalgae, coral and sponge microbiomes were predominantly structured by the host organism and less influenced by the environment[40]. Exposure of coral and sponge species to non-lethal stressors (temperature, acidification and salinity) in controlled experimental systems has similarly demonstrated that host factors strongly influence host-associated microbiomes[45–49], possibly limiting their use as early indicators of stress in reef monitoring. These trends were recently also documented at scale, using 16S rRNA gene amplicon sequencing data from the Tara Pacific Expedition[50]. Only 4–11% of variance in the coral microbiomes surveyed (*Millepora*, *Porites* and *Pocillopora*) was explained by physicochemical properties of the seawater, compared with ~30% variance in planktonic microbial communities explained by water chemistry[50]. Considering that seawater microbes provide accurate diagnostics of temperature and eutrophication states in the reef environment[40] and that seawater can be easily collected alongside *in situ* reef health surveys in a cost-effective and non-destructive manner, we assert there is realistic scope to incorporate microbial observations of seawater microbes alongside ongoing *in situ* reef health surveys (**Fig. 1.1**). To do so, it is important that monitoring programs incorporate temporal sampling designs that account for day-night cycles to minimise potential confounding effects, as recent studies have documented significant diel shifts in both the taxonomic structure and functional potential of reef bacterioplankton[51,52].

Apart from early detection of environmental changes (**Fig. 1.2**), seawater microbes are also important to predict reef functioning as environmental perturbations can destabilise reef bacterioplankton and alter their ecosystem services (**Fig. 1.3**), resulting in adverse implications on future reef dynamics via cascading effects and feedback loops[43,53,54]. For example, cumulative effects of

nutrient eutrophication and elevated temperature can trigger heterotrophic microbial activity in seawater, resulting in harmful algae blooms and hypoxia at reef scales causing rapid coral mortality[54–56]. Heterotrophic seawater microbes were also proposed to be central to large-scale reef declines caused by chronic stressors such as elevated nutrients and overfishing via the DDAM (disease, dissolved organic carbon (DOC), algae and microbes) model[53,57,58]. The DDAM positive feedback loop begins with eutrophication and overfishing facilitating growth of fleshy macroalgae, which confers a competitive advantage to other macroalgae over coralline algae and calcifying corals by preventing settlement of coral larvae[57]. At the same time, ocean warming caused by climate change stimulates the release of dissolved organic carbon (DOC) by fleshy macroalgae, which results in the proliferation of copiotrophic and potentially pathogenic bacterial communities in seawater, a process referred to as microbialisation. Increased abundance and activity of opportunistic and potentially pathogenic microbes in the water column further fuels the DDAM positive feedback loop by causing additional coral decline through increased coral disease prevalence, which ultimately maintains algal competitive dominance[57]. This concept of microbialisation links changes in seawater reef microbes to reef health decline and is therefore important from a predictive monitoring perspective (**Fig. 1.3**).

Currently most of what we know about the potential for seawater microbes to predict reef health has been inferred from microbial taxonomy rather than microbial function[22]. Amplicon sequencing of the 16S rRNA gene (the universal taxonomic marker gene in Bacteria and Archaea) has identified opportunistic and potentially pathogenic microbes persistently associated with degraded reefs across independent meta-omics studies[30,43,58–61]. However, it is still unclear how seawater microbes can be applied as indicators of coral reef ecosystem health as (1) functional characterisation (i.e. survey of functional potential via metagenomic sequencing) of reef bacterioplankton is still largely lacking (but see Tara Pacific Expedition[62–65]), and because (2) frameworks still need to resolve the ecologically important functions that seawater microorganisms provide to the reef ecosystem. Meta-omics approaches (e.g., metagenomics, metatranscriptomics, metaproteomics and metabolomics) that survey specific functional genes extend beyond taxonomy and would allow scalable investigation of adaptive (i.e. community turnover) and acclimatory (e.g. physiological and gene expression changes) responses of microbes to their environment[66]. For example, undertaking meta-omic surveys to document microbial functional potential as part of reef microbial observations can elucidate how environmental change affects ecosystem services that seawater microbes provide to coral reefs, and to predict how this may translate to future reef dynamics (**Fig. 1.3**). However, when applying meta-omics for ecosystem monitoring, there are numerous strengths and weaknesses associated with the various technologies that need to be considered. These methodological considerations have been extensively reviewed previously[66–71] and are therefore not covered in detail here. In this review, we demonstrate that there is realistic scope to extend reef monitoring efforts by including microbial meta-omic surveys that capture the metabolic and functional potential of free-living microbes in seawater. We discuss the current state of using seawater microbial indicators in reef monitoring programs in the context of a five-step framework (**Fig. 1.4**). We hope this review will accelerate the

shift from fundamental towards applied research to develop rapid and cost-effective microbial-based assays for assessment of reef health, which would be invaluable in predictive reef monitoring and proactive management.

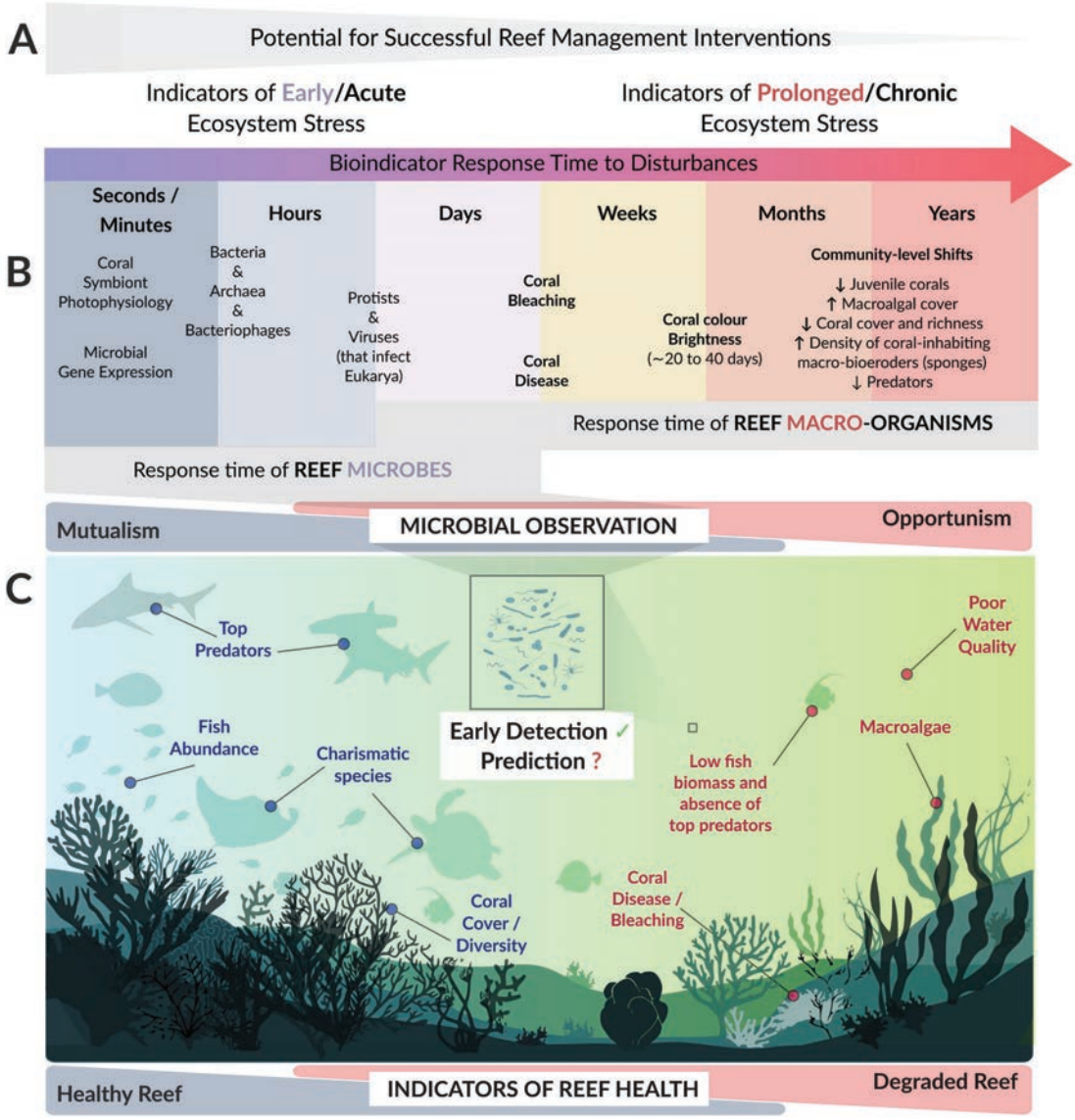

Figure 1.2. Potential of reef seawater microbes to inform on reef health status. Successful reef management interventions need to rely on acute and early identification of changes in the reef, before ecosystem 'tipping points' are reached (A). However, most reef monitoring programs are based on visual signs to assess ecosystem stress (e.g., coral disease, bleaching and community-level shifts), which become evident only after prolonged environmental disturbances (B). Due to their short generation times, seawater microbes respond rapidly to environmental changes, and it has therefore been well established that marine bacterioplankton allows accurate and early diagnostics of environmental fluctuations in the reef (C, middle). However, the predictive potential of seawater microbiome has been largely unexplored and it remains unclear how environmental changes will alter microbial functioning of reef bacterioplankton, and how this may translate to reef ecosystem functioning via cascading effects and feedback loops (C, middle). Fig. 1C was adjusted from Vanwonterghem and Webster (2020) with permission from authors.

1.10 Seawater microbes are essential to predict ocean and reef health

Enormous amounts of molecular and environmental data have been collected in recent years on oceanic microbes, both through various global sampling efforts that collected snapshots of marine microbes in time and space[12,14–16,18,62–64,72–74] as well as various long-term microbial observatory stations[17]. These large microbial oceanography initiatives aim to predict how environmental change alters the distribution patterns as well as taxonomic and functional diversity of ocean plankton at global scales[11,13,18,75,76]. To successfully integrate these large and complex datasets, novel computational approaches such as multi-omics data integration[4], ecological niche modelling[77,78], network analysis[78,79], multivariate statistics and supervised learning needed to be applied, and these modelling and data integration efforts have already enabled marine scientists to transition from hindcasting to forecasting. For example, tropical marine biogeographical provinces are predicted to expand towards the poles due to climate change (at the expense of temperate and polar zones), followed by a compositional shift in marine plankton which is projected to decrease carbon export fluxes and affect nitrogen cycling[76]. Furthermore, marine viruses were identified as the best predictors of global ocean carbon flux in comparison with archaea, bacteria and eukaryotes[79]. Further integration of seawater microbial meta-omics data into models of Earth system functioning will be crucial to improve such models, as marine bacterioplankton are directly involved in the processes being modelled such as biogeochemical cycling, primary production, and carbon efflux under climate change scenarios[80].

Statistical learning models have also been applied to coral reefs to identify microbial indicators that inform ecosystem health, though at comparatively smaller scales[40,43]. For example, random forest machine learning identified that seawater surface temperature in the Great Barrier Reef can be accurately predicted from reef bacterioplankton community structure[40], and a linear discriminant analysis (LDA) model was developed that accurately predicts reef categories (e.g. inshore, mid-shelf or offshore) in the GBR based on seawater microbial community profiles[43]. Importantly, vast amounts of meta-omic and multi-omic data streams have recently been collected on free-living and host-associated reef microbes in the Pacific Ocean (e.g. the Tara Pacific Expedition), which will allow further development of models to incorporate seawater microbes in predicting how climate change and human impact may affect reef functioning and health[62–64,72,81]. The Tara Pacific Expedition (2016-2018) has sampled seawater and coral for multi-omics sequencing in 32 island systems throughout the Pacific along with extensive environmental metadata, hence establishing spatial baselines of reef microbes in the Pacific Ocean[50,62–64,72,81].

These large-scale meta- and multi-omics datasets will, for the first time, provide the necessary basis to assess how functions of reef seawater microbes (e.g. photosynthesis, nitrification, ammonia oxidation, sulfate reduction, methanogenesis, virulence etc.) shift with the environment at global scales. Such datasets will be crucial to extend beyond localised studies that have already identified seawater microbes indicating poor reef health, by establishing a robust baseline of microbial indicators that are shared across wide spatial and temporal scales. A recent literature review provided such a

baseline, summarising how reef habitat degradation across regions in the Pacific Ocean and the Caribbean may alter microbe–DOM interactions, and the potential implications of shifts in microbial functioning contributing to further reef declines[82]. Further, analysis of large-scale meta-omics data of the seawater microbiome surrounding corals could provide insight about how local settings affect reef bacterioplankton, and how this may structure and affect dynamics of coral microbiomes[72,81], improving our understanding of the role of seawater microbes in reef resilience and acclimatisation (**Fig. 1.3**). Considering the implication of free-living seawater microbes in feedback loops and cascading effects, developing regulatory guidelines is needed to protect seawater microbial functions that are ecologically relevant to the reef, which would be invaluable in reef monitoring if early detection of how specific microbial functions are disrupted could be used to predict and avoid additional coral mortality and reef declines[83].

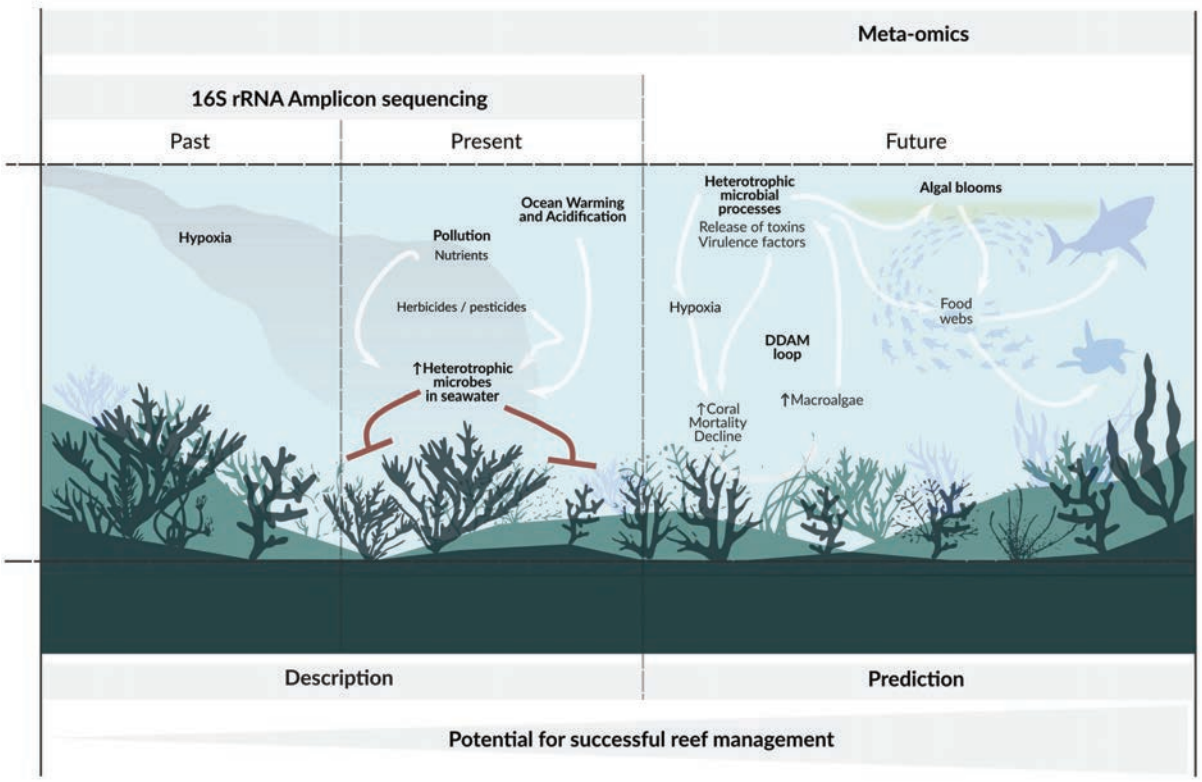

Figure 1.3: Reef microbial observation should extend beyond taxonomy and towards function, to move from descriptive to predictive reef monitoring. So far, 16S rRNA amplicon-seq data clearly showed that opportunistic and potentially pathogenic microbes robustly correlate to degraded reefs where we document poor water quality, increased macroalgae cover and coral disease/bleaching. However, amplicon-seq data has a limited resolution to go beyond description of past or present changes in the reef, as the consequences of the enrichment of particular microbial indicator taxa on reef health often cannot be inferred from microbial taxonomy alone (left, shown in red). Microbial meta-omics data would allow prediction of how environmental changes will affect the services microbes provide to coral reefs (e.g, primary productivity, nutrient/biogeochemical cycling, and exposure to pathogens), and how the altered microbial activity may translate to reef ecosystem dynamics (right). This predictive monitoring is needed for successful reef management and decision making (bottom).

1.11 Experimental validation of microbes that predict poor reef health

Large meta- and multi-omic data streams such as Tara Pacific will be vital for identifying the indicator taxa and genes that predict poor reef health, however, this large data collection exercise is just one step in incorporating microbial related processes in reef monitoring frameworks (**Fig. 1.4**, phase 1). In many instances the microbial features selected by the models may be unrelated to degraded ecosystem health despite high correlations identified by the model[84]. As an example, an environmental overlap may exist between microbial indicators that correlate to coral bleaching (which occurs under accumulated thermal stress) and warm-water-associated bacteria which also proliferate at elevated temperatures but do not impact reef health. Further, anaerobic bacteria and genes (which could indicate hypoxia in the reef) can still be present in the seawater days after a hypoxia-induced coral mortality event, and despite concentrations of dissolved oxygen having reverted to normal values[85]. Various additional factors may contribute to the noisy signal of microbial taxa-environment associations, including biotic interactions, limits to spatial dispersal and neutral demographic drift[86]. As there is a strong likelihood that only some microbial predictors of degraded reefs identified by the model will have a causal association with metrics of poor reef health, confirmatory experiments are needed to validate microbiome-environment associations[84] allowing interpretation of the suitability of the models and usefulness of the microbial indicator(s) (**Fig. 1.4**, phase 2), however this remains a distant goal. For example, while pathogenic microbes are hypothesised to play a role in coral disease, the causative microbial agents of coral disease have only been identified in a few cases, and often Koch postulates prove inconclusive[87–91]. The utility of seawater microbes as indicators of poor reef health should not be dismissed even when causal relationships are unknown, as there are promising microbes (most notably Flavobacteriaceae-affiliated taxa) that are persistently diagnostic of reef ecosystem degradation across independent omics studies[1,40,43,58–61,92]. Below we provide a shortlist of seawater taxa/genes that have been experimentally validated as indicators of poor reef health, and highlight they should be further studied to provide richer insights into their potential (causal) role in reef degradation.

Experimental validation of the DDAM loop and the microbialisation concept has confirmed bacterioplankton communities of coral- and algae-dominated reefs persist in laboratory conditions. A bottle experiment identified that rates of bacterioplankton growth and utilisation of DOC were elevated in algal exudate treatments compared to incubations with coral exudates and the control treatment, with macroalgae-derived sugars selecting for a less diverse community enriched in lineages of opportunistic Gammaproteobacteria including putative pathogens with known virulence factors (Pseudoalteromonadaceae and Vibrionaceae)[93]. Another bottle experiment identified that high DOC concentrations (as observed in macroalgae-enriched reefs) correlated to an enrichment of bacterial genera

Alteromonas, *Oceanicola*, *Erythrobacter*, and *Alcanivorax*, which shifted in their metabolic capabilities from mutualistic towards pathogenic states via up-regulation of genes encoding for metalloproteases, siderophores, toxins, and antibiotic resistance factors[94]. Similar trends were identified by *in situ* mesocosm studies which placed benthic chambers over coral-, sand-, and macroalgae-dominated communities, to identify that macroalgae exudates facilitated a shift to a net heterotrophic system, with pelagic microbial communities displaying elevated consumption of macroalgae-released DOC as well as increased oxygen consumption[58]. Experimental validation of the DDAM loop clearly shows that the addition of macroalgae-derived nutrients (under laboratory conditions) causes microbial proliferation and a shift towards pathogenesis and carbon metabolism pathways that are less energetically efficient[58,93,94], as also observed in the field[53]. However, the cause-and-effect understanding of the DDAM mechanism still needs to be teased apart to understand if and how the increased abundance and activity of microbial copiotrophs and putative pathogens in seawater at macroalgae-enriched reefs directly contribute to coral disease and reef declines, before applying the concept of microbialisation in predictive reef monitoring and proactive management.

Hypoxia also represents a crucial mechanism in the DDAM feedback loop, as macroalgae-released DOC fuels heterotophic activity and respiration by seawater microbes, which can create localised hypoxic regions at the coral-macroalgae interaction zones[53,82,95–97]. Experimental studies show that the addition of antibiotics may eliminate hypoxia in coral–algal interfaces[95–97]. As hypoxia predominantly occurs at coral–algal interaction zones, reef water away from the benthos may not be enriched in anaerobic microbial taxa and genes (but see Walsh et al. (2017)[98]). We therefore propose that monitoring for anaerobic microbes and functions is particularly relevant at the benthic and pelagic boundary layer, which represents a potential valuable environmental niche for reef monitoring to predict rapid declines in benthic organisms caused by hypoxia.

Another forecasting potential of seawater microbial monitoring is to predict coral disease outbreaks. The concept that animal disease outbreaks are driven by environmental change (e.g. climate warming) is well accepted across many terrestrial, freshwater and marine ecosystems[91,99–103], and acquisition of microbial pathogens from the environment has been documented in some food- and waterborne diseases[41,104]. For corals, there are many examples of putative microbial pathogens (isolated from diseased coral tissues) also being identified in the surrounding reef seawater, which often increase in their abundance and activity at elevated seawater temperatures[92,105,106]. The bacterial genus *Vibrio* sp. is particularly prominent as a potential causative agent of coral disease[90,91]. A survey to elucidate *Vibrio* diversity in surrounding reef seawater using the well-curated pyrH (uridylate kinase) gene sequence identified that putative coral pathogens (i.e. *V. coralliilyticus, V. neptunis*, and *V. owensii*) were persistently present in the seawater of the Ishigaki coral reef system (Japan) across the entire 3-year survey period, with increased abundances correlated with elevated seawater temperatures for the majority of *Vibrio* species[92]. Another study identified a significant enrichment in *Planctomycetota* (lineages OM190 and

CL500-3) and bacteria within genera *Synechococcus* and *Vibrio* during the marine heatwave on the GBR in April 2016, alongside an enrichment of *Vibrio*-derived virulence factors (i.e. metalloprotease genes vcpA, vcpB and vchA in *V. coralliilyticus*)[106]. Though as the authors highlight, a link between large-scale changes observed in the plankton-associated microbial community and reef ecosystem health could not be established based on their observations[106], which warrants for robust experimental validation to gain a cause-and-effect understanding between the presence of potentially pathogenic water-born microorganisms and coral disease outbreaks.

Interestingly, the signal of copiotrophic and potentially pathogenic microbes in seawater often persists even after the environmental disturbances have passed[85,106], and this concept is known as the microbial 'legacy effect'[84]. Anaerobic microbes can persist in seawater for days after dissolved oxygen concentrations revert to normoxic values[85], and the microbial signal of potentially pathogenic microbes (e.g. *Planctomycetota* and *Vibrio*) and functions (e.g. *Vibrio*-derived metalloproteases) that were enriched during the marine heatwave on the GBR in April 2016 remained apparent until August 2016, months after the marine heatwave had dissipated[106]. This microbial 'legacy' effect may be important to explore from a monitoring perspective of free-living seawater microbes, as a shift in microbial functioning may cause additional coral decline even after the disturbance has passed. As an example, a number of studies have documented that coral disease outbreaks can exacerbate the impacts of bleaching events[107–112]. The cumulative stress corals face during thermal stress may make them susceptible to opportunistic microbial pathogens that persist in the seawater following marine heatwaves[106], further compromising their health and increasing mortality. Experimentally validating this 'legacy effect' may be crucial from a monitoring perspective to understand how long opportunistic microbes and functions persists after different environmental disturbances, and to predict if this may affect future reef dynamics.

1.12 Formulation of seawater microbial indices for reef monitoring

Once experimentally validated, a list of microbial taxa and/or functions can be compiled to formulate robust microbial indicators which associate to metrics of poor reef health across both field and laboratory studies (**Fig. 1.4**, phase 3). Such efforts have already been made to formulate microbial indices for reef monitoring based on free-living seawater microbes. Most notably, microbialisation scores (defined as the ratio of metabolic rates between bacterioplankton and reef fish) have been proposed within the DDAM model as a metric of human impact on coral reefs[53,82,98,113,114]. This concept of microbialisation (a shift in biomass production and metabolic rates from macro to micro-organisms) has been well documented in macroalgae-enriched reefs in field observations[36,114], across local and regional scales[53], and also validated experimentally in laboratory bottle experiments[93,94,115] and *in situ* mesocosm studies[58]. Despite their potential, microbialisation scores have not been implemented into standard reef monitoring efforts to date, primarily since the scores represent a metric relevant to shifting coral-algal dynamics,

which is not universally applicable to reefs under environmental pressures that still maintain high coral cover and/or high fish biomass.

A recent meta-analysis of reef bacterioplankton identified several microbial indices of poor reef health across the Great Barrier Reef[43]. It was proposed that Prochlorococcaceae and Synechococcaceae families represent potential indicators of cross-shelf nutrient levels with an increasing Synechococcaceae:Prochlorococcaceae abundance ratio being a proxy for increased nutrient loads in reef waters[43]. In addition, Flavobacteriaceae-affiliated taxa were potentially diagnostic of reef ecosystem degradation, with an increasing abundance ratio of copiotrophic (e.g. OCS155, Flavobacteraceae, Cryomorphaceae and Rhodobacteraceae) taxa relative to oligotrophic taxa (e.g. Pelagibacteraceae (SAR11) and SAR86) as an index of eutrophication[43]. Finally, an increasing prevalence of opportunistic and potentially pathogenic taxa (Rhodospirillaceae, Rhodobacteraceae and Vibrionaceae) were also indicative of degraded inshore reef systems[43]. Another study proposed that increased Bacteroidota (prevalent in waters of inshore reefs in the GBR) relative to Alphaproteobacteria (more abundant in offshore GBR reefs) in reef surface waters could indicate enhanced macroalgae growth, elevated nutrient loads, and the start of microbial proliferation for inshore coral reefs on the GBR, which aligns with the concept of the DDAM loop[1]. Recently, a detailed overview has been provided on indicator microbes across different reef benthic habitats (i.e. nearshore and offshore, or coral- and macroalgae-dominated), highlighting the ubiquity of these patterns in the Great Barrier Reef, Caribbean, and Pacific Ocean regions (Table 1 within Nelson et al. (2023)[82]). This baseline knowledge is relevant as it provides a list of putative indicator microbes (with their expected relative abundances in different habitats) as stable predictors of poor reef health at broad spatio-temporal scales[82], which can be used as a starting point to create applied assays for rapid reef health assessment in the field using our framework (**Fig. 1.4**).

Despite their potential, seawater microbes are still largely overlooked by reef health surveys and these seawater microbial indices are yet to be validated as useful monitoring assays in the field (**Fig. 1.4**, phase 4). Such applied research should be pursued as the seawater microbiome possesses numerous additional characteristics that align with criteria of good indicators[34], in addition to its utility to infer and predict environmental fluctuations[40,50]. Seawater sampling and processing is simple, non-destructive and can be performed with minimal training required, which facilitates large-scale sampling alongside ongoing *in situ* coral health surveys that already collect metrics on water chemistry and benthic cover in the reef. Furthermore, seawater collection and processing protocols are largely standardised due to global plankton sampling expeditions such as Tara Ocean and Tara Pacific[11,15,18,62–64,72,81] which also ensures comparability of data streams from different studies. Lastly, while monitoring whole-community dynamics is preferred compared to focusing on a subset of indicators, it is simply not feasible for macro-organisms in highly biodiverse ecosystems such as coral reefs. However, whole-community monitoring can be done when working with seawater microbial communities, and tolerance thresholds to environmental disturbances can be determined for individual microbial species from seawater. Such

information could be utilised to construct cumulative species and functional sensitivity distributions, allowing to quantify the proportional impact of environmental stress on the entire microbial communities[83].

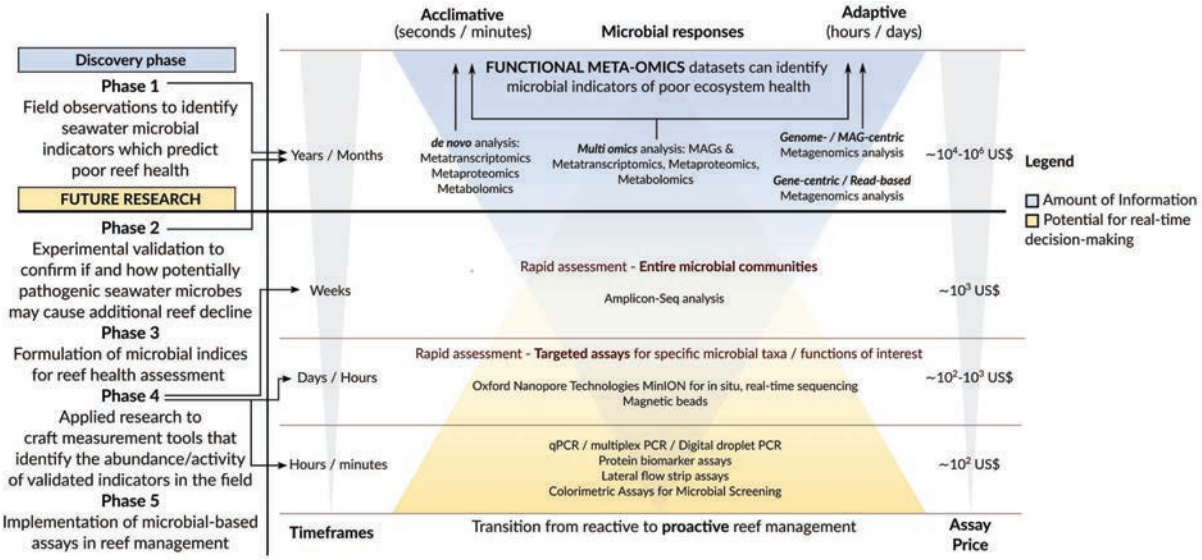

Figure 1.4. The proposed five-step framework of research and innovation to move from descriptive to predictive reef microbial monitoring, and from reactive to proactive reef management. Functional meta-omics datasets are critical to discover microbial indicators of poor reef health in the field (**Phase 1**), however high costs (see 'Assay price') and long bioinformatics processing times (see 'Timeframes') of microbial meta-omics datasets suggest their limited utility for rapid decision-making in reef management. We highlight that this milestone has been largely achieved through various localised studies, though in the years to come, the integration of recently generated datasets obtained in large-scale surveys (most notably the Tara Pacific Expedition) will be crucial to understand the ubiquity of identified microbial indicators at global scales. Once microbial indicators of poor ecosystem health are identified based on functional meta-omics datasets, experimental validation (**Phase 2**) is needed to confirm the same patterns occur in laboratory conditions, as well as to identify the causality of microbiome-environment associations from the field, which we predict still remains a distant goal and will require years of research. Once experimentally validated, microbial indices can be formulated (**Phase 3**) and applied research can commence to develop rapid (within weeks, days or minutes, see 'Timeframes') and cost-effective (see 'Assay price') assays to quickly assess reef health in the field (**Phase 4**), which can be used in proactive reef management and rapid decision-making (**Phase 5**).

1.13 Conclusions - A framework for incorporating microbial indicators into coral reef management

This review discusses the current state of using seawater microbes as indicators in predictive reef monitoring in the context of a five-step framework (**Fig. 1.4**), building on frameworks proposed by Parkinson et al. (2020) and Webster and Gorsuch (2020)[116]. Many small-scale field studies combined with the emerging global studies (e.g. Tara Oceans) have identified candidate microbial taxa and genes that predict poor ecosystem and reef health[29,35–38,40,43], hence the barrier does not lie in the indicator discovery phase (**Fig. 1.4**, phase 1) but largely in subsequent phases of experimental validation (**Fig. 1.4**, phase 2), formulation of microbial indices (**Fig. 1.4**, phase 3), applied research to generate microbial assays that can be used in the field (**Fig. 1.4**, phase 4), and implementation of microbial indicators in reef management and decision-making processes (**Fig. 1.4**, phase 5). Phase 1 (indicator discovery) is an ongoing process with large-scale multi-omics data streams, many that are yet to be published (e.g. Tara Pacific), fundamental to identify novel indicators and generate spatio-temporally coherent baselines of seawater

microbial predictors that associate with metrics of poor reef health. Phases 2 (Experimental validation) and 3 (Formulation of seawater microbial indices to predict poor reef health) are currently a work in progress (**Fig. 1.4**) with a few studies experimentally validating microbial indicators (**Fig. 1.4**, phase 2), although primarily in the context of the DDAM loop, and we still lack conclusive experimental evidence that water-born microbial pathogens can indeed cause coral disease. Further, some indices based on seawater microbes (most notably the microbialisation scores) have been formulated to assess reef health (**Fig. 1.4**, phase 3), but these indices are still not used in standard reef monitoring.

Validated microbial-based molecular assays for rapid screening of seawater microbial indicators to predict reef decline are yet to be crafted (e.g. screening for anaerobic microbes to predict hypoxia-induced coral mortality events), hence phase 4 (applied research to craft microbial diagnostic and predictive tools) still needs to be developed. Such rapid and cost-effective assays based on seawater microbes (PCR, magnetic beads, and proteomic/colorimetric assays) have been successfully applied for environmental management (**Fig. 1.4**, phase 5), although generally inform on single stressors or have a narrow focus on risks to human health and well-being. Some examples include the presence/increased abundance of coliforms in public swimming waters indicating faecal pollution[117,118], enrichment of antibiotic resistance genes indicating human impact[119], enrichment of hydrocarbon-degrading taxa and genes tracking oil spills[120] and anaerobic genes from sulphur-oxidising bacteria as indicators to trace the spread of oxygen minimum zones in the ocean[121]. Instead of assaying the reef environment for individual microbial indicators, it is potentially more productive to compile a list of target microbial taxa and functions that associate to poor reef health. For example, potentially pathogenic microbes increase in abundance and/or activity at elevated temperature and nutrient concentrations[91,92,106,122]. Therefore screening for these taxa is a necessity during the summer period, particularly before, during and after bleaching events when coral health becomes compromised.

To move towards proactive reef management, improved communication between researchers and practitioners is needed to determine whether microbial indicators are desired in reef monitoring, as well as a cost/benefit analysis to identify which putative markers should be prioritised in applied research to develop targeted microbial-based assays. By reviewing current knowledge gaps, we highlight that seawater microbes should not be overlooked in reef monitoring efforts as marine plankton is an essential proxy of reef health, and we hope this review will catalyse further research towards predictive reef microbial monitoring and proactive management, which can be achieved if objectives are aligned between scientists, managers, and funding bodies.

1.13.1 List of abbreviations

IMOS – Australia's Integrated Marine Observing System. GBR – Great Barrier Reef. DDAM – disease, dissolved organic carbon, algae, microbes. DOC - dissolved organic carbon. LDA - linear discriminant analysis. rRNA – ribosomal ribonucleic acid.

1.14 Declarations

1.14.1 Ethics approval and consent to participate

Not applicable.

1.14.2 Consent for publication

Not applicable.

1.14.3 Availability of data and material

Not applicable.

1.14.4 Competing interests

The authors declare no conflict of interest.

1.14.5 Funding

This study forms part of the Australia's Integrated Marine Observing System (IMOS) Great Barrier Reef Microbial Genomic Database sub-facility, funded by the Queensland Research Infrastructure Co-investment Fund (RICF) by the Department of Environment and Science, Queensland. IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent. This study was also funded by an AIMS@JCU PhD Scholarship to M.T. The funders had no role in performing the literature review, preparation of the manuscript, or decision to publish.

1.14.6 Authors' contributions

M.T performed the literature search. M.T, D.G.B and P.W.L conceived the review topic and scoped the sections. M.T, D.G.B and P.W.L wrote the manuscript, and S.R, Y.K.Y, P.R.F, B.G, and N.S.W reviewed and edited the manuscript, and provided input on the full manuscript. M.T and B.G created the figures.

1.14.7 Acknowledgements

This work took place at the Australian Institute of Marine Science (AIMS) headquarters at Cape Ferguson, and we wish to acknowledge the Wulgurukaba and Bindal peoples as the Traditional Owners of that land. This research was also undertaken at the JCU Townsville Bebegu Yumba campus, and the authors acknowledge that the Australian Aboriginal and Torres Strait Islander peoples are the original

inhabitants and traditional custodians of this continent and have unique cultural and spiritual relationships to the land and waters. We pay our respects to their Elders past, present, and emerging.

Data Analysis | Gene content of seawater microbes is a strong predictor of water chemistry across the Great Barrier Reef

This chapter is published in *Microbiome* as:
Terzin, M., Robbins, S.J., Bell, S.C., Lê Cao, K.-A., Gruber, R.K., Frade, P.R., Webster, N.S., Yeoh, Y.K., Bourne, D.G., Laffy, P.W., 2025. Gene content of seawater microbes is a strong predictor of water chemistry across the Great Barrier Reef. *Microbiome* 13, 11.
https://doi.org/10.1186/s40168-024-01972-0

2.1 Abstract

Background: Seawater microbes (bacteria and archaea) play essential roles in coral reefs by facilitating nutrient cycling, energy transfer, and overall reef ecosystem functioning. However, environmental disturbances such as degraded water quality and marine heatwaves, can impact these vital functions as seawater microbial communities experience notable shifts in composition and function when exposed to stressors. This sensitivity highlights the potential of seawater microbes to be used as indicators of reef health. Microbial indicator analysis has centred around measuring the taxonomic structure of seawater microbial communities, but this can obscure heterogeneity of gene content between taxonomically similar microbes, and thus microbial functional genes have been hypothesised to have more scope for predictive potential, though empirical validation for this hypothesis is still pending. Using a metagenomics study framework, we establish a functional baseline of seawater microbiomes across outer Great Barrier Reef (GBR) sites to compare the diagnostic value between taxonomic and functional information in inferring continuous physico-chemical metrics in the surrounding reef.

Results: Integrating gene-centric metagenomics analyses with 17 physico-chemical variables (temperature, salinity, and particulate and dissolved nutrients) across 48 reefs revealed that associations between microbial functions and environmental parameters were twice as stable compared to taxonomy-environment associations. Distinct seasonal variations in surface water chemistry were observed, with nutrient concentrations up to 3-fold higher during austral summer, explained by enhanced production of particulate organic matter (POM) by photoautotrophic cyanobacteria, primarily *Synechococcus*. In contrast, nutrient levels were lower in winter and POM production was also attributed to *Prochlorococcus*. Additionally, heterotrophic microbes (e.g., Rhodospirillaceae, Burkholderiaceae, Flavobacteriaceae, and Rhodobacteraceae) were enriched in reefs with elevated dissolved organic carbon (DOC) and phytoplankton-derived POM, encoding functional genes related to membrane transport, sugar utilisation, and energy metabolism. These microbes likely contribute to the coral reef microbial loop by capturing and recycling nutrients derived from *Synechococcus* and *Prochlorococcus*, ultimately transferring nutrients from picocyanobacterial primary producers to higher trophic levels.

Conclusion: This study reveals that functional information in reef-associated seawater microbes robustly associates with physico-chemical variables than taxonomic data, highlighting the importance of incorporating microbial function in reef monitoring initiatives. Our integrative approach to mine for stable seawater microbial biomarkers can be expanded to include additional continuous metrics of reef health (e.g., benthic cover of corals and macroalgae, fish counts/biomass) and may be applicable to other large-scale reef metagenomics datasets beyond the GBR.

2.2 Introduction

Coral reefs globally are increasingly subjected to the impacts of climate change and anthropogenic activity[7,123,124], driving declines in the health of these critical ecosystems[9,125]. Early identification of adverse environmental conditions and declining reef health is important for the development of management strategies that can effectively mitigate the effects of environmental pressures[34,35,126–128]. Free-living seawater microorganisms are the first responders to environmental change on reefs owing to their rapid turnover rates measured in hours or days[22,40,129]. The utility of microbes for reef monitoring has been previously proposed[34,35,82,130], with many studies documenting rapid changes in the structure of seawater microbial communities on reefs subjected to environmental stress[43,54,55,60,92,131]. Seawater microbiomes have been shown to be up to 5-fold more accurate compared to sediment and host-associated (coral, sponge, and macroalgae) microbiomes in predicting temperature and nutrient concentrations on reefs[40]. This was attributed to planktonic communities being more uniform in their spatial and temporal distribution across reef waters in contrast to sediment microbes, which were highly site-specific (i.e. influenced by sediment grain size and chemical composition), and host microbiomes strongly influenced by host-genotype[40,50]. Moreover, seawater can be easily and non-destructively collected alongside *in situ* reef health surveys, hence there is realistic scope to complement ongoing reef monitoring programs with seawater microbial observations[128,130].

Microbial communities in seawater are influenced by various oceanographic processes such as transport, mixing, resuspension, and shelf upwelling, in addition to niches associated with water chemistry and/or interactions with surrounding benthic and pelagic communities. As such, the challenge with using seawater microbial communities as indicators of reef health is in assessing their associations to different environmental factors (e.g. temperature, salinity, nutrient concentrations, and local biodiversity) and whether the identified microbial indicators associate to the same environmental factors consistently across broad spatial and temporal scales. Further, associations between pelagic microbes and the environment are often documented as stochastic, which is partly explained by 'functional redundancy' within the microbiome[77,132,133], whereby genes for many metabolic functions are present across broad classes of microorganisms[75,134–137] and microbial communities therefore likely have many compositional alternatives for carrying out the same process in any given environment. This phenomenon raises the possibility that microbial metabolic function could more reliably reflect environmental metrics than taxonomic identity, and this has been reported across plant[86,138], soil[139], human gut[140,141], and marine microbiomes in pelagic waters[75,77,132,133,142]. Genes for metabolic cellular functions like photosynthesis, nitrification, ammonia oxidation, sulfate reduction, and virulence have also been proposed as having higher utility in predicting environmentally induced changes that translate to shifts in reef health[43,128,130]. However, it is important to note that recent findings indicate that in Florida reef waters, the taxonomic microbiome (16S rRNA gene) was a stronger predictor of both physico-chemical and benthic reef properties compared to the functional microbiome (metagenome) and metabolome of the reef water[143].

This highlights the need for further research to fully understand the potential contributions of functional genes in different reef ecosystems.

Previous studies documenting community composition of reef bacterioplankton (seawater bacteria and archaea) across the Great Barrier Reef (GBR) have indicated a large influence of geography and season[43] with different explanatory drivers identified across the GBR. Using 16S rRNA gene sequencing, reef bacterioplankton in inshore GBR reefs of the Wet Tropics region were shown to predominantly respond to riverine inputs characterised by declining salinity and elevated organic and inorganic nutrients[61]. In comparison, the main drivers on inshore reefs in the central GBR were temperature, total suspended solids, particulate organic carbon, and macroalgae[1,40]. Due to these differences in geographical sites and/or different times of sampling, potentially in addition to methodological variations in field sample collection and laboratory processing, these independent meta-omics studies have also identified somewhat inconsistent seawater microbial indicators for the same environmental metric. For example, Rhodobacteraceae and Flavobacteriaceae were identified as indicative of elevated nutrients in degraded inshore reefs of the central GBR[1,40] however they were not identified as indicators of nutrient enrichment and poor water quality in the Tully River region of the northern GBR[61]. While there have been attempts to consolidate microbial community composition and environmental data sets spanning the GBR (i.e. meta-analysis by Frade et al. (2020)[43]), associations between reef bacterioplankton composition and nutrients were largely partitioned by cross-shelf spatial variation, with heterotrophic microbes and reduced bacterial diversity documented in inshore reefs, in contrast to more diverse and autotrophic bacterioplankton communities in oligotrophic mid- and outer-shelf GBR surface waters[43]. These findings suggest that putative indicator taxa were unique to their respective region, and may not serve as a general indicator of a specific continuous environmental metric stably across the GBR. Importantly, it remains unknown how microbial functional potential changes across the broad spatio-temporal scales of the GBR as previous studies predominantly focused on taxonomically profiling reef bacterioplankton communities through 16S rRNA gene sequencing (notable exception: Glasl et al. (2020)[1]), which may mask variation hidden by functional redundancy. Therefore, here we measure microbial functional genes directly to assess their reliability as indicators of metrics relevant to reef health.

In this study, we perform a gene-centric analysis on surface seawater metagenomes collected from 48 offshore reefs (at ~5 m depth) across the length of the GBR, integrating microbial metagenomic and physico-chemical data to (1) identify stable microbial indicators (both taxonomic and functional genes) that consistently respond to specific physico-chemical variables (e.g., nutrient loads, temperature, salinity) across broad spatio-temporal scales in the GBR, and (2) to assess whether microbial taxa or functional genes exhibit greater stability in their associations with these environmental factors. To achieve objective (1), we extended a Sparse Partial Least Squares analysis (sPLS[144,145]) widely used in microbial oceanography to correlate microbial data with continuous environmental metrics[79,146,147] with a Multivariate INTegrative method (MINT[3]) to integrate data from four independent sampling trips. This omics integration approach

aimed to uncover microbial indicators that are stable/shared across trips, hence persistently correlating to the same physico-chemical variables across space and time in offshore GBR reefs. To achieve objective (2), we applied data perturbation with cross-validation (CV) to first quantify indicator statistical stability—measured as the reoccurrence of microbial indicator taxa or GO terms across independent CV runs—and subsequently evaluate the diagnostic potential (i.e. higher stability scores = higher diagnostic value) of microbial functional information in surpassing taxonomy for reef health assessments, which we hypothesised based on the principles of functional redundancy. Our results demonstrate the potential of reef seawater microbes to accurately inform nutrient concentrations, contributing to the potential to link seawater microbes and reef health.

2.3 Materials and Methods

2.3.1 Seawater collection and field processing

Surface seawater (at 5m depth, approximately 5-15 meters from the reef benthos) was collected for water chemistry analysis and microbial community profiling at 48 reefs spanning the GBR, with each sample being collected once in time (**Fig. 2.1**). Sampling was performed from the RV Solander and RV Cape Ferguson alongside AIMS Long-term Monitoring Program *in situ* reef health surveys across four trips between November 2019 and July 2020 (**Fig. 2.1**). The first three sampling trips occurred during the austral (i.e. in the Southern hemisphere) summer (wet season) in the far Northern GBR (trip 1: November-December 2019, Cape Grenville and Princess Charlotte bay sectors, see **Appendix A: Fig. S1**), the southern GBR (trip 2: January 2020, Swains and Capricorn Bunker sectors, see **Appendix A: Fig. S2**), and in the central GBR (trip 3: February 2020, Cairns and Innisfail sectors, see **Appendix A: Fig. S3**), while the last trip was performed during austral winter (dry season) and also in the central GBR (trip 4: July 2020, Townsville sector, see **Appendix A: Fig. S4**) (**Fig. 2.1**). The coordinates of the 48 surveyed reefs were visualised as maps in R Studio (R version 4.3.2)[148] as per: https://open-aims.github.io/gisaimsr/articles/examples.html, which used the following R packages as dependencies: raster[149], tidyverse[150], ggspatial[151], sf[152,153], dataaimsr[154], gisaimsr (https://github.com/open-AIMS/gisaimsr), and ggrepel[155].

Triplicate 5 L seawater samples were collected using Niskin bottles or by divers for analysis of water chemistry variables. A total of 14 water chemistry variables were measured using established methods[156], including ammonia (NH_4^+), nitrite (NO_2^+), nitrate (NO_3^+), total dissolved nitrogen (TDN), phosphate (PO_4^{3-}), total dissolved phosphorus (TDP), dissolved organic carbon (DOC), silicate (Si), total suspended solids (TSS), chlorophyll *a* (Chl-*a*), phaeophytin *a* (Phaeo), particulate organic carbon (POC), particulate nitrogen (PN), and particulate phosphorus (PP). Samples for dissolved nutrient (NH_4^+, NO_2^+, NO_3^+, PO_4^{3-} – hereinafter used without specifying the electron charge for clarity, as well as Si, TDN, TDP, and DOC) analysis were immediately filtered through a 0.45 µm syringe filter (Sartorius Minisart N) into

10 mL acid-washed vials, which were pre-rinsed three times with filtered site seawater. Dissolved inorganic (NH_4, NO_2, NO_3, PO_4) and total dissolved (TDN, TDP) samples were stored frozen (-18° C) until analysis. Samples for DOC analysis were acidified with 100 µL of AR-grade hydrochloric acid; DOC and Si samples were stored refrigerated (4° C) until analysis. Samples for particulate nutrient (POC, PN, PP) and Chl-*a* analysis were manifold filtered through pre-combusted (450° C for 4 h) 25 mm diameter filters (Whatman GF/F, nominal pore size 0.7 µm), folded, placed in pre-combusted aluminium foil envelopes, and stored frozen (-18° C) until analysis. Samples for TSS analysis were manifold filtered onto pre-weighed 47 mm diameter polycarbonate filters (GE Water & Process Technologies, pore size 0.4 µm), which were then triple rinsed with ultrapure water to remove residual salt from the filter. TSS filters were stored at room temperature while onboard the vessel and were immediately placed in a drying oven (60° C) overnight upon return to AIMS (Townsville, Queensland).

In addition to the water chemistry variables listed above, temperature, salinity, and Chl-*a* fluorescence measurements were also retrieved from the underway sampling systems on the RV Solander and RV Cape Ferguson, which are part of Australia's Integrated Marine Observing System (IMOS) Ships of Opportunity Sensors on Tropical Research Vessels sub-facility[157]. Temperature and salinity data were measured at 10 sec intervals using a SBE 38 digital oceanographic thermometer and SBE 21 SeaCAT thermosalinograph (Sea-Bird Scientific), while fluorescence was measured using an ECO FLNTU-RT (WET Labs). Intake depths for underway systems were 1.9 m (RV Cape Ferguson) and 2.5 m (RV Solander). For temperature, salinity, and Chl-*a* fluorescence, a single value that was closest to the sampling time was recorded at each site. Hereinafter, we use the term "physico-chemical variables" to encompass the 17 variables measured in this study, which include 14 water chemistry variables, as well as temperature, salinity, and Chl-*a* fluorescence.

Seawater for metagenomic sequencing was collected concurrently with water chemistry samples in four 5 L replicates. Collected seawater was immediately passed through a 5 µm Minisart® NML syringe pre-filter (Sartorius, Goettingen, Germany) to remove large debris and eukaryotic cells, and subsequently through a 0.22 µm Millipore® Sterivex-GP™ Pressure Filter (Merck Millipore, Darmstadt, Germany) using a peristaltic pump on board the research vessel. The Sterivex filters were snap-frozen in liquid nitrogen and stored at -75 ℃ until processed in the laboratory.

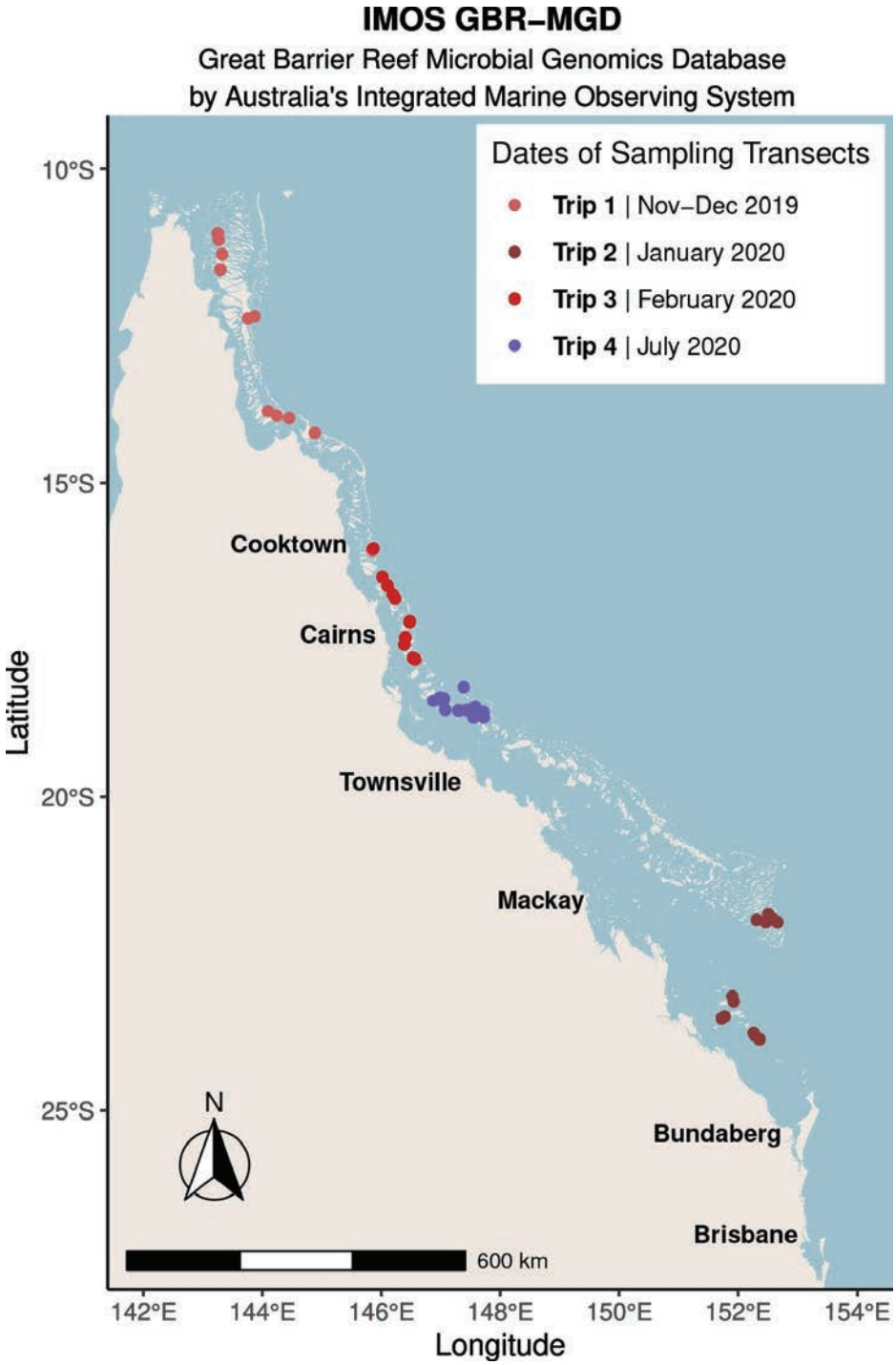

Figure 2.1: Field sampling design for the GBR-MGD (Great Barrier Reef Microbial Genomics Database) dataset by Australia's Integrated Marine Observing System (IMOS). Seawater was collected from 48 offshore GBR reef sites for microbial community metagenomic sequencing and analysis of 17 physico-chemical variables over 4 trips between November 2019 and July 2020. Reef sites are colored in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. Trip-specific maps with individual reef names are provided in Supplementary Material (Appendix A, Figures S1–S4).

2.3.2 Laboratory processing for water chemistry and metagenomic sequencing

Laboratory analyses of water chemistry samples were conducted at the AIMS Analytical Technology and Water Quality Laboratories within one month (Chl-*a*, DOC, and TSS) or three months (all other variables) of collection. Inorganic dissolved nutrient concentrations (NH_4, NO_2, NO_3, PO_4, Si) were determined using standard wet chemical methods[158–160] on a Seal AA3 segmented flow analyser. Total dissolved samples (TDN, TDP) were persulfate digested[161] and analysed for inorganic concentrations as above. Concentrations of DOC, POC, and PN were determined via high temperature catalytic combustion using a Shimadzu TOC-L carbon analyser with a solid sample module (SSM-5000A) for POC filters and a nitrogen module (TNM-L) for PN filters. Concentration of PP was determined spectrophotometrically[159] following digestion in hot acid persulfate[162]. Concentration of Chl-*a* was determined by grinding filters in 90% acetone (with a 2-hour incubation period in the dark) and reading the supernatant on a fluorometer (Turner Designs 10AU); samples were then acidified and re-read to determine the concentration of Phaeo and correct Chl-*a* measurements for its interference[163]. Concentrations of TSS were determined gravimetrically based on pre- and post-sampling filter weights.

DNA was extracted from 0.22 µm Sterivex filters using a phenol:chloroform:Iso-amyl alcohol extraction with ethanol precipitation (as in Botté et al. (2019)[164]; with the addition of 18 µL (100 mg mL^{-1}) lysozyme to the lysis buffer). DNA was quality-checked with a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Australia) and quantified using a Qubit 3 fluorometer (Thermo Fisher Scientific, Australia) before submission for Illumina Nextera FLEX sequencing using the NovaSeq at Microba Life Sciences Ltd. (Brisbane, QLD, Australia). An average of 17,464,769 ± 4,075,366 of 150 bp reads was sequenced from each of the 191 samples (47 sites x four replicates at each site, and three replicates at Hedley reef) (**Appendix A: Table S1**). The three negative controls had a low number of sequenced reads (173,749 ± 49,755; **Appendix A: Table S1**).

2.3.3 Metagenomic data processing

A read-based metagenomics analysis was applied to separate taxonomic and functional profiling of seawater microbiomes in offshore GBR reefs and elucidate the role of environmental filtering and functional redundancy in shaping reef bacterioplankton communities separately at taxonomic and functional levels[132,133]. Demultiplexed raw reads were first quality-checked in FastQC[163] (version 0.11.3) and quality-filtered in Trimmomatic[164] (version 0.38) to trim barcodes/adapters and remove low-quality bases (Phred <20). In total, 78.84% of reads were retained after quality filtering in Trimmomatic (an average of 13,853,993 ± 3,324,976 reads per sample) (**Appendix A: Table S1**). Quality-filtered reads were then aligned against the NCBI nr database using the DIAMOND (version 2.0.9) alignment tool[165]. For each read, the top match reported by DIAMOND with e-value of $<10^{-5}$ was retained to exclude poor annotations. Resulting DIAMOND files (in daa format) were then imported into MEGAN[166] (version 6.23.0) for community profiling, which assigned taxonomy using the NCBI reference system (see **Appendix D** for

comparison with GTDB nomenclature). Raw microbial abundance counts were exported from MEGAN for genus-level taxonomic and functional (GO-terms) composition, and subsequently imported into R Studio[146] (version 4.3.2) using the phyloseq R package[166]. Using R, further filtering steps included the removal of (1) non-annotated reads; taxa annotated as (2) eukaryotic (774 hits) or (3) viral (35 hits); and removal of (4) prokaryotic reads annotated to the Domain level only (Bacteria or Archaea), leaving 48% of the total data set. The last filtering step included the removal of (5) rare/spurious reads (relative abundance < 0.0001%), resulting in a total of 621 of the initial 1257 prokaryotic taxa (collapsed at genus level or above) for the final dataset on microbial taxonomy, while for gene annotation dataset, this filtering resulted in 4287 GO terms. This gave a final range of sequences of 3,752,207 ± 1,402,666 per sample (**Appendix A: Table S2**). Microbial abundance data was then Center-Log-Ratio (CLR) normalised in the microbiome R package[167] to account for sparsity and compositional nature of microbial metagenomic sequencing data. Pseudo counts were introduced prior to CLR normalisation as log 0 is undefined. These CLR-transformed counts or relative abundance data were used in downstream statistical analysis and visualisation in R Studio. Final composite plots were made in Inkscape 0.92.5.

2.3.4 Summarising water chemistry data and microbial community data

Principal Components Analysis (PCA) was applied in the R package mixOmics[4] as an unsupervised approach to visualise the main clustering patterns between reef sites based on physico-chemical variables. The number of optimal PCA components was determined using the mixOmics *tune.pca()* function. The PCA biplot was complemented with a heatmap to visualise the level of change in physico-chemical variables in more detail - across each reef site - by centering (median = 0) and scaling (standard deviation (SD) = 1) each of the 17 physico-chemical variables across sites.

PCA was used to visualise the main clustering patterns of reef sites based on seawater microbial communities (both for microbial taxonomy and GO terms, using CLR-normalised counts to account for compositionality and sparsity of metagenomics sequencing data), following the same approach as detailed in section above. Pairwise permutational multivariate analysis of variance (PERMANOVA) implemented in the *pairwise.adonis()* R wrapper function[168] was applied to test if distances between PCA group centroids (i.e. between the four trips) were statistically significant. Stacked bar charts were used to visualise microbial taxonomy profiles collapsed at (1) genus level (by showing the top 20 most abundant microbial genera), (2) at phylum level, and (3) at genus level but only within phylum *Bacteroidetes* which increased in relative abundances during summer. Microbial diversity was also compared between the four trips by computing a Shannon index (1) for the overall community profiles, and (2) only within phylum *Bacteroidetes*. Shannon diversity results were visualised as boxplots, and the variation in alpha diversity scores across trips was compared with pairwise Wilcoxon Rank-sum tests in R, which were integrated within microbial diversity boxplots.

2.3.5 Integrating microbial and physico-chemical data

Partial (geographic distance-corrected) Mantel tests with 10,000 permutations and Bonferroni correction were applied to identify physico-chemical variables that significantly correlated with seawater microbial communities[133,169]. In the partial Mantel tests, Bray-Curtis dissimilarities were computed within partial Mantel tests from relative abundances of microbial data with Euclidean distances of physico-chemical variables, while controlling for the effect of geography by including a third distance matrix of spatial distances between reef sites, expressed in km. A total of 34 partial Mantel tests were computed for both the taxonomy and functional genes datasets with each of the 17 physico-chemical variables.

Indicator microbes and GO terms were identified for each of the 17 physico-chemical variables using MINT sPLS - Multivariate INTegration Sparse Partial Least Squares[3,4,144,145]. sPLS[144,145] fits a linear relationship between multiple predictors (physico-chemical variables) with multiple continuous responses (microbial taxa or GO terms), while MINT is based on multi-group PLS that includes information about samples belonging to independent subsets of samples[3]. In this context, MINT sPLS integrated samples from independent subsets to remove unwanted sources of variation due to trips (i.e. confounding effects between season and geography), identifying microbial indicator taxa and GO terms that are shared/universal across the sampling trips. Prior to correlating metagenomic and physico-chemical data in MINT sPLS, median values per reef site were computed for each of the 17 physico-chemical variables as the number of Niskin deployments differed for molecular (four replicates) and water chemistry (three replicates) sampling. MINT sPLS selected 100 key features (i.e. seawater microbial taxa and GO terms; spanned across the first two MINT sPLS dimensions, with 50 features per dimension) that show the highest covariance with the 17 physico-chemical variables. MINT sPLS partial correlations were visualised as heatmaps for indicator taxa and GO terms using mixOmics[4].

Leave-One-Group-Out Cross-Validation (LOGOCV)[3] was applied to investigate the stability of microbial indicator taxa/GO terms identified in MINT sPLS dimension 1 across sampling trips. LOGOCV performed cross validation (CV) where one CV fold equals one study (sampling trip), hence four times until each of the four sampling trips was left out once. Indicator taxa/GO terms shared across different sampling trips were assigned stability scores of either 1 (selected in each of the four LOGOCV iterations), 0.75 (selected in 3/4 of the LOGOCV iterations), or 0.5 (selected in 2/4 of the LOGOCV iterations). A stability score of 0.25 indicates trip-specific microbiome/environment associations being identified in 1/4 LOGOCV iterations, hence these indicators were considered unstable (i.e. not shared across sampling trips). These stability scores were integrated with MINT sPLS heatmaps as barplots, visualised in the ggplot2 R package[170].

2.3.6 Comparing the potential of microbial indicator taxa and genes to infer reef physico-chemical metrics

The Bray-Curtis Similarity Index (expressed as 1 - Bray-Curtis dissimilarity, computed with the *vegdist()* function in vegan[171] R package) was used to compare within-site similarity (0 = dissimilar, 1 = identical) of reef bacterioplankton communities at functional and taxonomic levels. Bray-Curtis Similarity scores were computed within each of the 48 reefs and at various hierarchical levels, both for microbial taxonomy (genus, family, order, class, phylum) and functions (GO terms collapsed at levels 5, 4, and 3). For each of these levels, Bray-Curtis similarity scores (0 - low similarity; 1 - high similarity) were visualised as boxplots, with the higher similarity scores being indicative of the lower community variability in the microbiome composition within one reef site.

To identify if microbial indicator taxa or GO terms associate more robustly with physico-chemical variables in the surrounding reef, we used the same principles presented for MINT sPLS (i.e. inferring indicator stability using LOGOCV), but instead of removing one group during LOGOCV iterations (samples belonging to one trip), within each CV iteration, a random subset of samples from each trip was removed as a single subset of data. In more detail, sPLS was applied within each of the trips to account for confounding effects of geography and time, with microbial taxa and GO terms selected as predictor datasets, and physico-chemical variables as the response dataset. This resulted in a total of eight sPLS models (four trips x two datasets, for microbial taxa and GO terms). For each of the eight sPLS models, a 4-fold CV with 50 repeats was applied to assess reproducibility of the microbiome/environment signatures when the training set was subsampled via cross-validation, and each of the 50 indicator taxa/GO terms selected by sPLS on component 1 were assigned a stability score averaged across the 200 CV runs (4-fold CV x 50 iterations), ranging from 0 (i.e. low stability) to 1 (i.e. high stability). These stability scores were visualised as boxplots, and the variation between stability scores from indicator taxa and GO terms (within each of the four sampling trips) was tested with a Wilcoxon Rank-sum test in R, which were integrated within stability boxplots.

2.4 Results

2.4.1 Higher nutrients in GBR surface waters during summer

To identify drivers of microbial community variation for reef bacterioplankton (**Fig. 2.2, Table 2.1**), a total of 17 physico-chemical variables were derived from seawater samples from 48 offshore reefs across the length of the GBR (**Fig. 2.1**), including temperature, salinity, fluorescence, and particulate and dissolved nutrients. The largest source of variation in reef water chemistry was the timing of sampling trips across the austral summer or winter periods (41% of explained variance, PCA dimension 1), with samples collected in the peak of summer (Trip 3 - February 2020; SST = 30.16 ± 0.39 ℃) additionally separating from early summer sampling in trip 1 (Trip 1 - Nov-Dec 2019; SST = 27.78 ± 0.43 ℃ and Trip 2 -

January 2020; SST = 27.16 ± 0.61 ℃; 18% of explained variance, PCA dimension 2) (**Fig. 2.2a, Table 2.1, Appendix A: Fig. S5**). Overlaying physico-chemical data in a PCA visualisation showed that summer trips 1-3 were characterised by elevated temperature (median of 28.30 ± 1.51 ℃ across summer trips 1-3 vs 24.4 ± 0.95 ℃ in winter trip 4), and higher concentrations of particulate nutrients which were on average 3-fold higher in comparison to the winter trip (PP = 0.06 ± 0.01 µM for summer trips 1-3 vs 0.02 ± 0.01 µM in winter trip 4, ~3.4-fold increase in summer; PN = 1.27 ± 0.05 µM vs 0.50 ± 0.10 µM, ~2.5-fold increase in summer; POC = 8.54 ± 1.25 µM vs 3.67 ± 1.00 µM, ~2.3-fold increase in summer) (**Fig. 2.2a-b, Table 2.1, Appendix A: Fig. S5**). Chlorophyll fluorescence, Chl-*a*, and Phaeo were highest at sites collected in the central GBR in February 2020 (fluorescence = 0.32 ± 0.05 µg L^{-1}; Chl-a = 0.23 ± 0.18 µg L^{-1}; and Phaeo = 0.36 ± 0.15 µg L^{-1}; **Fig. 2.2a-b, Table 2.1, Appendix A: Fig. S5**). In contrast, reefs sampled in the austral winter had a 2-fold increase in dissolved phosphorus (PO_4 = 0.09 ± 0.02 µM; and TDP = 0.26 ± 0.02 µM) in comparison to the summer trips 1-3 (PO_4 = 0.04 ± 0.01 µM; and TDP = 0.20 ± 0.03 µM) (**Fig. 2.2a-b, Table 2.1, Appendix A: Fig. S5**). Notably, chemistry profiles of samples collected in the early austral summer were comparable despite being >1500 km apart in the far north (Cape Grenville and Princess Charlotte bay sectors) and far south (Swains and Capricorn Bunker sectors) of the GBR, whereas samples collected during the peaks of austral summer and winter were the most distinct although they were geographically close in the central GBR (~200 km apart, Cairns and Cooktown / Lizard island sectors for austral summer samples, and Innisfail and Townsville sectors for austral winter samples). This highlights that water chemistry measurements in offshore GBR surface waters are predominantly driven by seasonality and less influenced by geography.

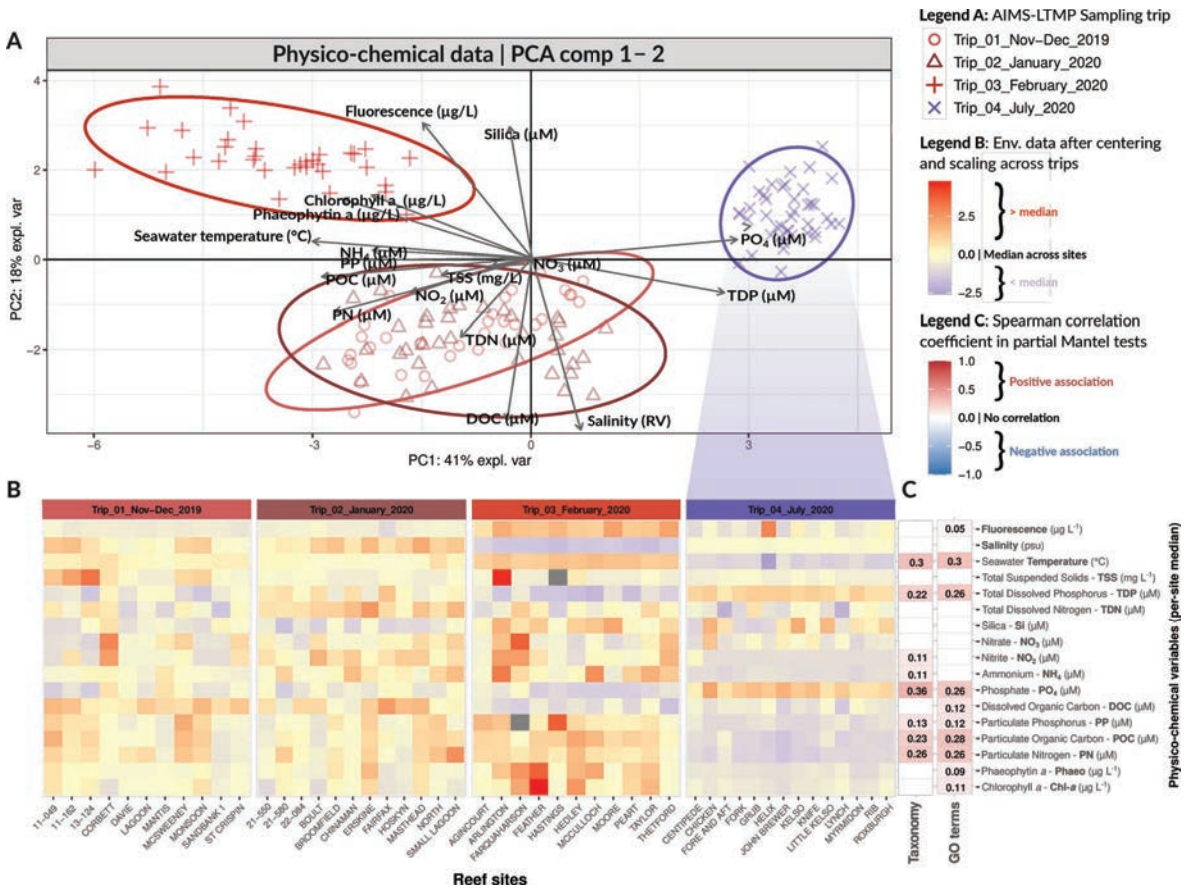

Figure 2.2: Summarising water chemistry data and identifying drivers of seawater microbial communities. (A) Principal Components Analysis (PCA) shows the main clustering patterns of reef sites based on physico-chemical variables. Reef sites use specific shapes and are coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. (B) The heatmap shows changes in physico-chemical variables (y axis) across the reef sites (x axis). Physico-chemical variables were centered (median = 0) and scaled (standard deviation (SD) = 1) across reef sites, and values that deviate from the median (0) were shown in red (> median) and blue (< median). (C) A total of 34 partial Mantel tests (corrected for geographic distance) were conducted for each of the 17 physico-chemical variables, and for both microbial datasets on taxonomy and GO terms. Non-significant results (p value > 0.05, Bonferroni correction) are shown as white cells, while coloured cells denote statistically significant trends (p value < 0.05, Bonferroni correction), indicating positive (red) or negative (blue) associations (Spearman's rank correlation coefficients ρ shown as the numeric value) between microbial and environmental distance matrices, while corrected for geographic distance between reefs (expressed in km).

Table 2.1: Physico-chemical data. Median ± SD values of 17 physico-chemical variables (rows) collected across 48 offshore GBR reefs. The values are collapsed across four sampling trips (columns).

Physico-chemical variables	Trip 1 (median ± SD)	Trip 2 (median ± SD)	Trip 3 (median ± SD)	Trip 4 (median ± SD)
Chlorophyll a (µg L^{-1}) — **Chl-*a***	0.18 ± 0.06	0.16 ± 0.08	0.32 ± 0.18	0.11 ± 0.03
Phaeophytin (µg L^{-1}) — **Phaeo**	0.18 ± 0.04	0.20 ± 0.08	0.36 ± 0.15	0.10 ± 0.02

Particulate nitrogen (μM) — **PN**	1.23 ± 0.35	1.27 ± 0.46	1.32 ± 0.22	0.50 ± 0.10
Particulate organic carbon (μM) — **POC**	8.06 ± 2.86	7.60 ± 1.89	9.95 ± 2.29	3.66 ± 1.00
Particulate phosphorus (μM) — **PP**	0.05 ± 0.02	0.05 ± 0.02	0.07 ± 0.03	0.02 ± 0.01
Dissolved organic carbon (μM) — **DOC**	84.51 ± 5.99	81.92 ± 9.89	67.22 ± 4.60	69.30 ± 4.67
Phosphate (μM) — **PO_4**	0.05 ± 0.03	0.04 ± 0.02	0.02 ± 0.02	0.09 ± 0.02
Ammonium (μM) — **NH_4**	0.39 ± 0.16	0.58 ± 0.27	0.74 ± 0.44	0.12 ± 0.06
Nitrite (μM) — **NO_2**	0.03 ± 0.02	0.04 ± 0.01	0.04 ± 0.02	0.01 ± 0.01
Nitrate (μM) — **NO_3**	0.30 ± 0.25	0.33 ± 0.15	0.35 ± 0.31	0.23 ± 0.16
Silica (μM) — **Si**	1.41 ± 0.30	1.30 ± 0.44	2.10 ± 0.55	1.78 ± 0.65
Total dissolved nitrogen (μM) — **TDN**	5.47 ± 0.83	6.62 ± 0.82	5.64 ± 0.72	5.18 ± 0.75
Total dissolved phosphorus (μM) — **TDP**	0.20 ± 0.03	0.23 ± 0.04	0.16 ± 0.03	0.26 ± 0.02
Total suspended solids (mg L^{-1}) — **TSS**	0.48 ± 0.41	0.15 ± 0.15	0.36 ± 0.52	0.11 ± 0.10
Temperature (℃)	27.78 ± 0.43	27.13 ± 0.61	30.01 ± 0.39	24.22 ± 0.95
Salinity (psu)	35.35 ± 0.21	35.52 ± 0.17	34.71 ± 0.05	35.16 ± 0.04
Chl-a fluorescence (μg L^{-1})	0.10 ± 0.01	0.10 ± 0.02	0.34 ± 0.05	0.13 ± 0.12

2.4.2 Microbial community composition and functional gene profiles differ by season

A total of 29 bacterial and archaeal phyla were identified in the seawater communities of the 48 surveyed offshore GBR reefs, of which three dominant phyla, Cyanobacteria (average 68% relative abundance), Proteobacteria (26%) and Bacteroidetes (2.6%) together comprised an average of 96.6% relative abundance of retrieved sequences (**Fig. 2.3d**). At the genus level, three genera dominated the seawater communities: *Synechococcus* with 54.99% average relative abundance, *Pelagibacter* (also known as SAR11) at 15.89% relative abundance, and *Prochlorococcus* at 11.93% (**Fig. 2.3c**).

PCA showed that seawater microbial communities differed between seasons, with samples primarily clustered by collections during the austral summer (trips 1-3) vs winter (trip 4) with around 26% of total variance attributable to the first two principal components (**Fig. 2.3a**). PCA clustering was supported by pairwise PERMANOVA indicating that seawater community composition significantly differed when comparing summer vs winter trips ($p < 0.05$, Bonferroni correction) but not between summer trips (**Table 2.2**). These differences in community composition were primarily driven by increased relative abundances of *Prochlorococcus* during the winter trip (average 32.93% vs 3.23% relative abundance in summer trips) and decreased *Synechococcus* (average 37.02% in winter vs 62.38% in summer trips) (**Fig. 2.3c**). Several members within the Bacteroidetes phylum were also more dominant in the three summer trips (mean relative abundances for summer trips: Trip 1 = 2.13 %; Trip 2 = 1.80 %; Trip 3 = 5.31 %; and the winter Trip 4 = 1.14%, see **Fig. 2.3d, 2.3e**), particularly the family Flavobacteriaceae which were the most discriminatory taxa in samples collected during the peak of summer in February 2020 (**Fig. 2.3e, Appendix A: Fig. S7**). Apart from increasing in relative abundance, members of the Bacteroidetes

phylum were also more diverse during the summer trips (median Shannon Index for Trip 1 = 2.67 ± 0.41; Trip 2 = 2.64 ± 0.36; and Trip 3 = 2.58 ± 0.63) compared to the winter (Trip 4 : median Shannon Index 2.24 ± 0.25) (**Fig. 2.3e, Fig. 2.3f**), with pairwise Wilcoxon rank sum tests only being significant (p adjusted <0.05) for summer/winter trip comparisons (**Fig 2.3f**). When comparing the Shannon Index computed for overall microbial communities, we identified no significant difference (p adjusted >0.05, Wilcoxon rank sum test) in median Shannon diversity between trips (**Appendix A: Fig. S6, Table S3, Table S4**).

Microbial functional profiles (GO terms) were also primarily clustered by sampling during the austral summer (trips 1-3) vs winter (trip 4), although with stronger separation compared with taxonomic composition (54% of variance attributable to the first two PCA dimensions vs. 26%) (**Fig. 2.3b**). Seawater microbial communities collected during summer trips 1-3 were characterised by elevated transporters (i.e. ABC transporters, TRAP transporter permease proteins, and UAA transporters, as well as various ion transporters) and GO terms encoding for oxidative phosphorylation (NADH:ubiquinone oxidoreductase, complex 1 of the respiratory chain), which were comparatively underrepresented in samples collected in the winter trip 4 (**Appendix A: Fig. S8**).

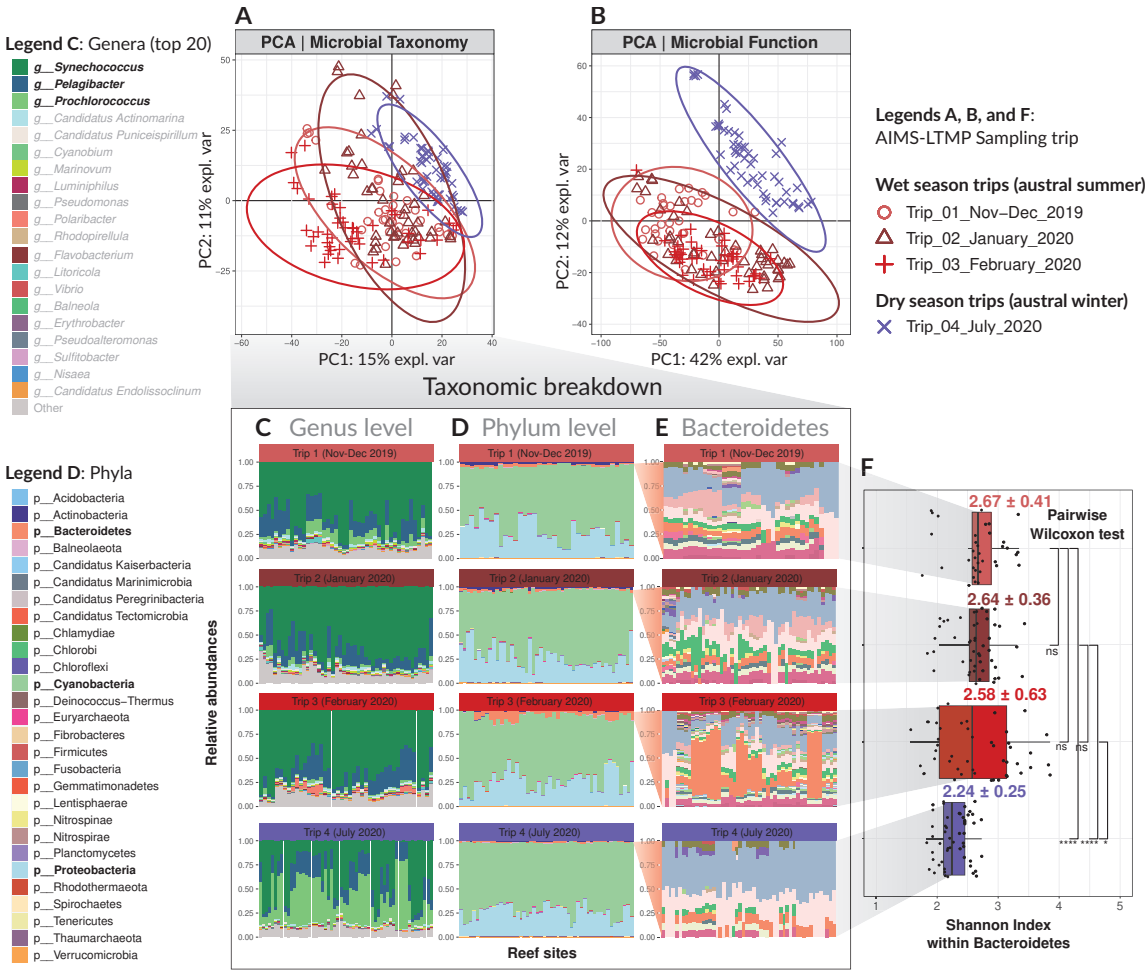

Figure 2.3: Main clustering patterns of seawater microbial communities. Principal Components Analysis (PCA) plots show the main clustering patterns of reef sites based on microbial community structure, both for microbial taxonomy (A) and microbial GO terms (B). Reef sites are coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively. Stacked bar plots illustrate microbial relative abundances (y-axis) for each sample (x-axis), with reef sites grouped by their corresponding sampling trip. These plots represent: (C) the top 20 most abundant microbial genera, (D) all 29 identified microbial phyla, and (E) all microbial genera within the phylum Bacteroidetes. The top three most abundant genera (C) and phyla (D) are highlighted in bold and the legend for genera within Bacteroidetes (E) was excluded due to the large number of taxa. (F) Boxplots illustrate microbial diversity (Shannon Index) for genera within phylum Bacteroidetes, across sampling trips. The symbols *, **, ***, and **** denote levels of statistical significance in pairwise Wilcoxon rank sum tests when testing variation of Bacteroidetes Shannon diversity scores across the four sampling trips: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$, indicating increasing levels of significance. 'ns' indicates non-significant results, where $p \geq 0.05$.

Table 2.2: The pairwise permutational multivariate analysis of variance (PERMANOVA) test for microbial communities (taxonomic level). Significant results (p-value < 0.05, Bonferroni correction) are highlighted in bold.

Pairwise comparison	SumSqrs	MeanSqrs	F	R^2	p-value	P-value (Bonferroni corrected)
Trip 1-Trip 2	0.111	0.111	3.168	0.037	0.057	0.344

Trip 1-Trip 3	0.203	0.203	5.765	0.066	0.009	0.055
Trip 1-Trip 4	1.856	1.856	28.370	0.248	0.000	0.001
Trip 2-Trip 3	0.091	0.091	2.688	0.028	0.076	0.453
Trip 2-Trip 4	2.933	2.933	48.630	0.332	0.000	0.001
Trip 3-Trip 4	3.210	3.210	52.889	0.353	0.000	0.001

2.4.3 Particulate and dissolved nutrients drive seawater microbial community composition

Partial Mantel tests identified nine- and 11-physico-chemical variables which were associated with taxonomic composition and gene-based microbial profiles respectively, while accounting for geographic distance between reefs ($p < 0.05$, Bonferroni correction; **Fig. 2.2c**). The highest Spearman's rank correlation coefficients (ρ, ranging from -1 to 1, with negative values indicating an inverse relationship, and positive values denoting the same trajectory), and therefore the strongest physico-chemical variables influencing both taxonomic composition and functional profiles were computed for phosphate (ρ = 0.36 for microbial taxonomy; and ρ = 0.26 for microbial functional genes), seawater temperature (ρ = 0.3 and 0.3), and particulate nutrients (POC: ρ = 0.23 and 0.28; PN: ρ = 0.26 and 0.26; PP: ρ = 0.13 and 0.12) (**Fig. 2.2c**). Physico-chemical variables that were significantly associated only to microbial genes but not microbial taxonomy included Chl-*a* (ρ = 0.11), Phaeo (ρ = 0.09), fluorescence (ρ = 0.05), and DOC (ρ = 0.12). In contrast, NH_4 (ρ = 0.11) and NO_2 (ρ = 0.11) positively associated only to microbial taxonomy, but not functional gene profiles (**Fig. 2.2c**). Only positive Spearman correlations were calculated for the physico-chemical variables significantly associated with reef bacterioplankton, indicating that both taxonomic and functional composition of seawater microbes become increasingly dissimilar as associated physico-chemical variables change. This suggests that seawater microbes exhibit a deterministic response to their surrounding environment, with microbial population dynamics or community structure being directly influenced by specific nutrient conditions and changing in proportion to variations in measured nutrients.

Using Multivariate INTegration Sparse Partial Least Squares (MINT sPLS) to identify which indicator microbial taxa and GO terms consistently associated with the same physico-chemical variables in more than one sampling trip, we selected 100 key indicator seawater microbial taxa and GO terms (spanned across the first two MINT sPLS dimensions, with 50 features per dimension) that show the highest associations with 17 physico-chemical variables stably across trips. Since low MINT sPLS correlation scores (i.e. below the absolute value of 0.22) were observed for the 50 indicator microbial taxa and genes selected on the second MINT sPLS dimension, a Leave-One-Group-Out Cross-Validation (LOGOCV) was applied to mine for stable indicators only on MINT sPLS dimension 1, ultimately identifying 33 microbes and 34 GO terms that are shared across 2, 3, or 4 trips (i.e. indicators assigned LOGOCV stability scores of 0.5, 0.75, and 1, respectively). All 100 indicator features (microbes in **Fig. 2.4**,

and GO terms in **Fig. 2.5**) were then grouped into three "community type" clusters based on Euclidean distance clustering (marked with dashed lines), and the clusters containing the 33 microbes and 34 GO terms as stable indicators were termed "Cluster 1" to highlight their importance, and were the main focus in results interpretation and discussion. Microbial indicators in MINT sPLS clusters 2 (34 indicator taxa in **Fig. 2.4, Cl. 2**, and 37 indicator GO terms in **Fig. 2.5, Cl. 2**) and 3 (13 indicator taxa in **Fig. 2.4, Cl. 3**, and 12 indicator GO terms in **Fig. 2.5, Cl. 3**) were not considered in downstream discussion.

The 33 stable taxonomic indicators from cluster 1 collectively showed positive associations with particulate nutrients (Median ± SD of MINT sPLS positive partial correlation scores for POC: 0.44 ± 0.04; PN: 0.41 ± 0.03; and PP: 0.34 ± 0.02), Chl-*a* (0.39 ± 0.03), and DOC (0.35 ± 0.02), and negative associations with dissolved inorganic nutrients (Median ± SD of MINT sPLS negative partial correlation scores for NO_3: -0.50 ± 0.03; NO_2: -0.34 ± 0.04; NH_4: -0.26 ± 0.03; PO_4: -0.46 ± 0.03; and TDP: -0.27 ± 0.04) (**Fig. 2.4a, Cl. 1**). These microbial indicators were consistent across either three trips (LOGOCV stability score = 0.75) for 17 taxa, including members of Synechococcales, two Rhodobacteraceae, and Rhodospirillaceae, or two trips (LOGOCV stability score = 0.5) for 16 taxa, including two Oceanospirillaceae, two Rhodospirillaceae, and two Burkholderiaceae (**Fig. 2.4b, Cl. 1**). The second cluster contained 34 taxa, and while also largely composed of Alphaproteobacteria (four Rhodospirillaceae), Gammaproteobacteria (five Cellvibrionales) and Deltaproteobacteria similar to the first cluster, these microbes were up to a 5-fold less strongly associated with particulate nutrients (POC: 0.09 ± 0.06; PN: 0.10 ± 0.06; and PP: 0.09 ± 0.05) and phosphorus (PO_4: -0.11 ± 0.07) compared with the first cluster, but still show positive associations with DOC (0.21 ± 0.07) (**Fig. 2.4a, Cl. 2**). The third cluster (13 taxa), predominantly composed of Flavobacteriaceae (12 taxa), showed two distinct subgroups. Both were positively associated with dissolved nitrogen (NH_4: 0.12 ± 0.04; NO_2: 0.19 ± 0.06; and NO_3: 0.14 ± 0.09), but one cluster (**Fig. 2.4a, Cl. 3a**) positively associated with particulate nutrients (POC: 0.17 ± 0.07; PN: 0.14 ± 0.06; and PP: 0.10 ± 0.05) and negatively associated with dissolved phosphorus (PO_4: -0.15 ± 0.07), and the other cluster (**Fig. 2.4a, Cl. 3b**) negatively associated with particulate nutrients (POC: -0.08 ± 0.04; PN: -0.08 ± 0.04; and PP: -0.06 ± 0.06) and positively associated with dissolved phosphorus (PO_4: 0.10 ± 0.04). Overall, these patterns indicate that particulate nutrients, dissolved N and P were the main physico-chemical variables driving partitioning of seawater microbial communities in the surveyed offshore reefs. Most of the indicator taxa were positively associated with particulate nutrients and negatively associated with dissolved N and P, with the exception of several genera in the Flavobacteriaceae family that were positively associated with both particulate nutrients and dissolved N and P (**Fig. 2.4**).

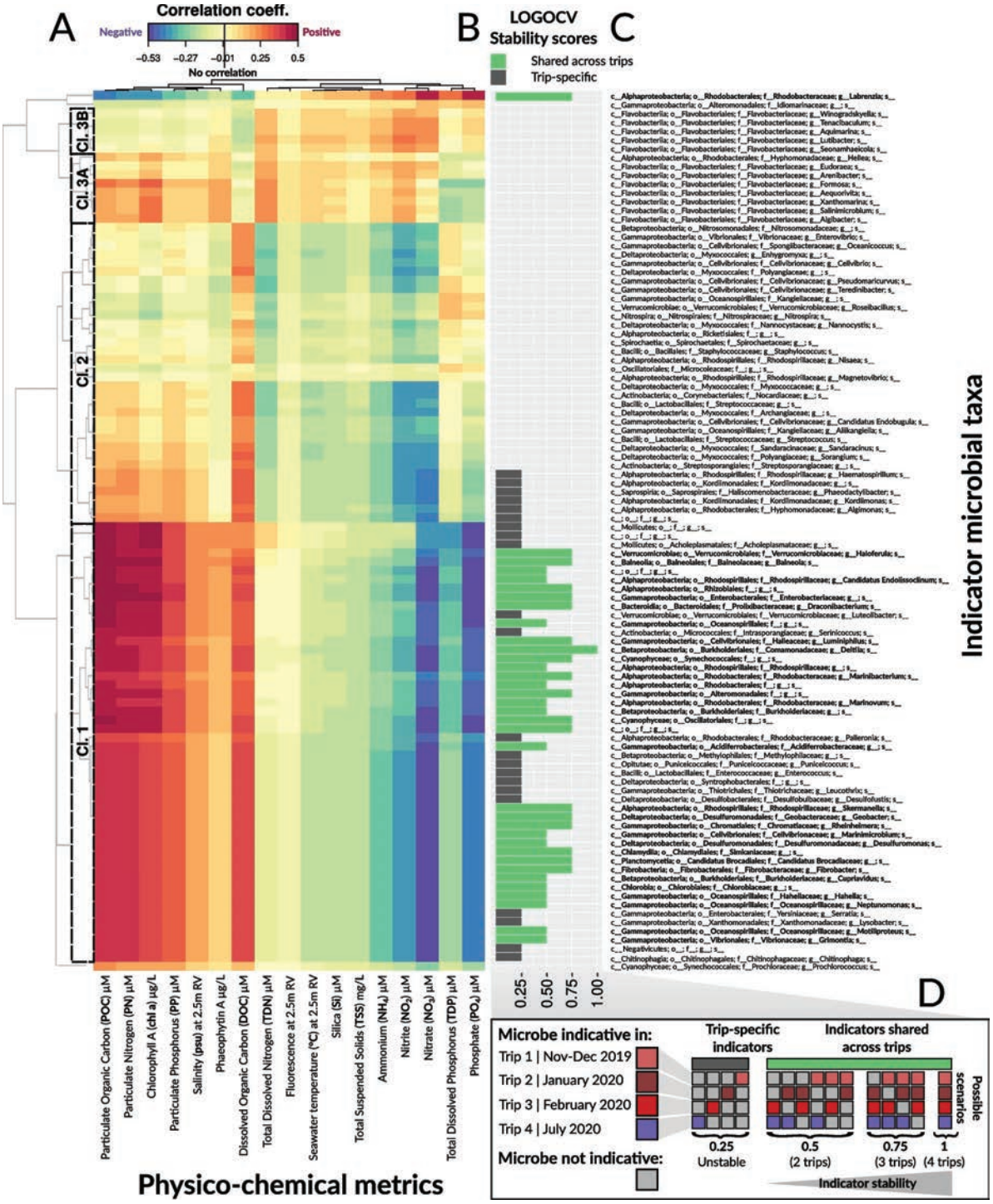

Figure 2.4: MINT sPLS - Associations between microbial taxa and physico-chemical variables. (A) The heatmap shows similarity values (partial correlations) between 17 continuous physico-chemical variables (predictor dataset) and 100 microbial taxa (response dataset) selected across the first two MINT sPLS dimensions. Heatmap cells are coloured to indicate either positive (red) or negative (blue) correlation. Heatmap rows and columns were clustered with a complete Euclidean distance method, with three clusters highlighted with a dashed line and numbered as they were discussed in the text. (B) Indicator stability barplots as determined by Leave-One-Group-Out Cross-Validation - LOGOCV. Microbial indicator taxa are colored in green if they are shared across sampling trips, or in grey if they are trip-specific. (C) Taxonomic breakdown of indicator microbes, with indicator taxa shared across

different sampling trips (as inferred by LOGOCV) further highlighted in bold. (D) Explanation of LOGOCV stability scores through 15 possible scenarios. Indicator microbes were assigned colours if indicative in a particular trip (with colouring scheme for trips corresponding to Fig. 2.1), while non-indicator taxa are colored in grey (D, left). The lowest LOGOCV stability score of 0.25 indicates trip-specific microbial indicators (selected in 1/4 LOGOCV iterations, with four possible scenarios), which were therefore considered unstable as these indicators are not reproducible across sampling trips (D, middle). Stable microbial indicators (shared across trips) were assigned LOGOCV stability scores of either 0.5 (selected in 2/4 of the LOGOCV iterations, with six possible scenarios), 0.75 (selected in 3/4 of the LOGOCV iterations, with four possible scenarios), or 1, which indicated the highest indicator stability score (selected in each of the four LOGOCV iterations) (D, right). Only shared microbial indicator taxa (with LOGOCV stability scores of 0.5, 0.75, and 1) were considered in downstream interpretation and discussion of results.

The 34 microbial GO terms identified in MINT sPLS as stable indicators (i.e. reproducible across sampling trips) collectively showed positive associations with particulate nutrients (Median ± SD of MINT sPLS positive partial correlation scores for POC: 0.42 ± 0.05; PN: 0.35 ± 0.05; and PP: 0.24 ± 0.07), Chl-*a* (0.29 ± 0.05), DOC (0.30 ± 0.06), and salinity (0.34 ± 0.08), and were negatively associated with dissolved nutrients (NH_4: -0.18 ± 0.05; NO_2: -0.27 ± 0.05; and NO_3: -0.29 ± 0.06; TDP: -0.24 ± 0.06; and PO_4: -0.33 ± 0.06, see **Fig. 2.5a, Cl. 1**). These stable indicator GO terms were involved in (1) transmembrane nutrient uptake, including permease proteins PstB - phosphate transport system permease protein (LOGOCV stability = 0.5) and PstC (LOGOCV stability = 0.5) as subunits of a Pst system for phosphate transport; ion transmembrane transport - Na^+/H^+ antiporter subunit G (LOGOCV stability = 0.75); and assimilation of external ammonium – alanine dehydrogenase (LOGOCV stability = 1); (2) utilisation of N-acetylglucosamine (N–acetylglucosamine–6–phosphate deacetylase, LOGOCV stability = 1); (3) oxidative phosphorylation, such as chain I of the NADH–quinone oxidoreductase (LOGOCV stability = 1); as well as synthesis of (4) fatty acids - enoyl-acyl carrier protein reductase (NADH) (LOGOCV stability = 1); and (5) vitamins - pyridoxal kinase for biosynthesis of pyridoxal phosphate, an active form of vitamin B6 (LOGOCV stability = 0.5) (**Fig. 2.5, Cl. 1**). The second cluster (**Fig. 2.5a, Cl. 2**) consisted of 37 GO terms positively associated with Phaeo, salinity, PP, and dissolved nitrogen variables, and negatively associated with dissolved phosphorus and DOC (**Fig. 2.5**), while the third cluster (**Fig. 2.5a, Cl. 3**) consisted of 12 GO terms only positively associated with dissolved phosphorus (TDP) (**Fig. 2.5**). Collectively, the 34 GO terms identified as stable indicators were implicated in processes including nutrient uptake, ion transport, ammonium assimilation, oxidative phosphorylation, and synthesis of fatty acids and vitamins.


A
Correlation coeff.
Negative
Positive
−0.39
−0.16
0.06
0.29
0.52
No correlation
B
LOGOCV
Stability scores
Shared across trips
Trip-specific
C
Cl. 2
Cl. 1
Cl. 3
Indicator microbial genes (GO terms)
Dissolved Organic Carbon (DOC) µM
Total Dissolved Nitrogen (TDN) µM
Silica (Si) µM
Fluorescence at 2.5m RV
Phaeophytin A µg/L
Seawater temperature (°C) at 2.5m RV
Total Suspended Solids (TSS) mg/L
Particulate Nitrogen (PN) µM
Chlorophyll A (chl a) µg/L
Particulate Organic Carbon (POC) µM
Salinity (psu) at 2.5m RV
Particulate Phosphorus (PP) µM
Total Dissolved Phosphorus (TDP) µM
Phosphate (PO4) µM
Nitrate (NO3) µM
Ammonium (NH4) µM
Nitrite (NO2) µM
0.25
0.50
0.75
1.00
Physico-chemical metrics
D
Microbe indicative in:
Trip 1 | Nov-Dec 2019
Trip 2 | January 2020
Trip 3 | February 2020
Trip 4 | July 2020
Microbe not indicative:
Trip-specific indicators
Indicators shared across trips
Possible scenarios
0.25
Unstable
0.5
(2 trips)
0.75
(3 trips)
1
(4 trips)
Indicator stability

Figure 2.5: MINT sPLS - Associations between microbial genes/functions (GO terms) and physico-chemical variables. (A) The heatmap shows similarity values (partial correlations) between 17 continuous physico-chemical variables (predictor dataset) and 100 microbial GO terms (response dataset) selected across the first two MINT sPLS dimensions. Heatmap cells are coloured to indicate either positive (red) or negative (blue) correlation. Heatmap rows and columns were clustered with a complete Euclidean distance method, with three clusters highlighted with a dashed line and numbered as they were discussed in the text. (B) Indicator stability barplots as determined by Leave-One-Group-Out Cross-Validation - LOGOCV. Microbial indicator genes are colored in green if they are shared across sampling trips, or in grey if they are trip-specific. (C) GO functional annotation of indicator genes/functions, with indicator GO terms shared across different sampling trips (as inferred by LOGOCV) further highlighted in bold. (D) Explanation of LOGOCV stability scores through 15 possible scenarios. Indicator genes were assigned colours if indicative in a particular trip (with colouring scheme for trips corresponding to Fig. 2.1), while non-indicator genes are colored in grey (D, left). The lowest LOGOCV stability score of 0.25 indicates trip-specific microbial indicators (selected in 1/4 LOGOCV iterations, with four possible scenarios), which were therefore considered unstable as these indicators are not reproducible across sampling trips (D, middle). Stable microbial indicators (shared across trips) were assigned LOGOCV stability scores of either 0.5 (selected in 2/4 of the LOGOCV iterations, with six possible scenarios), 0.75 (selected in 3/4 of the LOGOCV iterations, with four possible scenarios), or 1, which indicated the highest indicator stability score (selected in each of the four LOGOCV iterations) (D, right). Only shared microbial indicator genes (with LOGOCV stability scores of 0.5, 0.75, and 1) were considered in downstream interpretation and discussion of results.

2.4.4 Microbial functional genes correlate more stably to physico-chemical variables than taxonomy

To test our hypothesis that reef-associated bacterioplankton, due to functional redundancy, would exhibit higher community similarity at the functional rather than taxonomic level within a single reef site (i.e. under similar environmental conditions), we computed the Bray-Curtis Similarity Index (as a metric of overall compositional similarity: 0 = dissimilar, 1 = identical) between four replicates within each of the 48 surveyed reefs. This resulted in a total of 288 reef-specific Bray-Curtis similarity values (six pairwise comparisons per reef × 48 reefs) for each hierarchical level tested: for microbial taxonomy (genus, family, order, class, and phylum) and genes (GO terms at Ranks 5, 4, and 3). The Bray-Curtis similarity scores for taxonomic communities showed consistent and high median values across different hierarchical levels. Specifically, the median ± SD Bray-Curtis similarity scores were: 0.9 ± 0.15 at the genus level, 0.9 ± 0.14 at the family level, 0.92 ± 0.09 at the order level (**Fig. 2.6a**, microbial taxonomy), 0.93 ± 0.09 at the class level, and 0.93 ± 0.08 at the phylum level (**Appendix A: Fig. S9**). These results indicate high within-site taxonomic similarity for most of the surveyed offshore reefs. The lowest observed similarity scores were 0.37 (genus-level) and 0.38 (family-level) indicating that replicates within some reefs can be dissimilar at the lower taxonomic levels, although minimum similarity remains higher at higher taxonomic levels (0.56 at order-level; and 0.57 at class- and phylum-level communities) (**Fig. 2.6a**, microbial taxonomy). Gene profiles for reef bacterioplankton communities showed comparable median similarity scores to taxonomic communities, although with lower SD (median ± SD Bray-Curtis similarity for GO terms at Rank 5: 0.90 ± 0.08; Rank 4: 0.95 ± 0.04; and Rank 3: 0.97 ± 0.02) and higher minimum similarity scores (0.57, 0.76, and 0.86 for GO terms collapsed at Ranks 5, 4, and 3, respectively) (**Fig. 2.6a**, microbial function). Overall,

replicates within a single reef site are similar both at taxonomic and functional gene levels, though this similarity is increased for functional traits.

To compare whether seawater indicator GO terms or indicator microbes have a higher stability to infer continuous physico-chemical variables in the outer GBR reefs, we generated eight sPLS models (computed for four trips x two datasets, for microbial taxa and GO terms) and perturbed them with a 4-fold cross-validation repeated 50 times, resulting in 200 independent CV runs for each sPLS model. In this we introduced a measure of statistical stability[3,172,173] calculated as the averaged re-occurrence of microbial indicators (taxa and GO terms, selected on sPLS dimension 1) across 200 sPLS CV runs, and the stability scores ranged from 0 (low indicator stability) to 1 (high stability). In each of the four trips, the same microbial genes/functions were more frequently re-selected as indicators of physico-chemical variables compared with microbial taxa, with stability scores for indicator GO terms consistently higher (median ± SD stability for the 50 indicator GO terms on sPLS dimension 1: Trip 1 = 0.74 ± 0.18; Trip 2 = 0.66 ± 0.30; Trip 3 = 0.47 ± 0.14; and Trip 4 = 0.71 ± 0.18) compared with indicator microbes (median ± SD stability for the 50 indicator microbes on sPLS dimension 1: Trip 1 = 0.66 ± 0.18; Trip 2 = 0.53 ± 0.24; Trip 3 = 0.10 ± 0.08; and Trip 4 = 0.63 ± 0.03) (**Fig 2.6b**, trips 1-4). Pairwise Wilcoxon rank sum tests confirm these trends were significant (p adjusted <0.05) for trips 1, 3, and 4, but the results were not significant in trip 2 (p adjusted >0.05, Wilcoxon rank sum test) (**Fig 2.6b**, trips 1-4). Overall, these results suggest that microbial genes/function is more robustly associated with physico-chemical variables compared to microbial taxonomy.

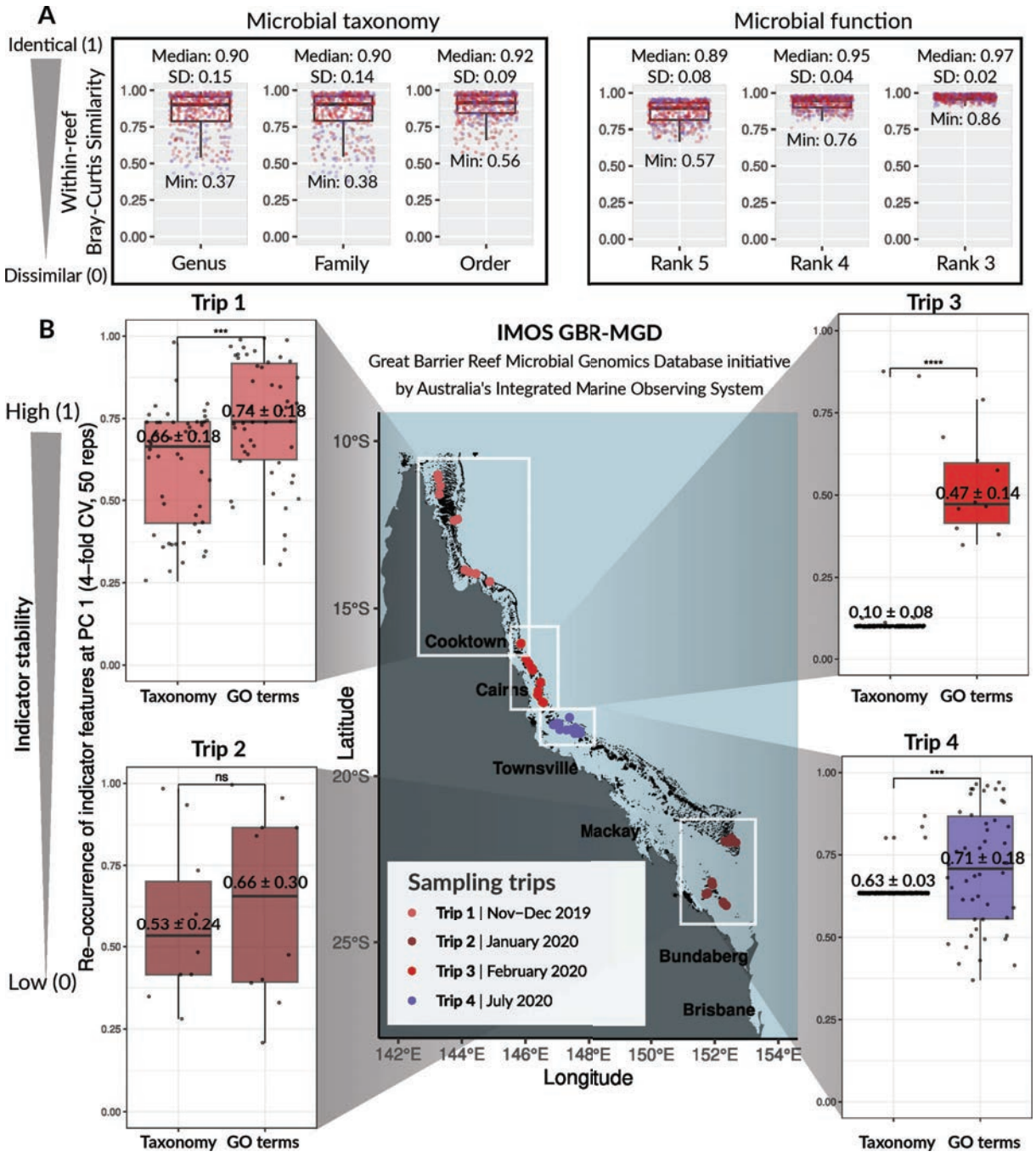

Figure 2.6: Differing diagnostic potential of microbial taxonomy and function to inform changes in continuous physico-chemical variables in the surrounding reef. (A) Bray-Curtis Similarity Index shows within-reef community similarity (0 = dissimilar; 1 = identical) for microbial taxonomy (at genus, family, and order-level classifications) and microbial functions (GO terms collapsed at Ranks 5, 4, and 3). (B) Comparison of how frequently indicator microbes and indicator genes (left and right boxplots, respectively) are re-selected across 200 independent sPLS cross validation runs (4-fold CV x 50 repeats), across all four sampling trips (Trips 1-4). Higher stability scores are a proxy of robustness of the indicator signal for a corresponding microbe/gene (i.e. the stability score of 1 would mean that the indicator microbe/gene was re-selected in sPLS on component 1 in each of the 200 CV runs). The symbols *, **, ***, and **** denote levels of statistical significance in pairwise Wilcoxon rank sum tests when testing variation between stability scores from indicator taxa and GO terms within each of the four sampling trips: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$, indicating increasing levels of significance. 'ns' indicates non-significant results, where $p \geq 0.05$.

2.5 Discussion

The composition of reef-associated bacterioplankton undergoes significant shifts in response to environmental stressors and poor reef health conditions (reviewed within[34,35,82,130]) Numerous opportunistic seawater microbes—such as Flavobacteriaceae, Rhodobacteraceae, and Vibrionaceae—that increase in abundance during disturbances, along with their functions (e.g., virulence factors, toxin production, antibiotic resistance), have been proposed as candidate indicators of poor reef health[43,55,92,143]. However, analysis efforts are lacking to evaluate if reef-associated seawater microbial taxa or genes/functions have a higher diagnostic potential in microbial monitoring, and to determine whether seawater biomarkers will consistently be indicative of a specific physico-chemical metric across broad spatio-temporal scales. By employing integrative omics approaches, specifically P-integration (sensu[3,4,173]) and introducing the measure of statistical stability (i.e. reoccurrence of microbial indicators across independent cross-validation runs) into microbiome-environment associations, here we identify microbial markers stably associated with nutrient concentration across reefs and season in offshore GBR surface waters. We also show that a greater proportion of variance in gene content was attributable to physico-chemical variables compared to taxonomic composition, with genes/environment associations being more than twice as stable.

2.5.1 Deriving seawater microbial indicators for GBR reef health monitoring

Functional redundancy proposes that environmental filtering primarily selects for functional traits in pelagic microbes[16,77,133,136]. Computing reef-specific Bray-Curtis similarity scores (at various levels for microbial taxonomy and GO terms) as a metric of overall community similarity, we show that across the surveyed reefs, reef-associated bacterioplankton exhibit higher community similarity at the functional rather than taxonomic level within a single reef site, where similar environmental conditions prevail. As the observed patterns may include core genes encoding for essential functions that are critical to life and thus shared across diverse taxa[174], we further explored the robustness of these findings by focusing only on the stability of indicator microbial taxa and GO terms associated with specific physico-chemical variables in the reef environment, using sPLS analysis complemented with cross-validation. The sPLS stability scores for indicator microbial genes/functions were approximately twice as high as those for microbial taxa consistently across different regions and seasons on the GBR, further highlighting that indicator gene targets offer greater precision in monitoring environmental metrics within reef ecosystems.

These observations are consistent with the concept of functional redundancy in pelagic microbial communities where multiple members of the community possess overlapping metabolic capabilities and are able to functionally replace one another[77,132,136,142]. For example, an analysis of N cycling seawater

microbial communities using data from the TARA Oceans expedition reported 30.1% of variance in the composition of functional traits statistically attributed to environmental measures compared with 16.3% of variance in taxonomic composition. Further, stochastic (i.e. random) processes had ~1.4-fold increase in relative importance of shaping the taxonomic compared to functional compositional variance, suggesting N-cycling microbial functions are more influenced by deterministic processes (i.e. environmental filtering) compared to taxonomy[133]. This explains why genes encoding for the same N-cycling pathways were consistently enriched in the epipelagic (N_2 fixation, organic decomposition, and assimilatory nitrite reduction to ammonia) and mesopelagic (nitrification, dissimilatory nitrate reduction to nitrite, and annamox) zones, whereas the taxonomic composition of N-cycling microbes between depth layers varied substantially, even at phylum level[133]. Taken together, the findings indicate that functional traits in seawater microbial communities are tightly linked to environmental measures and thus more likely to reflect the environment than taxonomy does. Functional redundancy may broadly contribute to ecosystem resilience against perturbations[16,132,175–177]. Since resilience is a key measure in ecosystem monitoring and management[178], we posit that gene content could conceivably serve as an indicator of ecosystem resilience, and changes in gene content coupled with contextual metadata could better reveal insight into the state of reef ecosystems compared with taxonomic indicators.

2.5.2 *Synechococcus* and *Prochlorococcus* are central to the production of particulate nutrients

The chemistry of GBR surface waters is characterised by a cluster of five collinear physico-chemical variables (Chl-*a*, Phaeo, POC, PN and PP) and a weak collinearity of dissolved nutrients (DOC, NH_4, NO_2, NO_3, PO_4, and Si) consistently elevated by 10-50% during summer, and with the lowest nutrient concentrations typically observed during winter and early spring in August-September[61,179,180]. Largely consistent with published data, we observed that temperature and nutrient concentrations were consistently higher in austral summer than in winter, apart from TDP and PO_4 which were higher in austral winter potentially due to seasonal upwelling of nutrient-rich water from the Coral Sea into reefs on the outer continental shelf, via intrusive upwelling events that are documented to occur in the central GBR[181,182]. We also observed collinearity between particulate nutrients (POC, PN, and PP) and Chl-*a* (proxy of phytoplankton biomass), indicating that particulate nutrients (≥0.7 μm in diameter) in the studied microbial size fraction (0.2-5 μm) may originate from the picoplankton biomass (**Fig. 2.7a**) - most likely from picocyanobacteria *Synechococcus* (~1 μm) and *Prochlorococcus* (~0.5 μm) which cumulatively comprised 66.92% of annotated sequences in our data. *Synechococcus* and *Prochlorococcus* usually dominate phytoplankton biomass in GBR waters[20,43,181,183], benefiting from favorable light conditions in offshore GBR reefs that facilitate photosynthesis. Therefore, particulate organic matter (POM) in the outer GBR predominantly originates from marine phytoplankton[184], contrasting with the terrestrial origin of POM found in riverine zones, inner estuarine mixing zones, and inshore reefs, with minimal amounts reaching the outer GBR[184]. Our results further suggest that POM in the outer GBR is predominantly

produced by *Synechococcus* during summer (average 62.38 % and 3.23 % relative abundance in summer trips for *Synechococcus* and *Prochlorococcus*, respectively), whereas during winter, we also observe considerable contribution of *Prochlorococcus* to POM production (average 37.02 % and 32.93 % relative abundance in the winter trip for *Synechococcus* and *Prochlorococcus*, respectively) (**Fig. 2.7a**). These picocyanobacteria have relevance to prospective monitoring since an increasing *Synechococcus*:*Prochlorococcus* abundance ratio was proposed as an index for elevated cross-shelf nutrient loads in reef waters[43], and we posit extending this index to a wider swath of offshore reefs, with *Synechococcus* indicative of high particulate nutrient loads broadly across the GBR. To further validate our proposed model, it would be beneficial to incorporate cell count data for *Synechococcus* and *Prochlorococcus* in future sampling efforts, as well as consider benthic cover organisms since emerging evidence suggests that corals exhibit preferential feeding on *Synechococcus*, potentially affecting their abundances[185,186]. Such approaches will enhance our understanding of picocyanobacterial contributions to POM dynamics and nutrient cycling in reef ecosystems.

Negative correlations were observed between particulate nutrients (POC, PN, PP) and Chl-*a* with dissolved inorganic nutrients (NH_4, NO_2, NO_3, PO_4, and TDP) (**Fig. 2.4a**), likely because this production of phytoplankton-derived POM from newly fixed carbon requires the uptake and assimilation of dissolved nutrients such as N, P, and trace metals[20]. Dissolved inorganic nutrients in shelf waters are rapidly taken up by growing phytoplankton (i.e. ~8-24 h for dissolved nitrogen, and ~24 h for dissolved phosphorus, see e.g. [20,187,188]) including *Synechococcus* and *Prochlorococcus*, which are highly efficient at using dissolved nutrients and exhibit a capacity for near-maximal growth down to available DIN levels of ≤0.02 μM - concentrations similar to the minimum detection levels[20]. Based on our findings and existing literature, we propose a mechanistic explanation, whereby picocyanobacteria *Synechococcus* and *Prochlorococcus* initially uptake dissolved nitrogen and phosphorus (resulting in decreased DIN concentrations), subsequently producing POM (POC, PN, and PP) during photosynthesis (**Fig. 2.7**, 1a). This process leads to increased phytoplankton biomass, as indicated by elevated Chl-*a*, and ultimately results in the observed collinear relationship between Chl-*a* and POM, negatively correlating with the uptake of dissolved inorganic nutrients (**Fig. 2.4, Fig. 2.5, Fig. 2.7**).

Several GO terms identified in our analysis are potentially involved in the uptake of dissolved nitrogen and phosphorus. For instance, alanine dehydrogenase, indicative of low ammonium concentrations across all sampling trips, likely plays a role in ammonium assimilation by catalyzing the synthesis of L-alanine from pyruvate and external ammonium[189]. Additionally, two subunits of the phosphate transport system (Pts), PstB (the catalytic subunit) and PstC (the transmembrane portion), were consistently enriched in low phosphate environments across sampling trips, suggesting an adaptive response to increase the uptake of limited inorganic phosphate[190]. While these GO terms positively correlated to Chl-*a* (proxy for phytoplankton biomass), further research is necessary to attribute these genes/functions to *Synechococcus* and *Prochlorococcus*, lineages well-documented for their genomic

heterogeneity, making genomic reconstructions from environmental metagenomics problematic[191–195]. Lastly, we also observe strong collinearity between phytoplankton-derived POM with DOC, and while DOC can be a product of extracellular release from actively photosynthesizing phytoplankton[196–198], alternative microbial processes may also produce DOC[199], including (1) senescing and dead phytoplankton[200,201], (2) sloppy feeding during zooplankton grazing[197,198,202,203], (3) POM dissolution by heterotrophic microbes[204–206], and (4) viral lysis[79]. Both DOC (regardless of its origin) and phytoplankton-derived POM can then enter the microbial loop[207–209] where a diverse consortium of seawater heterotrophic bacteria will benefit from nutrient-rich conditions (**Fig. 2.7**, 2a, b).

2.5.3 Phytoplankton-derived nutrients fuel the microbial loop and support higher trophic levels

Free-living pelagic microorganisms surrounding coral reefs enable the capture, retention, and recycling of nutrients and trace elements, essential to maintaining reef ecosystems in oligotrophic environments often likened to 'nutrient deserts'[19,22,210]. Heterotrophic seawater microbes positively associated to elevated POM and DOC in the surveyed offshore reefs included members of Gammaproteobacteria (two Oceanospirillaceae), Alphaproteobacteria (three Rhodospirillaceae, three Rhodobacteraceae), and two Burkholderiaceae (**Fig. 2.7b**) These microbes are frequently documented as enriched under elevated nutrients within coral reefs[1,40,43,60,61]. For example, Rhodobacteraceae are noted for their association with dissolved nutrients in inshore GBR reefs dominated by macroalgae[1,40,43]. Our findings show Rhodobacteraceae to be consistently enriched with elevated particulate nutrients in outer GBR reefs, indicating their role as versatile heterotrophic marine bacteria[211] potentially capable of utilising both dissolved and particulate nutrients in the GBR. Various members of Rhodospirillaceae also indicated high levels of particulate nutrients, although previous studies have reported their association with decreasing nutrient levels[43]. This discrepancy likely stems from their broad metabolic potential, which includes diazotrophic capabilities, opportunistic pathogenesis, and adaptation to various aerobic and anaerobic conditions[43,212], ultimately allowing Rhodospirillaceae to adapt to various niches across reef environments. Lastly, Flavobacteriaceae, known for their capacity to degrade complex polysaccharides and utilise diverse carbon sources[213], were the only group in our data enriched when both particulate and dissolved nutrient concentrations were elevated. Interestingly though, the MINT sPLS LOGOCV stability scores suggest that the signal of Flavobacteriaceae as indicators was not consistent across trips. This instability, coupled with low MINT sPLS correlation scores, suggests that Flavobacteriaceae are summer-specific indicators of elevated nutrients in the offshore GBR as Flavobacteriaceae were the most discriminatory of summer trips in our data, increasing both in relative abundance and diversity. This is in addition to the relevance of Flavobacteriaceae as indicators of labile polysaccharides released from macroalgae in inshore GBR reefs[1,40,43] where macroalgae cover is comparatively higher than in the outer GBR.

Numerous genes encoding for nutrient uptake systems were enriched in the GBR samples when DOC and phytoplankton-derived POM are available (**Fig. 2.7**, 1b) including ABC (ATP-binding cassette) transporters, TRAP (Tripartite ATP-independent periplasmic) transporter permease proteins, UAA (Uncharacterised Amino Acid) transporters, and various ion transporters. Concurrently, we found an enrichment of microbial Gene Ontology (GO) terms related to energy metabolism and cellular respiration (**Fig. 2.7**, 2b), such as NADH-quinone oxidoreductase (IPR010226, complex I of the respiratory chain), and cytochrome c oxidase subunit III (IPR024791, subunit of the terminal complex IV in the respiratory chain). These gene pathways drive electron transport and are coupled to proton transmembrane transport, generating a proton motive force for ATP synthesis, ultimately contributing to increased energy metabolism[214–217]. The energy generated from nutrient uptake and cellular respiration can then be directed towards anabolic metabolism and synthesis of various complex compounds[218–220] and we observed high representation of gene pathways involved in biosynthesis of vitamins, fatty acids, amino-acids, and proteins (**Fig. 2.7**, 3b). For example, vitamin B6 biosynthesis appears widespread in GBR bacterioplankton, facilitated by the 4-hydroxythreonine-4-phosphate dehydrogenase PdxA (IPR005255)[221] and pyridoxal kinase enzymes (IPR004625)[222–224], which were consistently enriched in samples from at least two trips (LOGOCV stability = 0.5), indicating that elevated particulate nutrients promote this biosynthesis, as also observed in pelagic microbes[220]. Further, the consistent presence of genes associated with fatty acid and amino acid biosynthesis indicates that elevated DOC and phytoplankton-derived POM in the GBR supports increased synthesis of these compounds. Specifically for fatty acid biosynthesis, NADH-dependent enoyl-acyl carrier protein reductase (IPR014358) was stably indicative of elevated nutrients in each sampling trip (LOGOCV stability score = 1), facilitating fatty acid biosynthesis by reducing the enoyl-acyl carrier protein (ACP) intermediate to produce saturated acyl-ACP[225,226]. For amino-acid biosynthesis, chorismate synthase (IPR000453), catalysing the final step of the shikimate pathway used by prokaryotes to synthesise aromatic amino acids[227] was stably (i.e. in three sampling trips, LOGOCV stability = 0.75) enriched with elevated DOC and phytoplankton-derived POM, suggesting enhanced amino acid biosynthesis in reef bacterioplankton under nutrient-rich conditions. As building blocks of proteins, these amino-acids are likely used in subsequent protein synthesis since various ribosomal proteins as essential components of protein-translation organelles ribosomes[228] were indicative of elevated POM and DOC. The indicator ribosomal proteins were: S6 (IPR000529) and L16 (IPR000114) identified as stable indicators across three and two sampling trips (respectively), and the trip-specific L5 (IPR002132, IPR020930), and S5 (IPR000851, IPR005712) (**Fig. 2.5, Fig. 2.7**, 3b).

Enhanced biosynthesis of complex compounds in nutrient-rich conditions can support the cellular growth and proliferation of reef-associated heterotrophic seawater microbes (**Fig. 2.7**, 4b). Crucial to the bacterial cell cycle (elongation and division) is the synthesis of bacterial cell walls which consist of peptidoglycans, composed of alternating units of N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM) connected via the β-(1,4)-glycosidic bond[229,230]. The NagA gene (N–acetylglucosamine–6–phosphate deacetylase - IPR003764) was persistently indicative of elevated DOC and phytoplankton-derived POM in

the outer GBR (i.e. in each of the four sampling trips), potentially facilitating the NAG breakdown to produce glucosamine-6-phosphate for synthesis of bacterial cell walls via the peptidoglycan recycling pathway[231]. NAG are among the largest pools of amino sugars in the ocean[232] and NAG utilisation is consistent with a previous metagenomic study in inshore reefs of the Central GBR, where NAG transporters were identified in reef water microbes, though absent from sponge and macroalgae microbiomes[1]. Further, anhydro-N-acetylmuramic acid kinase (IPR005338) was also enriched under elevated POM and DOC, another enzyme crucial for peptidoglycan recycling through phosphorylation of the anhydro-N-acetylmuramic acid (anhMurNAc) to produce MurNAc-6-phosphate, an intermediate in peptidoglycan metabolism during cell wall remodelling and turnover [233,234]. Consistent enrichment of these two enzymes across the GBR highlights that both bacterial cell wall biosynthesis and maintenance (indicative of microbial growth and proliferation) are widespread in heterotrophic seawater microbes when DOC and phytoplankton-derived POM are available (**Fig. 2.7**, 4b). Further investigation into the metabolic activities of these indicator microbes and genes, using techniques such as metatranscriptomic and metaproteomic analyses, as well as stable isotopes, could provide richer insights into how nutrient availability influences the composition and metabolism of GBR seawater microbial communities.

Five physico-chemical variables, salinity, total suspended solids (TSS), total dissolved nitrogen (TDN), silica (Si), and nitrate (NO_3), did not significantly influence the overall community composition or functional potential (**Fig. 2.2c**). This is likely due to our sampling design where all sites are offshore reefs and therefore some metrics have a low explanatory value as they are highly consistent across this longitudinal gradient. Salinity, for example, has been well-documented as one of primary factors shaping community composition in aquatic microbes[169] and was reported to explain 4.2% of community variation (according to Variation Partitioning Analysis) in the GBR seawater microbiomes[43]. However, inshore sites influenced by freshwater input and therefore a stronger salinity gradient were investigated[43], while our data captured a low degree of variation in salinity (34.6 to 35.8 practical salinity units – PSU, Table 1), which is likely why salinity was not significant in our study (**Fig. 2.2c**). Moving forward, reevaluation of these specific physico-chemical variables should occur in the future if additional sampling introduces a broader range of sites, particularly areas of inshore reefs, where proximity to land and human activities contributes to a wider range of environmental variation.

2.6 Conclusions

Our study provides a functional baseline for reef-associated bacterioplankton in the outer GBR, demonstrating that microbial functional genes have a higher stability than taxonomy in inferring physico-chemical variables across broad spatio-temporal scales. When dissolved organic carbon (DOC) and phytoplankton-derived particulate organic matter (POM) are elevated in offshore and mid-shelf GBR reefs, microbial genes and functions we found as consistently enriched in heterotrophic seawater microbes collectively point towards enhanced microbial nutrient uptake (**Fig. 2.7**, 1b) and energy generation

through cellular respiration (**Fig. 2.7**, 2b), supporting anabolic metabolism and synthesis of complex compounds (**Fig. 2.7**, 3b) to ultimately increase growth and biomass of heterotrophic seawater microbes (**Fig. 2.7**, 4b). Members of reef bacterioplankton that increased in relative abundances with elevated POM and DOC consistently across seasons/sectors in the offshore GBR included Rhodospirillaceae, Rhodobacteraceae, and Burkholderiaceae, whereas Flavobacteriaceae were enriched when both dissolved and particulate nutrients were elevated, although predominantly during summer (**Fig. 2.7b**). These heterotrophic marine microorganisms can then be grazed by flagellates and microzooplankton which in turn support larger macroorganisms, ultimately transferring nutrients derived from *Synechococcus* and *Prochlorococcus* to higher trophic levels in the outer GBR (**Fig 2.7**, 1c). Phytoplankton-derived POM (i.e. retained on a filter with a pore size of approximately 0.7 µm) not immediately metabolised by heterotrophic seawater microbes will escape the microbial loop, also becoming available to benthic and pelagic organisms at higher trophic levels through direct uptake (**Fig 2.7**, 2c). In summary, *Synechococcus* and *Prochlorococcus* are crucial components of the marine food web in the outer Great Barrier Reef, supporting various levels of the ecosystem through their role as primary producers and their contributions to nutrient cycling and carbon sequestration.

Since microbial genes had higher indicator stability scores and functional redundancy is a well-established phenomenon for pelagic microbes[132,133,235,236], we assert that microbial functions have a higher utility than microbial taxa for rapid assessment of reef ecosystem health. It is worth noting, however, that this study was conducted using taxonomic annotations derived from metagenomic reads, which may provide less resolution than 16S rRNA gene based taxonomic annotations. Microbial transcriptomic profiling assays and biosensors, already used in environmental toxicity testing[237] to detect heavy metal pollution[238] and track hydrocarbon degradation from oil spills[239], would benefit from improved collaboration between researchers and reef managers to identify the most suitable microbial markers (taxa or genes/functions) for developing targeted microbial-based assays for rapid reef health assessment. Lastly, as reef metagenomes become more widely available[63,82,143], it will be possible to cross examine data sets across global scales and integrate microbial responses to generate spatio-temporally coherent baselines of microbes indicating reef health, however care will be needed to distinguish microbial biomarkers from confounding factors such as geography and season. To complement these emerging large-scale surveys of reef seawater microbes, it will be crucial to capture the state of reef bacterioplankton over time as is being recorded for pelagic microbes[18], for example, at (1) long-term ocean time-series stations (which are yet to be established, unlike the 72 microbial observatories catalogued so far for pelagic microbes[240]), at (2) day-to-day resolution[52,241] and across (3) mesoscale processes[242]. Such baselines of reef-associated (bacterio)plankton will be invaluable in facilitating identification of deviations that could signal impending disturbance events[43,103] and link how microbial community shifts contribute to ecosystem stability and transition to alternative stable states.

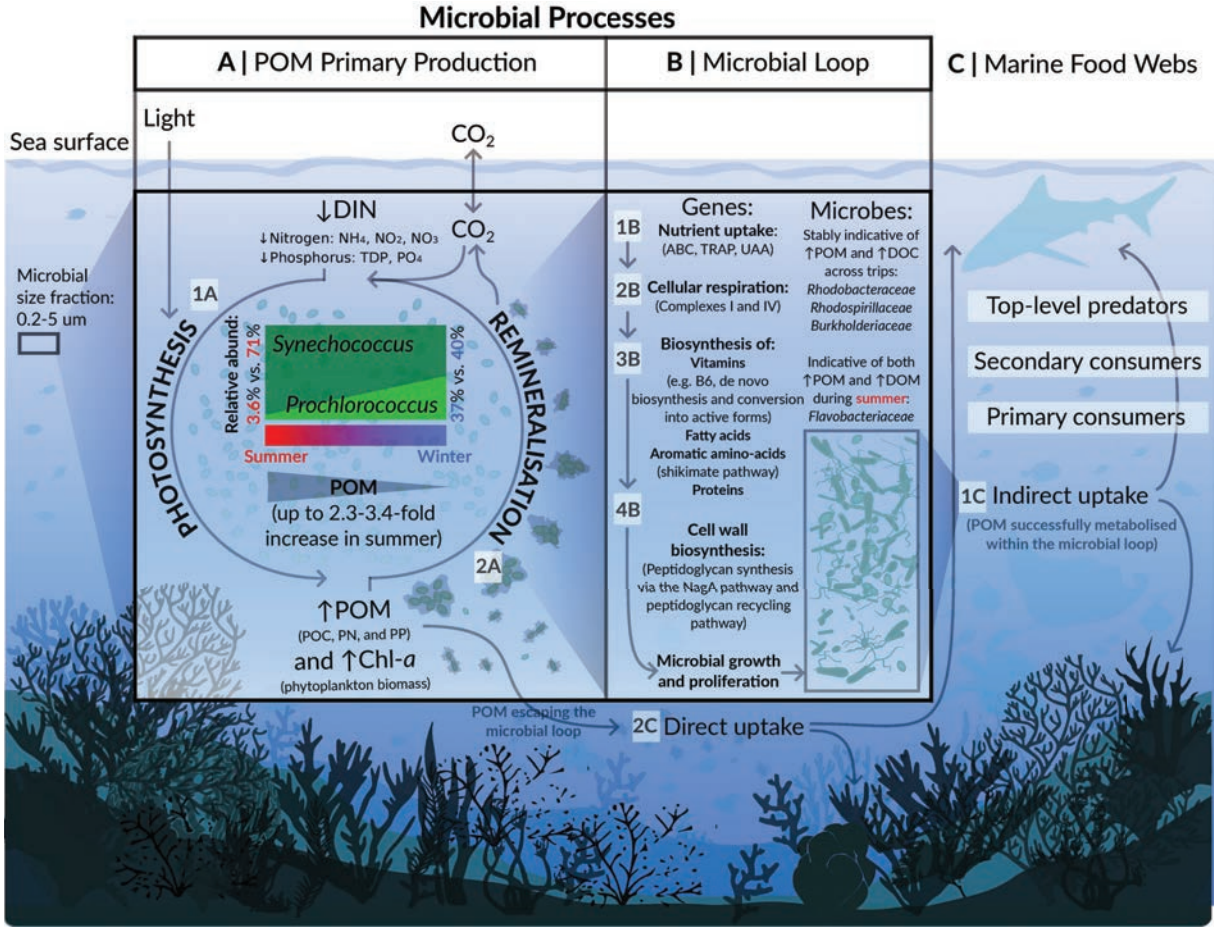

Figure 2.7: Conceptual overview summarising the roles of seawater microbiomes in nutrient cycling in offshore Great Barrier Reef (GBR) surface waters. (**A**) Planktonic picocyanobacteria *Synechococcus* and *Prochlorococcus* play key roles in nutrient cycling in the outer GBR: they uptake dissolved inorganic nutrients (DIN) such as nitrogen (ammonium - NH_4, nitrite - NO_2, and nitrate - NO_3) and phosphorus (phosphate - PO_4), reducing DIN concentrations (1A). In the presence of light and carbon dioxide (CO_2) the uptaken DIN will be used during photosynthesis to produce particulate organic matter (POM) including organic carbon (POC), nitrogen (PN), and phosphorus (PP), overall resulting in elevated POM concentrations and higher biomass of these picocyanobacteria, indicated via elevated chlorophyll a (Chl-*a*) (**2A**). During summer, elevated photosynthesis rates primarily by *Synechococcus* result in up to a 3-fold increase in POM production, whereas during winter, nutrient concentrations are lower and we also observe notable contributions of *Prochlorococcus* to POM production (**2A**). A fraction of POM deriving from *Synechococcus* and *Prochlorococcus* will (**B**) enter the microbial loop. Here, a diverse consortium of seawater heterotrophic microbes, notably Rhodobacteraceae, Rhodospirillaceae, Oceanospirillaceae, Burkholderiaceae and Flavobacteriaceae, will benefit from nutrient-rich conditions by encoding genes for (**1B**) nutrient uptake and (**2B**) cellular respiration to generate energy, which can be directed towards (**2C**) synthesis of complex compounds and (**2D**) microbial growth. As a result of microbial activity on phytoplankton-derived POM, organic molecules originally present in particulate form are remineralised into DIN (NH_4, NO2, NO_3, PO_4), and dissolved organic carbon (DOC). These dissolved nutrients are then available for uptake by other organisms, including *Synechococcus* and *Prochlorococcus* which can photosynthesise again, ultimately recycling POM in the outer GBR and making it available to higher trophic levels (**C**). POM from these picocyanobacteria may enter marine food webs via two pathways: (**1C**) an indirect pathway, where heterotrophic seawater microbes that successfully integrated phytoplankton-derived POM into their biomass will be grazed by flagellates and microzooplankton, which in turn will support larger macroorganisms; or (**2C**) through direct uptake of POM that escapes immediate

metabolism by heterotrophic seawater microbes, thus bypassing the microbial loop.

2.6.1 List of abbreviations

Abbreviations used in this study include (ordered alphabetically): **ABC**: ATP-binding Cassette, **AIMS**: Australian Institute of Marine Science, **Chl-*a***: Chlorophyll-a, **CLR**: Center-Log-Ratio, **DIAMOND**: Alignment Tool, **DOC**: Dissolved Organic Carbon, **ECO FLNTU-RT**: Fluorescent and Turbidity sensor (Real-Time), **FastQC**: Quality Control Software, **GBR**: Great Barrier Reef**, GO**: Gene Ontology, **IMOS**: Integrated Marine Observing System, **Inkscape**: Vector Graphics Software, **IPR**: InterPro, **LOGOCV**: Leave-One-Group-Out Cross-Validation, **MEGAN**: Metagenome Analyzer, **Millipore Sterivex-GP**: Pressure Filter, **MINT**: Multivariate INTegrative method, **MINT sPLS**: Multivariate INTegration Sparse Partial Least Squares, **N**: Nitrogen, **NADH**: Nicotinamide Adenine Dinucleotide (NAD) + Hydrogen (H), **NAG**: N-acetylglucosamine, **NAM**: N-acetylmuramic acid, **NanoDrop**: Spectrophotometer, **NH_4**: Ammonium, **NO_2**: Nitrite, **NO_3**: Nitrate, **NovaSeq**: Sequencing System, **PCA**: Principal Component Analysis, **PERMANOVA**: Permutational Multivariate Analysis of Variance, **Phaeo**: Phaeophytin, **PLS**: Partial Least Squares, **PO_4**: Phosphate, **POC**: Particulate Organic Carbon, **PN**: Particulate Nitrogen, **PP**: Particulate Phosphorus,**Qubit** :Fluorometer for DNA Quantification, **R**: Programming Language, **RV**: Research Vessel, **Sartorius**: Filter Manufacturer, **Sartorius Minisart N**: Syringe Filter, **SBE**: Sea-Bird Electronics, **Shimadzu TOC-L**: Carbon Analyzer, **Si**: Silicate, **sPLS**: Sparse Partial Least Squares, **SST**: Sea Surface Temperature, **TARA**: Tara Oceans Foundation, **TDN**: Total Dissolved Nitrogen, **TDP**: Total Dissolved Phosphorus, **TRAP**: Tripartite ATP-independent Periplasmic, **Trimmomatic**: Quality Filtering Software, **TSS**: Total Suspended Solids, **UAA**: Uncharacterised Amino Acid, **VEGAN**: R Package for Diversity Analysis, **µg**: Microgram, **µm**: Micrometer, **mm**: Millimeter, **cm**: Centimeter, **h**: Hour, **L**: Liter.

2.7 Declarations

2.7.1 Ethics approval and consent to participate

Samples were collected under the permit G12/35236-1 issued by the Great Barrier Reef Marine Park Authority.

2.7.2 Consent for publication

Not applicable.

2.7.3 Availibility of data and material

Raw sequencing data and the associated physico-chemical variables have been uploaded to the IMOS-AODN repository and are available at: Australian Institute of Marine Science (AIMS). (2022). Great Barrier Reef

Genomics Database: Seawater Illumina Reads. https://doi.org/10.25845/Q4XH-YN10. The code to replicate the analysis is available at: https://github.com/mterzin/IMOS_GBR_MGD_read-centric_analysis.

2.7.4 Competing interests

The authors declare no competing interests.

2.7.5 Funding

This study forms part of the Australia's Integrated Marine Observing System (IMOS) Great Barrier Reef Microbial Genomic Database sub-facility (GBR-MGD), funded by the Queensland Research Infrastructure Co-investment Fund (RICF) by the Department of Environment and Science, Queensland. IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent. This study was also funded by an AIMS@JCU PhD Scholarship to MT. The funders had no role in sampling design, data collection, processing and interpretation, preparation of the manuscript, or decision to publish.

2.7.6 Authors' contributions

NSW obtained funding for the project. NSW, DGB, PWL, RKG, and SJR conceived the sampling design. SCB collected seawater in the field. SCB processed all samples in the laboratory for metagenomic sequencing. MT analysed the data, with the assistance of PWL, SJR, YKY, DGB, KALC, RKG, and PRF. MT wrote the original draft of the manuscript, and PWL, DGB, SJR, YKY, KALC, RKG, NSW, PRF, and SCB made substantial contributions to its form. All authors critically reviewed the manuscript before submission.

2.7.7 Acknowledgements

The seawater samples analysed in this study for metagenomics and physico-chemical variables were collected from a total of 48 offshore and midshelf reefs, in the (1) Far Northern GBR - the traditional sea Country of the Gudang Yadhaigana, Kuuku Ya'u, Lama Lama, Cape Melville, Howicks and Flinders Island Traditional Owners, (2) Northern GBR - the traditional peoples of Eastern Kuku Yalanji, Yirrgandji, Gunggandji, Gunggandji-Mandingalbay Yidinji sea country estates, (3) Central GBR - Mandubarra, Wulgurukaba and Bindal Traditional Owners and (4) Southern GBR - the traditional sea Country estate of the Port Curtis Coral Coast Traditional Owners. We pay our respects to their Elders past, present, and emerging. Further, our desktop / lab research took place at the Australian Institute of Marine Science (AIMS) headquarters at Cape Ferguson, and we wish to acknowledge the Bindal peoples as the Traditional Owners of the land. This research was also undertaken at the JCU Townsville Bebegu Yumba campus, and

the authors acknowledge the Wulgurukaba Peoples as the traditional owners of this site. We acknowledge Australian Aboriginal and Torres Strait Islander peoples are the original inhabitants and traditional custodians of this continent and their unique cultural and spiritual relationships to the land and waters. We acknowledge the AIMS Water Quality team, especially Ulysse Bove, Keeley Glasson, and Daniel Moran for logistics, training, and processing of water chemistry samples. We acknowledge the AIMS-LTMP team and others involved in field collection and preparation of samples including Michael Emslie, Emmanuelle Botté, Johnston Davidson, Veronique Mocellin, and Josephine Nielsen. We thank the crew of the RV Solander and RV Cape Ferguson for their excellent logistical support in the field. We also acknowledge Gene Tyson for his support in facilitating the use of the NovaSeq at Microba Life Sciences Ltd. (Brisbane, QLD, Australia). We extend our gratitude to Murray Logan for his insightful discussions on the appropriate statistical handling of the data. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2025648).

Data Analysis | No-take marine reserves promote oligotrophic reef bacterioplankton communities across the Great Barrier Reef

Manuscript in revision for *Nature Communications* (transferred from *Nature Microbiology*).
Terzin, M., Robbins, S.J., Lê Cao, K.-A., Bell, S.C., Dougan, K.E., Zaugg, J., Gruber, R.K., Emslie, M.J, Ceccarelli, D.M., Chaffron, S., Hugenholtz, P., Webster, N.S., Bourne, D.G., Yeoh, Y.K., Laffy, P.W., 2025. No-take marine reserves promote oligotrophic reef bacterioplankton communities across the Great Barrier Reef.

3.1 Abstract:

Australia's Great Barrier Reef is a biodiversity hotspot critical to ocean health, yet it faces increasing threats from climate change and localised impacts requiring effective conservation and management action. Rezoning of the Great Barrier Reef Marine Park in 2004 expanded No-Take Marine Reserves (NTMRs) to restrict extractive activities like fishing and collecting, creating one of the largest networks of marine reserves globally. Benefits like increased biomass of fisheries-targeted species and improved coral community health metrics have been reported, though the effects of zoning on water chemistry and seawater microbiology remain unexplored. Using seawater metagenomes data from the Great Barrier Reef Microbial Genomics Database, we investigated the structure of seawater microbiomes on 48 offshore reefs within NTMRs (green zones) and fished reefs (dark-blue and yellow zones). A supervised classification method identified 350 indicator species that predict zoning with ~71% accuracy. Microbial communities broadly reflected reef states, with green zones enriched in streamlined microbial oligotrophs (*Pelagibacter* and SAR86 MAGs) correlating with higher cover of hard coral, crustose coralline algae, and herbivore fish abundance under lower nutrient conditions. By contrast, fished reefs harbored opportunists (Flavobacteriales, especially UA16, and Pseudomonadales) associating with elevated nutrients and turf algae cover. Co-occurrence networks revealed stronger competitive interactions in fished reefs, where nutrient-responsive taxa may outcompete other microbes, underscoring the need to investigate how these shifts influence reef nutrient cycling and function. Our findings reveal ecosystem-wide effects of marine zoning beyond fish protection, with distinct seawater microbiomes between fished reefs and NTMRs, which will help build decision tools for more targeted reef health monitoring assessments.

3.2 Introduction

Marine Protected Areas (MPAs), particularly No-Take Marine Reserves (NTMRs), represent conservation tools aimed at protecting exploited species and ecosystems through restricting extractive activities such as fishing and mining[243–246]. Recent estimates indicate that coral reefs within these marine reserves support significant recovery and maintenance of exploited fish biomass, accounting for an estimated 10% of the existing fish biomass in reefs globally[247]. Fisheries management success was also documented for Australia's Great Barrier Reef (GBR) Marine Park which was rezoned in 2004, resulting in NTMRs increasing from ~5% to ~33% of the entire area[248] and thereby becoming the world's largest network of marine reserves at the time[249,250]. Benefits of rezoning accrued rapidly[249,251], with coral trout (*Plectropomus* spp., *Variola* spp.) density increasing by 57–75% within two years[252] and biomass by up to 89% within four years of rezoning[253]. NTMRs now support half of grouper biomass and 47% of fishery yield within the GBR Marine Park through spillover into adjacent fished areas[254].

In addition to supporting fish biomass recovery, NTMRs may also generate positive cascading effects that enhance broader ecosystem resilience, increasing the capacity of protected GBR reefs to resist and recover from disturbances while maintaining key ecological functions. For example, NTMRs display lower impacts and faster recovery from disturbances such as bleaching, cyclones, and crown-of-thorns starfish (CoTS) outbreaks[253,255], and they support more spatially heterogeneous benthic communities that promote higher biodiversity[256]. NTMRs also show fewer CoTS outbreaks, possibly linked to increased fish predators of juvenile starfish[257], in addition to having reduced coral disease prevalence, likely associated with less fishing-related tissue damage[258]. These GBR-specific trends align with global evidence demonstrating that NTMRs support healthier reef benthic communities, with higher cover of hard coral and crustose coralline algae (CCA), increased fish diversity and biomass, reduced macroalgal cover, and faster recovery rates[259,260], contributing to enhanced resilience against climate change and other disturbances[261,262].

Microbial communities provide key services for coral reef ecosystems such as nutrient cycling and maintaining host fitness vital to productivity and resilience[263], however little is known about reef zoning influence on pelagic microbiomes. Reef bacterioplankton diversity and community composition shift in response to environmental changes such as nutrient loads, temperature, and benthic cover, and may serve as indicators of ecosystem health[264–267]. Currently, only a few meta-omics studies have explored how zoning affects microbial dynamics of the surrounding bacterioplankton, but emerging evidence suggests: (1) NTMRs in Brazil's Abrolhos Bank supported distinct seawater microbial communities alongside higher fish biomass and lower macroalgal cover[268] and (2) in Kuwait's Sulaibikhat Bay MPA, protected sites showed higher microbial diversity correlated with elevated nitrogen, phosphorus and salinity levels compared to fished areas[269]. However, in both studies, protected and unprotected sites were located far apart and thus the effects of protection status remain challenging to isolate from inherent environmental differences, and may be partially confounded by the offshore placement of protected reefs.

Reef protection measures may influence seawater microbial communities by reducing human impacts and promoting more diverse bacterioplankton communities, however, more targeted research is needed to clarify the relationship between reef zoning and the health and functioning of seawater microbes. To address this gap, we examined the distribution of 5,283 prokaryotic metagenome-assembled genomes (pMAGs) assembled from 48 offshore reef-associated seawater metagenomes from the Great Barrier Reef Microbial Genomics Database (GBR-MGD)[270]. Here, we compared co-located fished and NTMR reefs to identify microbial indicators of reef zoning status, and clarify their relationships to protection status through associations with physicochemical variables (temperature, salinity, chlorophyll *a*, and dissolved/particulate nutrients), benthic cover (coral types, algae, and abiotic components) and fish abundances and biomass. Using supervised machine learning, microbial niche modeling, and network analysis, we assessed how reef zoning shapes reef bacterioplankton composition, functioning, and

microbe-to-microbe interactions, identifying seawater microbial signatures that could inform future monitoring.

3.3 Materials and Methods

3.3.1 Seawater sampling

Water sampling was conducted on four vessel-based field trips by the Australian Institute of Marine Science (AIMS) between November 2019 and July 2020 for (1) microbial community profiling and (2) analysis of physico-chemical variables. The field trips covered the Northern (Trip 1: Princess Charlotte Bay and Cape Grenville sectors; November–December 2019), Southern (Trip 2: Swains and Capricorn Bunker sectors; January 2020), and Central Great Barrier Reef (Trip 3: Cairns and Innisfail sectors; February 2020; and Trip 4: Townsville sector; July 2020). The first three trips were conducted concurrently with AIMS-LTMP *in situ* reef health surveys collecting data on (3) fish abundance and biomass, and reef benthic cover, with the fourth water sampling trip occurring 2 months after the AIMS-LTMP reef health surveys, at the same sites. At each reef site, surface seawater was collected at approximately five metres depth (2-10 m depending on site and tide) using SCUBA (Trips 1-3) or Niskin bottles (Trip 4) (**Fig. 3.1**).

During each trip, sampling sites were evenly distributed across fished reefs and NTMRs for a total of 48 offshore reefs (23 NTMRs and 25 fished reefs). Of the fished reefs, 20 were located in Habitat Protection zones (dark blue) where most fishing activities are allowed with the exception of trawling, while 5 were in Conservation Parks zones (yellow), where fishing for sea cucumbers, lobsters, and net fishing are also prohibited[271].

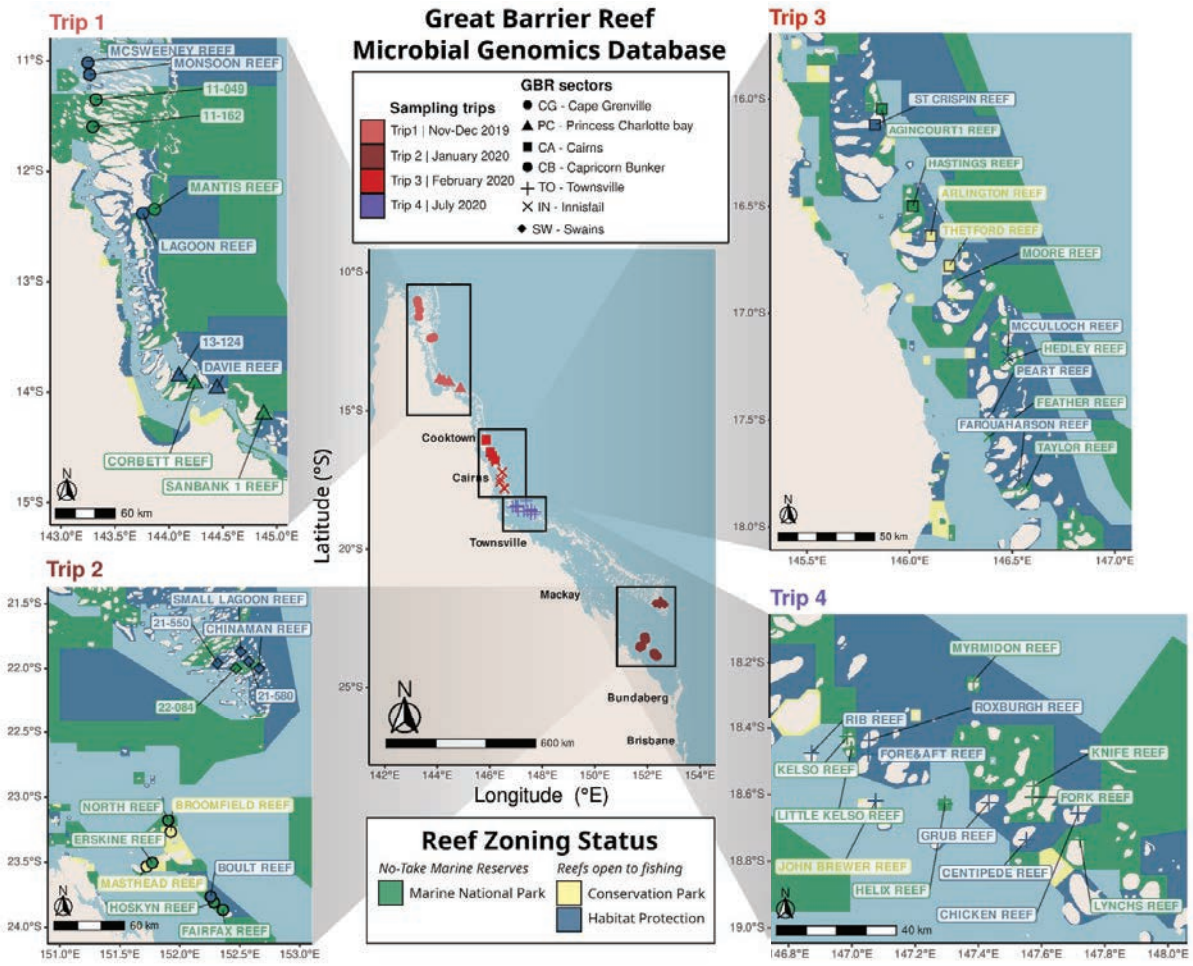

Figure 3.1: Field sampling design for the Great Barrier Reef Microbial Genomics Database (GBR-MGD) dataset. Seawater was collected from 48 offshore GBR reefs for microbial community metagenomic sequencing and analysis of physico-chemical variables, concurrently with AIMS-LTMP in situ estimates of benthic cover and fish abundance and biomass. Sampling occurred in four trips between November 2019 and July 2020, with red tones indicating Austral summer/wet season (trips 1-3, Nov 2019–Feb 2020) and blue indicating winter/dry season (trip 4, July 2020). Samples were taken across seven GBR sectors, denoted on the maps with different symbols. Trip-specific map insets show that reefs in No-Take Marine Reserves (NTMRs, green zones) and fished reefs (dark-blue and yellow zones) were sampled in pairs to minimise confounding effects of geography.

Seawater processing for metagenomics

For microbial sampling, four 5 L seawater replicates from each reef were first filtered through 5 µm (Minisart® NML syringe pre-filter [Sartorius, Goettingen, Germany]) followed by 0.22 µm filters (Millipore® Sterivex-GP™ Pressure Filter [Merck Millipore, Darmstadt, Germany]) immediately after collection, and the 0.22 µm filters were snap frozen and stored at -75 ℃. DNA was extracted from the frozen 0.22 µm filters using phenol:chloroform:iso-amyl alcohol extraction with ethanol precipitation, followed by quality assessment using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Australia) and quantification using a Qubit 3 fluorometer (Thermo Fisher Scientific, Australia). DNA was sent to the Australian Centre for Ecogenomics (ACE, University of Queensland, Australia) and Microba

Life Sciences Ltd (Brisbane, Australia) for short-read metagenome sequencing (Nextera FLEX 2x150 bp library prep, Illumina NovaSeq). Hybrid metagenome assemblies were generated from 27 samples (one replicate from 27 of the 48 sites) that underwent both long-read sequencing on the Oxford Nanopore PromethION platform and deep (i.e., 40 Gbp) Illumina sequencing for polishing, also at the ACE sequencing centre[1]. High–molecular weight DNA was size-selected with the Circulomics SRE XS kit (PacBio, SKU 102-208-200), barcoded (Oxford Nanopore, LSK-109 with EXP-NBD104), and sequenced on PromethION R9.4 flow cells using MinKNOW (v20.06.18; default settings), with raw reads subsequently re-basecalled using Guppy (v5.0.16) in superaccuracy mode. Adapter and barcode trimming was performed with Porechop (https://github.com/rrwick/Porechop) under default parameters.

Seawater processing for physico-chemical variables

Water chemistry was measured in three of the four seawater replicates[2] using established methods[272]. Fourteen variables were measured, including dissolved nutrients (total dissolved nitrogen – TDN; ammonium – NH_4^+; nitrite – NO_2^-; nitrate – NO_3^-; total dissolved phosphorus – TDP; phosphate – PO_4^{3-}; dissolved organic carbon – DOC; silicate – Si), particulate fractions (particulate organic carbon – POC; particulate nitrogen – PN; particulate phosphorus – PP), pigments (chlorophyll *a* – Chl-*a*; phaeophytin *a* – Phaeo), and total suspended solids – TSS. Seawater for the measurement of dissolved nutrients (NH_4^+, NO_2^-, NO_3^-, PO_4^{3-}, TDN, TDP, DOC, and Si) was immediately filtered (0.45 **µ**m Sartorius Minisart N) into 10 mL acid-washed vials (triple pre-rinsed with filtered site water) and all analytes except DOC and Si were stored frozen (-18 ℃) until laboratory processing. Samples for DOC and Si analyses were stored refrigerated (4 ℃) until analysis, and DOC was acidified with 100 µL AR-grade hydrochloric acid immediately following sampling. Particulate variables (POC, PN, PP, and Chl-*a*) were filtered onto pre-combusted (450 ℃ for 4 h) 25 mm glass fibre 0.7 µm filters (Whatman GF/F), wrapped in foil, and stored frozen (-18 ℃) until analysis. TSS samples were filtered onto pre-weighed 47 mm polycarbonate filters with 0.4 **µ**m pore size (GE Water & Process Technologies), triple-rinsed with ultrapure water, and analysed gravimetrically upon return to AIMS in Townsville.

Laboratory analyses were conducted at the AIMS Analytical Technology and Water Quality Laboratories within one month (Chl-*a*, DOC, TSS) or three months (all other variables) of collection. Dissolved nutrient concentrations (NH_4^+, NO_2^-, NO_3^-, PO_4^{3-}, Si, TDN, and TDP) were determined using wet chemical methods[273–275] on a segmented flow analyser (Seal AA3). Persulfate digestion[276] was applied to TDN and TDP samples prior to analysis. Organic carbon (DOC and POC) and PN were analysed via high

1 The seawater samples, DNA, and a subset of the Illumina short-read data are derived from the same collection used in Chapter 2 (Terzin et al., 2025). Chapter 2 utilised only shallow short-read Illumina data (target 5 Gbp per replicate) for a read-based metagenomics analysis, whereas the pMAG generation in this chapter in addition relied on deeper Illumina sequencing (target 40 Gbp) and Oxford Nanopore long reads from a subset of 27 samples, enabling hybrid assembly and high-quality MAG recovery.

2 The physico-chemical data used herein are the same 17 variables generated and detailed in Chapter 2 (Terzin et al., 2025), including: temperature, salinity, chlorophyll-a fluorescence, particulate and dissolved nutrients (carbon, nitrogen and phosphorus), and photosynthetic pigments (chlorophyll-*a* and phaeophytin *a*). Analytical methods and quality control procedures follow those previously described.

temperature catalytic combustion (Shimadzu TOC-L) with solid sample module (SSM-5000A) and nitrogen module (TNM-L). Persulfate digestion[277,277] was used prior to spectrophotometric analysis of PP samples[274]. Chl-*a* filters were ground and incubated for 2 h in 90% acetone prior to reading on a fluorometer (Turner 10AU). Samples were then acidified and re-read to determine the concentration of Phaeo[278].

In addition, underway physico-chemical data (temperature, salinity, Chl-*a* fluorescence) were obtained from IMOS Ships of Opportunity sensors (SBE 38 thermometer, SBE 21 Thermosalinograph, WET Labs ECO-FLNTU-RT fluorometer) at 1.9 m (RV Cape Ferguson) or 2.5 m (RV Solander), recording the closest measurement to sampling time[279,280]. This resulted in 17 physico-chemical variables analysed in this study: 14 water chemistry parameters, temperature, salinity, and Chl-*a* fluorescence. A more detailed protocol can be found in[281].

Fish abundance and biomass, and reef benthic cover

Standardised protocols were used to survey benthic and fish assemblages at 48 reefs as part of the Long-Term Monitoring Program (LTMP) by the Australian Institute of Marine Science (AIMS). Trained divers conducted fish and benthic surveys on SCUBA at three sites per reef in a standard reef slope habitat, simultaneously with collections of seawater samples for metagenomic sequencing and physico-chemical analyses. At each site, five 50 m permanently marked transects were surveyed, set parallel to the reef crest, and at depths between 4 and 12 m (depending on site bathymetry).

Along each transect, large-bodied, mobile fishes were surveyed within a 5 m belt (250 m^2 area) to capture diurnal, non-cryptic reef fish species. Non-cryptic, diurnal small-bodied fishes (e.g., Pomacentridae, small Labridae) were also surveyed using a transect width of 1 m (50 m^2 area). Fish were identified to species and abundance counts were subsequently aggregated into trophic groups and at family level based on the latest available classification. For each individual fish, length was estimated using predefined size classes (5 cm bins for large mobile fishes counted on the 5 m belt, 2 cm bins for species counted on the 1 m belt). Fish biomass was calculated from length estimates using species-specific length-weight relationships, and based on the formula: Biomass = Abundance × a × $(\text{Midpoint})^b$, where Abundance is the number of fish, and a and b are species-specific coefficients derived from FishBase[282]. The midpoint refers to the average length of fish within each size category. Total biomass (kg per 1000 m^2) for each transect was summed to estimate the overall fish biomass for the surveyed area (i.e., per reef).

Benthic surveys were conducted concurrently along the same transects using digital imagery. Digital images of the substrate were taken on the up-slope side of each transect at 50 cm intervals. Estimates of proportional benthic cover were subsequently derived from the identification of the benthos beneath five fixed points arranged in a quincunx pattern digitally overlaid onto these images. A total of 40 images from each transect (n = 3000 points $reef^{-1}$) were randomly selected and analysed using a specialised point image classifier. Hard corals were identified to the lowest taxonomic resolution possible, usually genera[283]. A total of 37 *in situ* variables were used, encompassing benthic cover (abiotic substrates,

soft corals, hard corals by morphology, algae, sponges, and other biota), and fish community data (aggregated per trophic group, family-level taxonomy, and overall biomass).

3.3.2 Metagenome hybrid assembly, binning, taxonomic annotation, and abundance estimation

The prokaryotic metagenome-assembled genomes (pMAGs) analysed in this study were generated with hybrid (Illumina-Nanopore) and short-read-only assemblies as part of the Great Barrier Reef Microbial Genomics Database (GBR-MGD), with full methodological details and code previously described[270]. Briefly, metagenomes were assembled from the Illumina and Nanopore (27 sites) or Illumina-only data (21 sites), and pMAGs subsequently binned from the metagenome assemblies using the Aviary (https://github.com/rhysnewell/aviary; v0.3.3) assembly and genome binning pipelines. The resulting 5,283 high-quality pMAGs (**Appendix B: Table S1**; deposited under EBI BioProject PRJEB82623) were taxonomically classified with the Genome Taxonomy Database Toolkit (GTDB-Tk[284], release R214) and subsequently dereplicated at 95% ANI using CoverM[285] (v0.6) (see **Appendix D** for comparison with NCBI nomenclature). This resulted in a total of 876 "species-resolved" $pMAGs_{95\%ANI}$ (**Appendix B: Table S2**) used in downstream analysis (and hereinafter referred to as pMAGs for brevity). Read mapping counts and relative abundances were inferred by mapping Illumina short reads to the dereplicated set of pMAGs using minimap2[286] (v2.18; as implemented in CoverM). To account for sparsity and compositionality issues inherent to microbial metagenomics data, raw read counts were transformed using a center log-ratio (CLR) transformation in the microbiome[287] (version 1.24.0) R package. All statistical analysis and visualization were performed on CLR-transformed and relative abundance data, and final figures were compiled using Inkscape (v0.92.5).

3.3.3 Functional annotation of differentially enriched KEGG modules between microbes discriminating NTMRs and fished reefs

The de-replicated set of 876 pMAGs were imported into anvi'o[288] (v8) for functional annotation. Using the anvi'o *contigs* workflow (https://anvio.org/help/main/workflows/contigs/), open reading frames (ORFs) in each pMAG were first predicted using Prodigal[289] (v2.6.3) and then compared against anvio's set of protein hidden markov models (HMMs) for functional annotation with the *anvi-run-hmms* command. Using the *anvi-run-kegg-kofams* program, ORFs that matched to HMMs were then mapped against the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) database[290] using HMMER (v3.3.2, http://hmmer.org/). These mappings were used to assess the stepwise completeness of metabolic pathways in each pMAG with the *anvi-estimate-metabolism* program using KEGG 'modules'. The resulting KO annotations and KEGG module completeness statistics were exported from anvi'o (with *anvi-export-functions*) for downstream indicator analyses of NTMRs and fished reefs (**Appendix B: Table S3**). From 358 detected KEGG modules identified in our data (**Appendix B: Table S3**), Sparse Partial Least Squares Discriminant Analysis (sPLS-DA)[291,292] was used to identify the top 50 modules showing differential

completeness between NTMR-associated (n=236) and fished reef-associated (n=114) microbial indicators, which was further validated with Wilcoxon rank sum tests in R.

3.3.4 Seawater microbial indicators of reef zoning

Principal Components Analysis (PCA) in mixOmics[292] (v6.26.0) (R version 4.3.2; R Core Team, 2023) was first used to explore major sources of variation in microbial community composition, revealing that temporal (sampling time) and spatial (site proximity) factors explained more variation (**Appendix B: Fig. S1**) than reef zoning status (**Appendix B: Fig. S2**). To account for these spatiotemporal confounding effects, microbial community data were partitioned by GBR sector since AIMS-LTMP benthic cover and fish abundance data have historically been analysed across GBR sectors[249,293]. Multivariate INTegration Sparse Partial Least Squares Discriminant Analysis (MINT sPLS-DA)[291,292,294] was then used to identify microbial indicators that consistently discriminated NTMRs from fished reefs across sectors. For comparison, a conventional (i.e., without sector integration) sPLS-DA was also performed to identify microbial indicators of reef zoning, which yielded discriminatory features but retained strong spatiotemporal batch effects (**Appendix B: Figs. S3–S7; Tables S1–2**), supporting the use of MINT for this dataset.

MINT sPLS-DA was run on microbial abundance data (CLR-transformed within each sector) with model tuning performed via Leave-One-Group-Out Cross-Validation (i.e., training the model on six sectors and validating on the left-out seventh; iterated seven times across sectors) to determine (1) the optimal number of components; and (2) select the most informative microbial features (pMAGs) per component. The final model retained two components, with 350 and 180 microbes selected on the first and second component, respectively (**Appendix B: Fig. S8–S10; Tables S3–S5**). Results were visualised at the sample level with the sample plot, and heatmaps were used to show abundances of discriminatory microbes across sites, and the reproducibility of indicator microbes across sectors.

3.3.5 Correlating microbial abundance data with 54 continuous environmental variables

To further investigate environmental factors that may be driving microbial community patterns in NTMRs and fished reefs, we applied MINT Sparse Partial Least Squares (sPLS) analysis to identify important associations between 350 seawater microbes indicating reef zoning with continuous environmental data. Unlike MINT sPLS-DA, which is used for classification with a categorical outcome, MINT sPLS[292,294,295,295,296] integrates two continuous datasets (i.e., microbial abundances and environmental variables) while accounting for study-specific (i.e., GBR sectors) variation. First, median values were calculated per reef site for each of the 54 environmental variables (physico-chemical measurements and measures of benthic cover, fish assemblage abundances, and overall fish biomass) to account for differences in the number of replicates between microbial (n = 4) and environmental (n = 3) samples. Then,

MINT sPLS was used to identify important associations between 350 indicator seawater microbes (CLR-transformed abundances) and the 25 most covarying environmental variables. MINT sPLS results were visualised with a heatmap (displaying microbiome-environment associations) and a biplot reintroducing the context of the samples, thus visualising how well the latent components separate groups of interest (i.e., NTMRs and fished reefs).

3.3.6 Generalised Linear Mixed Models (GLMMs)

After identifying environmental variables significantly associated with microbial indicators of NTMRs versus fished reefs, we tested whether these key metrics themselves differed between protection zones using generalised linear mixed models (GLMMs). To analyse differences between NTMRs and fished areas, all models included protection status as a fixed effect (NTMR vs. fished) as well as the random terms of latitudinal sector and position across the continental shelf, with nested random effects including reef, site nested within reef and transect nested in site. Herbivore density (abundance numbers per 1000 m^2) was modelled against a negative binomial distribution to handle overdispersion, total fish biomass with a Gamma distribution with a log link function. Hard coral and turf algae were modelled against binomial distributions representing the number of points classified as hard coral or turf algae out of the total number of points surveyed. All models were implemented in R using the glmmTMB (v1.1.10) package[297], with significance assessed via Wald z-tests ($\alpha = 0.05$) and model diagnostics confirming appropriate fit using the DHARMa (0.4.7) R package[298].

3.3.7 Network Connectedness and Cohesion

Network connectedness and cohesion are defined as network metrics that quantify the degree of positive and negative connectivity within a microbial community, and were computed as in Herren and McMahon. (2017)[299]. Briefly, connectedness (microbe-specific metric) is derived by first computing Pearson pairwise correlations between taxa from relative abundance data. To correct for biases inherent in compositional data, a null model (e.g., "taxon shuffling") randomises abundances for all taxa except the focal one across 200 iterations, generating expected correlations that are subtracted from observed correlations. The resulting corrected correlations are then averaged separately for positive and negative values, yielding taxon-specific positive and negative connectedness with the broader community. Cohesion (sample-specific metric) is then calculated for each sample by weighting the relative abundance of each microbe by its connectedness values and summing these products across all microbial taxa, producing two metrics per sample (i.e., community): positive cohesion (sum of positive contributions) and negative cohesion (sum of negative contributions). In contrast to existing correlation detection methods which aim to identify significant pairwise associations (i.e., between two taxa), connectedness and cohesion therefore evaluate connectivity at the community level, and were used in our study to compare the prevalence of positive (i.e., mutualism, commensalism, co-dependency due to metabolic exchange) and

negative (i.e., competition, predator/prey, pathogen/host, parasite/host, and etc.) correlations in reef bacterioplankton between NTMRs and fished reefs. It should be noted however that these correlations may also reflect shared microbial responses to the same environmental drivers or, and are not a direct proxy of direct biological interactions.

Results were visualised as barplots for connectedness (expressed as positive/negative edges ratio, for the 350 indicator pMAGs discriminating NTMRs and fished reefs) and as boxplots for cohesion (expressed as positive/negative cohesion ratio). For each of the 350 indicator pMAGs, linear regression was used to find a relationship between positive/negative connectedness ratio (as a response variable) and (1) genome size, (2) GC content, and (3) potential for metabolic independence (expressed for each pMAG as an average completeness across obtained KEGG modules, similar to Veseli et al. (2024)[300]) as the three predictor variables.

3.3.8 Predictions of continuous environmental variables using seawater microbial data

To identify if microbial markers can predict environmental variables in the surrounding seawater, Random Forest (RF) models[301] were trained on CLR-transformed microbial abundances to predict continuous environmental response variables (physicochemical measurements, benthic cover, and fish abundance and biomass). RF models (500 trees; mtry = $\sqrt{p}$, where p = number of microbial features; node size = 5) were validated using stratified (strata = GBR sectors) and site-aware (all replicates from a given site were grouped within each split) 80/20 train-test splits repeated 50 times, and with fixed random seeds for reproducibility. This repeated subsampling approach provides robust performance estimates by mitigating geographic (sector and site-specific) and temporal biases, while also avoiding information leakage (i.e., training the data on a subset of site replicates and validating on left-out replicates of the same site). RF models were implemented via the randomForest (v4.7-1.2) R package[302].

For each environmental variable, model performance was quantified using mean R^2 values (i.e., by aggregating results across 50 permutations). While no universal R^2 interpretation exists for marine microbiome-environment prediction studies, we defined RF model predictive performance as high ($R^2 \geq 0.6$), moderate ($0.3 \leq R^2 < 0.6$), or poor ($R^2 < 0.3$) based on existing literature[303–305]. Predictor importance was quantified using the percentage increase in mean squared error (%IncMSE), with higher values indicating stronger predictive contributions (as shuffling the values of that microbe significantly lowered the RF model's performance). Microbe-specific %IncMSE scores were averaged across 50 permutations, and the top (i.e., the highest average %IncMSE) 50 microbes per variable were selected.

3.3.9 Inferring microbial niches

To further validate these predictions, we computed the microbial niche tolerance ranges for consensus markers, defined as the 25th–75th percentile of environmental values beyond which that

microbe is typically not observed. Microbial niches were calculated following[306], using the robust optimum (RO) method[307]. In this approach, the ecological optimum (Q2) represents the ideal living conditions for a microbe concerning a given environmental parameter (i.e., where a specific pMAGs will be found at its highest relative abundance), whereas the lower (Q1) and upper (Q3) niche bounds correspond to the minimum and maximum values of that environmental variable beyond which the microbial taxon is rarely observed. These bounds represent the environmental ranges within which each microbial taxon can be found at varying levels of relative abundance, and for each microbe, its tolerance (niche) range for a specific environmental variable is calculated as the interquartile range (Q3–Q1). Microbial niches (i.e., Q1, Q2, and Q3 values for each of the 876 pMAGs, and for each of the 54 environmental variables) were computed three times to estimate global microbial niche across all samples, and separately for NTMRs and fished reefs.

Relative abundance data for each of the 876 microbial pMAGs was normalised in a two-step manner: by introducing pseudocounts (i.e., adding 1 to abundance values to avoid issues with zero counts) and then dividing abundances by the geometric mean, scaling the data to account for variations between microbial features and ensuring the data was standardised across samples. Based on obtained relative abundance distributions across samples, niche tolerance ranges were computed for each of the 876 pMAGs—specifically the lower bound (Q1, i.e., 25th percentile), ecological optimum (Q2, i.e., 50th percentile), and upper bound (Q3, i.e., 75th percentile)—relative to each of the 54 continuous environmental variables. Niches were computed only based on the samples where the microbial feature was present, which was done by identifying prevalent samples where the microbial feature was observed (i.e., where the abundance value of that pMAG was greater than zero), and the environmental variables associated with these prevalent samples were then extracted to compute microbial niche bounds. Finally, the resulting niche bounds (Q1, Q2, and Q3) for each microbial feature (876 pMAGs) and each of the 54 continuous environmental variables were compiled into a comprehensive dataset, providing detailed information on the optimal environmental conditions for each microbe. We visualised the niche tolerance ranges for microbial predictors identified in RF models using boxplots generated within ggplot2[308] (v3.5.1).

3.4 Results and Discussion

3.4.1 Stable seawater indicators of GBR zoning status across season and geography

No-Take Marine Reserves (NTMRs) across the GBR have well-documented benefits for reef macroorganisms[249,254,255], however, less is known about how reef zoning influences surrounding bacterioplankton communities. To investigate whether seawater microbes can serve as indicators of reef zoning status, we assessed their ability to classify NTMRs and fished reefs across the GBR. Zoning status was consistently predicted across 48 reefs spanning seven sectors of the GBR (**Fig. 3.2a**), with an average

classification accuracy of ~71% (**Appendix B: Fig. S8, S11; Table S3-5**). This was achieved using a subset of 350 indicator pMAGs (**Fig. 3.2b**) identified through MINT sPLS-DA[291,292,294] as stable microbial markers of reef zoning that were spatially and temporally reproducible (**Fig. 3.2c; Appendix B: Fig. S10**).

Taxonomic analysis of these indicators revealed consistent differences in microbial assemblages between NTMRs and fished reefs. Of the 350 indicator pMAGs, 236 were consistently enriched in NTMRs (**Fig. 3.2b**; **Fig. 3.2d**, green) with 63 belonging to the order Pelagibacterales (of which 62 were Pelagibacteraceae, 44/63 *Pelagibacter*). Additional indicator taxa were classified within orders TMED127 (in the class Alphaproteobacteria) and Flavobacteriales, with families TMED127 (including genus GCA-002690875) and Flavobacteriaceae being the second most abundant indicator taxa of NTMRs, each with 17 pMAGs. The orders Poseidoniales (25 indicators), Marinisomatales (15 indicators) and SAR86 (14 indicators) were other notable taxa indicative of NTMRs. In contrast, 114 pMAGs were consistently more abundant on fished reefs compared to NTMRs (**Fig. 3.2b**; **Fig 3.2d**, blue). Of these, 44 indicators belonged to the order Flavobacteriales with 19 classified as Flavobacteriaceae (10 of which were classified as genus *Arcticimaribacter*, making this the most numerically abundant genus level indicator alongside UBA11663 in the family Sanyastnellaceae), and 16 from the family UA16 (at genus level: 10 from UBA11663, 2 from UBA8752, and one not identified to genus level) (**Fig. 3.2d**, blue). Other notable taxa indicative of fished reefs included 19 members of the archaeal order Poseidoniales (6 each belonging to the genera *Poseidonia* and MGIIa-L1), 18 in the order Puniceispirillales, 7 in the order Pseudomonadales and 4 in the Parvibaculales.

The average classification accuracy differed between reef zones (**Appendix B: Table S4**) and GBR sectors (**Appendix B: Fig. S10; Table S3, S5**). NTMR reefs were more accurately predicted compared to fished reefs (75% vs 68% accuracy; **Appendix B: Table S4**), a difference that may reflect variation in protection levels among the sampled fished reefs. Specifically, 20 reefs were sampled in the less protected dark blue 'Habitat Protection' zone, where most fishing is allowed except trawling, versus five reefs in the more protected yellow 'Conservation Park' zone[271] (**Fig. 3.1**). While finer-scale comparisons of protection levels (e.g., green vs. dark blue vs. yellow zones) could offer additional insights, the limited number of reefs sampled in the yellow zone (n=5) precluded robust statistical comparisons at this granularity.

Among sectors, the highest prediction accuracy at 85% was detected in the Innisfail sector followed by Princess Charlotte Bay (80%), while the lowest were observed for Cape Grenville and Swains sectors (50% and 55%, respectively; **Appendix B: Table S3**), which may be attributed to particular characteristics of these sectors. In Cape Grenville, reefs are located close to the coast and subject to influences from the Torres Strait and northern Australia, including the presence of fish species with northern-truncated distributions that extend from the north and north-west Western Australia[309]. Additionally, hydrodynamic characteristics in Cape Grenville differ from those further south due to a predominant northerly flow from incoming Coral Sea currents north of Cape Flattery[310,311], and fishing pressure is lower in the Far-North GBR due to its remoteness[312]. Poor zoning prediction within the Swains

is likely also due to the remoteness of this sector, with reefs situated between 150 and 250 km from the coast, further offshore than any other GBR reefs and therefore subjected to minimal coastal influence[249]. Fish and benthic community composition are also somewhat distinct in the Swains compared to other GBR regions[309,313], and oceanographic processes including southerly transport from the East Australian Current lagoonal branch and frequent upwelling events[314] may have contributed to distinct environmental conditions and thus poor zoning predictions in the Swains. Despite these sector-specific variations, our study demonstrates the potential of seawater microbes as stable and consistent indicators of reef zoning across the GBR (**Fig. 3.2c; Appendix B: Fig. S10–11**), with distinct microbial taxa associated with NTMRs and fished reefs, highlighting their utility for monitoring zoning status across broad spatiotemporal scales.

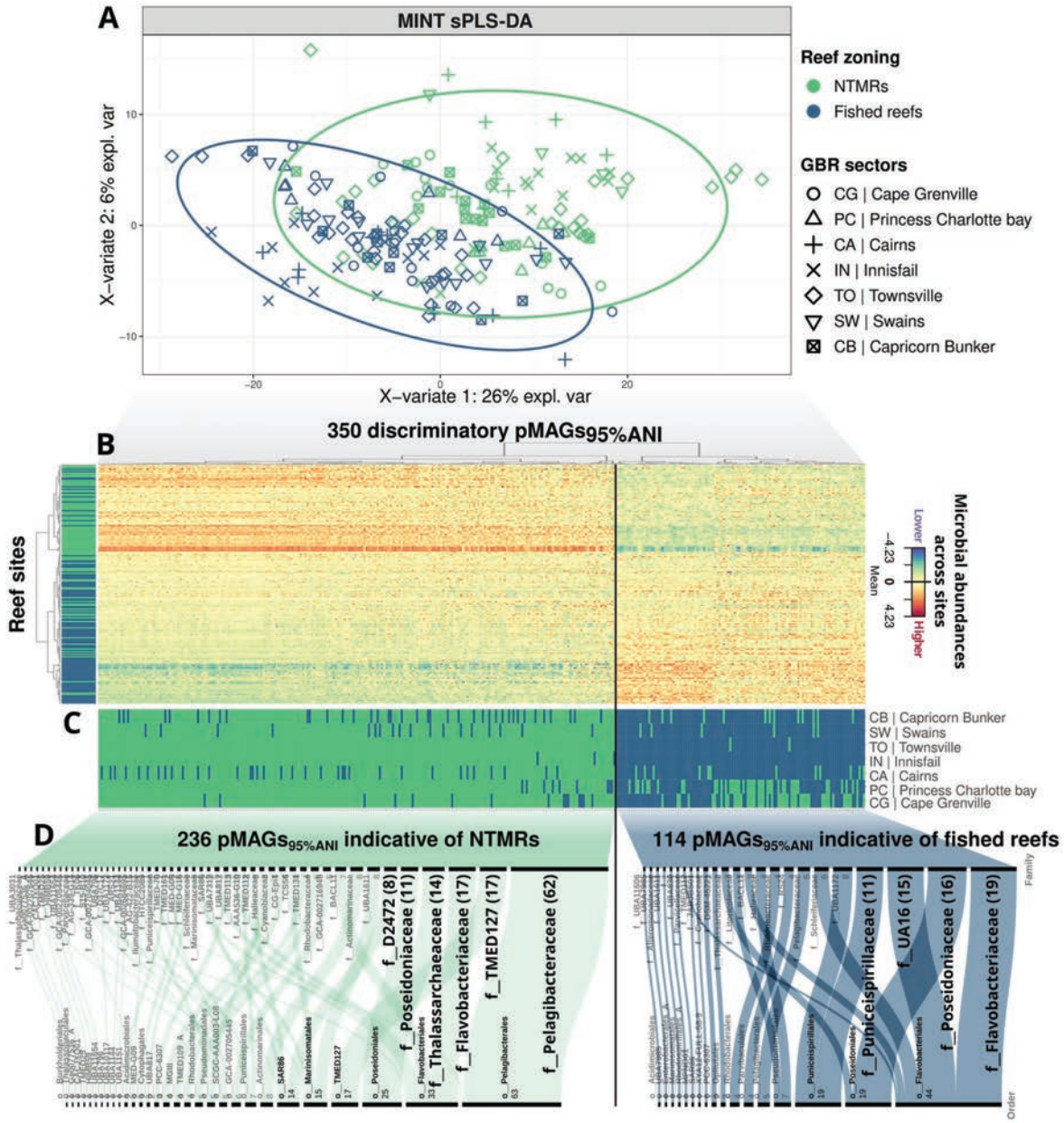

Figure 3.2: Visualisation of seawater microbes selected to distinguish NTMRs from fished reefs across seven GBR sectors. (A) Sample plot from MINT sPLS-DA showing clustering of reefs by reef zoning (fished vs NTMRs) based on indicator seawater prokaryotic metagenome-assembled genomes (pMAGs). Samples (48 reefs x four replicates) are projected in the first two components of the MINT sPLS-DA space, with ellipses indicating 95% confidence level. (B) Heatmap shows differential abundances for the pMAGs (columns) discriminating between NTMRs and fished reefs (rows). Euclidean clustering of microbes by abundance across reef sites splits the heatmap into two groups: 236 pMAGs indicating NTMRs (left), and 114 microbial indicators of fished reefs (right). (C) Heatmap indicating whether the pMAGs were more relatively abundant in NTMRs (green) or fished reefs (blue) in each GBR sector. (D) Alluvial diagrams summarise the taxonomy of indicator microbes for NTMRs (left) and fished (right) reefs, ordered by the most common taxa for each zone.

3.4.2 Linking environmental variables with indicator microbes

Environmental drivers that shape the differences in microbial assemblages between NTMRs and fished reefs were identified through an integrated analysis of reef bacterioplankton and ecosystem conditions at these offshore GBR sites. MINT sPLS analysis[294,295,295,296] identified that microbial indicators enriched in NTMRs were more abundant and correlated with environmental variables indicative of healthy reefs[315], including elevated fish biomass, increased coral and crustose coralline algae (CCA) cover, reduced turf algae, and lower nutrient concentrations (**Fig. 3.3a–b**; **Appendix B: Table S6**). In contrast, microbial indicators of fished reefs were enriched under environmental features characteristic of degraded reef conditions[266,316], including lower fish abundances, high turf-to-coral ratios, and nutrient enrichment (**Fig. 3.3a–b**; **Appendix B: Table S6**). Interestingly, even though the MINT sPLS approach is unsupervised and does not incorporate reef zoning information, it nevertheless revealed distinct clustering of microbial communities by zoning, with NTMR reefs (e.g., Myrmidon, Moore, and Hastings reefs) clustering separately from fished sites (e.g., Farquharson, John Brewer, and Masthead reefs), independent of reef sector (**Fig. 3.3a**).

The establishment of NTMRs in 2004 increased fish biomass in the GBR[252–254], a trend also reflected in our data, where biomass was approximately 1.28× greater in NTMRs (mean 65.57 ± S.E. 73.55 kg 1000 m^{-2}) than in fished reefs (mean 51.28 ± S.E. 41.34 kg 1000 m^{-2}; z=2.371, p=0.018; **Appendix B: Fig. S12a; Tables S7–S9**). While herbivorous fish densities (**Appendix B: Fig. S12b; Tables S10–S13**) and the cover of hard corals (**Appendix B: Fig. S12c; Tables S14–S17**) were not significantly different between NTMRs and fished reefs across all sites, fish abundances (including herbivores) and overall biomass were collinear and elevated together with CCA and hard coral cover in a subset of NTMR reefs (**Fig. 3.3a**; component 1), explaining why NTMR-associated microbial indicators were positively linked to all these variables simultaneously (**Fig. 3.3b**). Specifically, the 236 indicators enriched in NTMRs (for taxonomic annotations, refer to **Fig. 3.2d**, green) were positively associated with higher abundances of multiple fish groups including detritivores, corallivores, herbivores, invertivores, and piscivores (**Fig. 3.3b**; green cluster; **Appendix B: Table S6**). These NTMR-enriched microbes also showed positive relationships with hard coral cover, including different groups of *Acropora* (digitate, tabulate and encrusting), CCA and the hydrozoan *Millepora,* and were negatively associated with turf algae cover (**Fig. 3.3b**; green cluster; **Appendix B: Table S6**). Fish and coral are likely collinear because corals enhance reef 3D complexity to support diverse fish assemblages[317–320]. Additionally, elevated counts of herbivorous fish in some NTMR reefs (**Fig. 3.3a**; component 1) may have resulted in qualitative differences in grazing pressure and increased algae removal in NTMRs, thereby creating more space for coral and CCA to settle[309,321,322].

Microbes enriched in NTMRs were also correlated with reduced nutrient concentrations, including particulate organic matter (POM) and dissolved inorganic nitrogen (specifically NO_2 and NO_3, collectively termed NOx), while showing a positive association with dissolved organic carbon (DOC) (**Fig. 3.3a–b**). The negative associations between coral cover and water column nutrients (POM and NOx

concentrations, in addition to lower Chl-*a* and Phaeo; **Fig. 3.3a–b**) may reflect nutrient assimilation by coral and other benthic organisms[323] in addition to reduced POM remineralisation into NOx due to the lower POM concentrations[324]. Corals supplement their nutrition through heterotrophic feeding on phytoplankton, zooplankton, and detritus[323], with literature proposing that this efficient uptake of nutrients (including NOx) maintains the high productivity of reef ecosystems despite residing in nutrient-poor waters[263]. Viewed through the lens of the reef "microbialization" hypothesis, lower microbial depletion of DOC in NTMRs and the fact that DOC there likely originates from hard coral cover (given its collinearity with tabulate and digitate Acroporans and non-Acropora groups; Fig. 3B), suggests that microbial activity may be more tightly coupled to particulate and inorganic nutrient pools than to labile DOC. This is consistent with coral-dominated systems where DOC production is lower and microbial respiration is less stimulated than on algal-dominated reefs, where enhanced microbial respiration of DOC exudates from algae (including turfs) results in DOC drawdown[316]. Overall, nutrient cycling appears to be more efficient in protected coral-dominated reefs where higher benthic and fish biomass efficiently removes or limits the accumulation of organic particulates[325]. This aligns with findings from highly protected, near-pristine reefs like Jardines de la Reina, Cuba[326], where the most strictly enforced protection zones resulted in some of the highest fish biomass and coral cover in the Caribbean[326], maintaining oligotrophic conditions and picoplankton microbial communities with the highest alpha diversity and dominated by oligotrophic taxa like SAR11 and *Prochlorococcus*, despite geographic factors (i.e. site-specific patterns) being the primary driver of community variation[327].

Conversely, fished reefs with reduced herbivore counts may experience diminished grazing[328], facilitating turf algae proliferation[266,316,329]. While turf algae cover was not significantly different between NTMRs and fished reefs when considering all sites (**Appendix B: Fig. S12d**; **Tables S18–21**), it was significant in some reefs (**Fig. 3.3a**, component 1), explaining why the 114 microbial indicators of fished reefs positively associate with these metrics collectively (**Fig. 3.3b**, blue cluster; **Appendix B: Table S6**). With respect to nutrient dynamics, microbial indicators of fished reefs were negatively correlated with DOC, suggesting enhanced microbial uptake of labile DOC in reefs open to fishing, likely driven by heterotrophic microbial processes and consistent with early stages of reef microbialization[316]. DOC uptake may fuel microbial biomass production, potentially contributing both to POM accumulation and the release of NOx (as a byproduct of microbial metabolism) into the water column[330]. To explore the hypothesis that POM in reefs open to fishing is partly of microbial origin, we examined POC:PN ratios relative to the Redfield ratio, a well-established benchmark in oceanography describing the canonical C:N composition of marine phytoplankton[331]. While the average POC:PN ratios were close to the Redfield ratio of 6.6, suggesting a largely phytoplanktonic origin of POM across all reefs, the wider range in POC:PN values observed in fished reefs (**Appendix B: Fig. S14**) suggests greater variability in POM composition. This could reflect more site-specific or diverse POM inputs in fished areas (i.e., some sites had N-enriched POM, consistent with microbially processed or more labile detritus), potentially due to altered trophic dynamics and reduced detritivore fish abundance (**Fig. 3.3b**), which may affect detritus consumption

rates[332]. As further evidence for enhanced microbial remineralisation in fished reefs, higher DIN:DIP ratios were observed in fished reefs (9.9 ± 7.9) relative to NTMRs (7.2 ± 5.5; **Appendix B: Fig. S13**), suggesting more efficient nitrogen utilisation in NTMRs vs. enhanced release and retention of NOx in fished reefs. While the exact cause of this pattern remains uncertain, it could reflect both bottom-up (nutrient availability) and top-down (fishing pressure) effects on nutrient dynamics.

Our results provide support that marine microbes can be used as indicators of benthic-pelagic coupling in reef ecosystems, with the synergistic effects of reduced herbivory, declining hard coral cover, and increased turf algae collectively altering nutrient cycling and driving divergent microbial assemblages between some NTMRs and fished reefs. Most notably, the influence of these benthic-pelagic couplings are measurable in the microbial data despite subtle changes in individual parameters, confirming the proposed sensitivity of seawater microbes to integrated ecosystem states[333]. Given that microbial indicators proved informative even under the GBR's relatively low fishing pressure when compared to other reef ecosystems globally[271,312], their utility may be even greater in reef ecosystems where fishing is a dominant driver of ecosystem degradation. Such systems could leverage microbial monitoring as a sensitive, early-warning tool for changing trophic cascades and benthic community shifts.

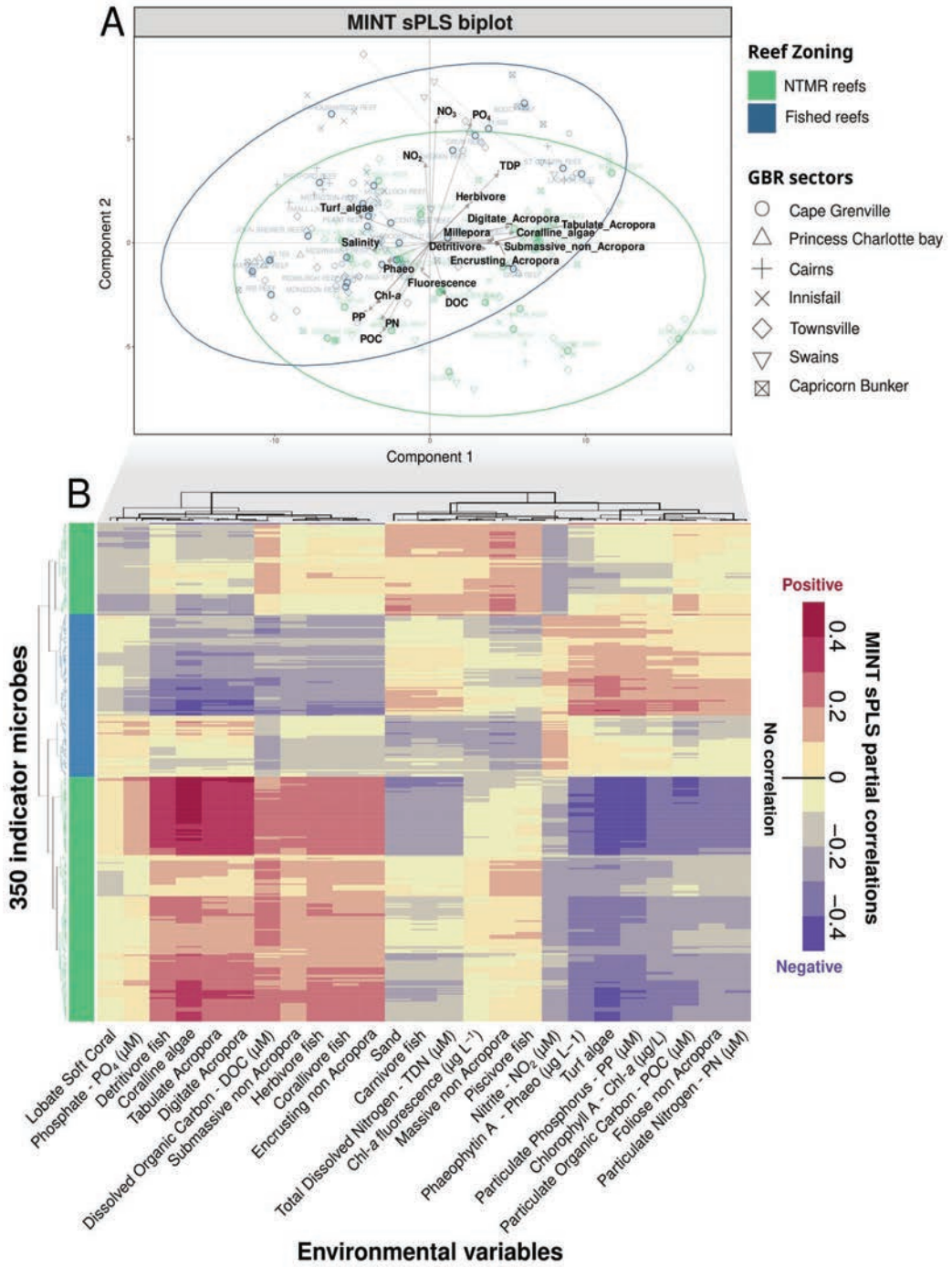

Figure 3.3: MINT sPLS integrates abundances of seawater microbes with continuous environmental data, while accounting for GBR sector-specific variation. (**A**) The biplot from MINT sPLS shows both samples (four replicates per reef; hollow circles represent site centroids connecting the four site replicates) and variables (environmental only, in black; microbial variables omitted for clarity) on the same plot. (**B**) A heatmap shows the MINT sPLS partial correlations between 350 pMAGs identified in MINT sPLS-DA as reef zoning indicators (rows; colored based on indicator status) and 25 most influential environmental variables (columns).

3.4.3 Microbial indicators prevalent in NTMRs harbor signatures of genome streamlining via gene loss

NTMR-enriched microbes exhibited genomic features consistent with microbial streamlining, an adaptation that facilitates survival in nutrient-poor pelagic systems[334–336]. Specifically, NTMR-enriched microbes such as Pelagibacterales, SAR86, and Marinisomatales (see **Fig 3.2d**, green) had significantly ($p < 0.0001$) smaller genomes (~1.57×; **Fig. 3.4a-b**) and lower GC content (~1.62×; **Fig. 3.4c-d**) compared to microbes prevalent in fished reefs, which were dominated by Flavobacteriales UA16 and Schleiferiaceae (**Fig. 3.2d**, blue). These genomic signatures are consistent with adaptation to oligotrophic conditions[334] and align with the lower nutrient concentrations measured in NTMRs (e.g. ~1.50–1.71× lower NO_2 and NO_3; **Appendix B: Fig. S15, Table S22**). NTMR indicator pMAGs additionally exhibited ~1.53× lower average KEGG module completeness compared to the 114 microbes indicative of fished reefs (**Fig. 3.4e-f**), consistent with gene loss and reduced metabolic capacity, while microbial indicators enriched on fished reefs appear to have an increased pathway completeness and thus may be more metabolically independent[300].

A key consequence of genome reduction and gene loss is that many streamlined microbes lose the ability to synthesise certain essential metabolites, rendering them dependent on neighboring community members for these compounds[337]. Cooperative interactions among free-living microbes are central to the Black Queen Hypothesis[338], which posits that oligotrophic microbes will form more positively connected communities due to enhanced mutualistic metabolic exchanges[336,338,339]. Using species co-occurrence network analysis[299], we found that NTMR indicator pMAGs with streamlined genomes were indeed ~1.61× more positively connected to the broader community (**Fig 3.4g–h**, green). In contrast, fished reef microbial indicators showed lower positive-to-negative connectedness ratios (**Fig 3.4g–h**, blue), potentially implying more antagonistic microbial interactions such as competition, amensalism, or parasitism[340]. Supporting the predicted genomic underpinnings of these network patterns, regression analysis revealed significant (albeit weak) negative correlations between positive-to-negative connectedness and genome size ($\beta = -2.67 \times 10^{-7}$, $R^2 = 0.096$, $p < 0.001$), GC content ($\beta = -0.021$, $R^2 = 0.143$, $p < 0.001$), and KEGG module completeness ($\beta = -0.265$, $R^2 = 0.078$, $p < 0.001$) (**Appendix B: Fig. S20a–c**). This further supports the pattern that genome streamlining (**Fig. 3.4a–d**), marked by gene loss and reduced metabolic capacity (**Fig. 3.4e–f**), is linked to increased cooperative microbial interactions in NTMRs (**Fig. 3.4g–h**).

Fished reefs showed stronger signatures of antagonistic microbial interactions across entire reef bacterioplankton communities, with co-occurrence networks (inferred from all 876 pMAGs) exhibiting ~12.9% lower positive:negative cohesion ratios (though with high variability; ±50% S.E.; **Fig. 3.4i**) compared to NTMRs. This pattern was consistent across GBR sectors (**Appendix B: Fig. S16**) suggesting a systemic shift toward negative (i.e., mutually exclusive) interactions in fished reef seawater microbiomes. As reef bacterioplankton indicators of fished reefs were shown to increase in prevalence

with elevated nutrient concentrations (NO_2^-, NO_3^-, and POM; see **Fig. 3.3b**), this microbial competition may be linked to competition for available nutrients. Correspondingly, negative microbe-to-microbe interactions were also more prevalent during the austral summer (**Appendix B: Fig. S17**) when all nutrients apart from phosphate were elevated (**Appendix B: Fig. S18**). These observations align with previous findings from oligotrophic reefs in Porto Seguro (Bahia, Brazil) showing that reefs further from a river mouth and exposed to less organic pollution harboured more complex microbial networks, with a higher proportion of positive associations in the more pristine oligotrophic reefs[339]. The increased microbial positive connectedness in NTMRs (**Fig. 3.4g–i**) may also be the reason why we detected a higher number of indicator microbes for NTMRs than for fished reefs (236 vs 114 markers; **Fig. 3.2**), as more co-dependent microbes are stably associated and thus more consistently enriched, potentially facilitating better predictions of NTMRs than fished reefs (~7% increase in classification accuracy; **Appendix B: Table S4**). In contrast, the enrichment of pMAGs on fished reefs may reflect competitive dominance at the expense of other microbes, likely contributing to the ~5% lower overall microbial diversity observed relative to NTMRs (**Fig. 3.4k**, consistent across most GBR sectors; **Appendix B: Fig. S19**). This is in line with findings by Hernandez et al. (2021) that negative microbial interactions strongly predict reduced microbial diversity[340].

Graph theory is increasingly used to quantify stability in microbial co-occurrence networks[306,341]. Recent frameworks[340] suggest that healthy microbial communities show stronger partitioning (high connectivity within network modules, i.e., tightly linked microbial subgroups), which limits environmentally induced shifts within these modules. In contrast, unstable microbial communities (undergoing stress) are predicted to exhibit higher between-module connectivity which was proposed as a metric of destabilised networks[340]. Environmental fluctuations may propagate widely across the community, potentially resulting in network collapse if critical thresholds are exceeded[306]. In addition to a minor decrease in microbial diversity (**Fig. 3.4k**), species co-occurrence networks in fished reefs also showed lower (albeit not significant) modularity scores (0.71 vs. 0.73 in NTMRs; **Fig. 3.4j**), suggesting higher between-module connectivity and potentially less stable (modular) communities in fished reefs compared to NTMRs. Further experimental validation, such as controlled manipulation of nutrient levels in mesocosms coupled with microbial interaction assays or stable isotope probing to track metabolic exchanges, is required to explore this hypothesis.

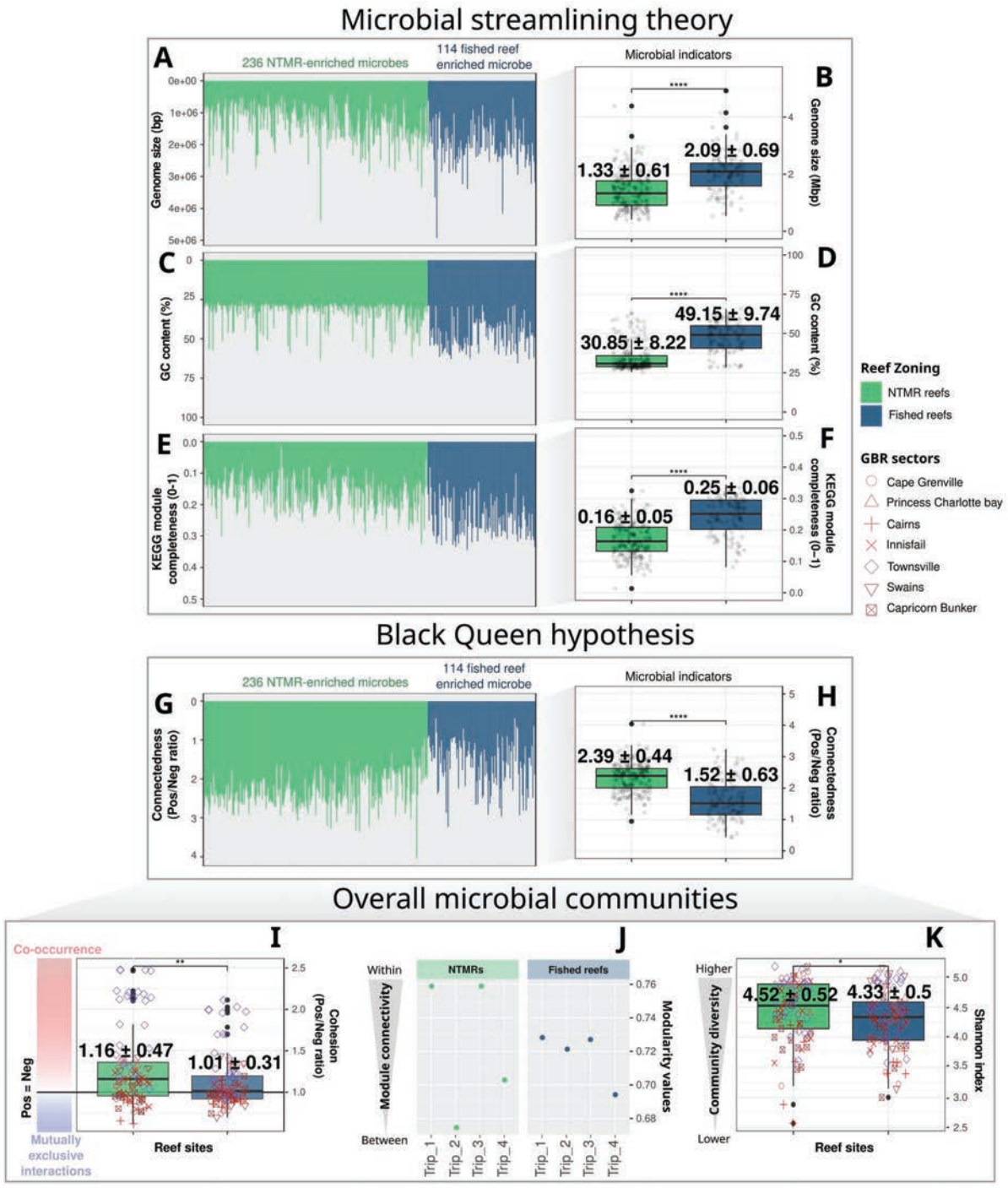

Figure 3.4: Genomic, functional, and network characteristics of indicator microbes associated with NTMRs (green) and fished reefs (blue). Genomic features (A–D): Barplots show genome size (A) and GC content (C) for indicator pMAGs, with boxplots comparing their distributions between NTMRs and fished reefs (B, D). Lower values are considered signatures of genome streamlining. Functional potential and interaction structure (E–I): (E) Barplots show average KEGG module completeness per pMAG, with (F) corresponding group-level comparisons. Lower values indicate gene loss and incomplete pathways. (G) Microbe-specific ratios of positive to negative co-occurrence edges represent cooperative versus antagonistic interactions, and (H) shows group-level comparisons of these ratios. (I) Sample-level ratios of positive to negative cohesion illustrate co-occurrence within overall

bacterioplankton communities. Higher positive to negative ratios (both for connectedness and cohesion; G-I) indicate a prevalence of positive interactions, while lower values are a proxy of negative interactions. Diversity and community structure (J–K): (J) Network modularity scores for microbial networks between reef zones, with higher and lower modularity values indicating increased connectivity 'within' and 'between' modules, respectively. (K) Alpha diversity (Shannon index) is shown for seawater microbiomes from NTMR and fished reef sites. For boxplots (B, D, F, H, I, K), significance levels from Wilcoxon rank sum tests are indicated as: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; "ns" = not significant.

3.4.4 Microbes enriched on fished reefs are functionally primed for rapid carbohydrate uptake, energy production, and biosynthesis of complex compounds

To test if microbial indicators of fished reefs possess metabolic traits to compete for nutrients, hypothesised based on their associations with elevated nitrate, nitrite, and POM (**Fig. 3.3**), and network properties indicating microbial antagonism (**Fig. 3.4**), we compared the completeness of metabolic pathways (KEGG modules) between fished-reef and NTMR microbial indicators. Microbial indicators of fished reefs overall showed an enhanced potential for rapid nutrient utilisation as they had higher completeness in several metabolic pathways involved in carbohydrate metabolism, energy generation, and anabolic reactions, including biosynthesis of cofactors, vitamins, amino acids, and lipids (**Fig. 3.5; Appendix B: Figs S22-S28**). This reflects true biological variation as genome completeness estimates (**Appendix B: Fig. S21**) and reference genomes were consistent with predicted genome size (and associated metabolic capacity) variation.

Compared to NTMR-enriched pMAGs (e.g., Pelagibacterales, SAR86; **Fig. 3.2d**, green), the 114 microbes prevalent on fished reefs (primarily Flavobacteriales; **Fig. 3.2d**, blue) had more complete pathways for sugar metabolism, including the pentose phosphate, Entner-Doudoroff, and galactose degradation pathways (**Fig. 3.5**; **Appendix B: Fig. S22**). This may be reflective of ecological changes between NTMRs and fished reefs (though minor in our data) influencing the quantity and composition of organic matter in reef waters, potentially creating conditions that favour microbial assemblages capable of exploiting more labile and diverse carbon sources in fished reefs. Fished-reef microbial indicators likely process sugars derived from benthic algae, inferred from the collinearity between nutrients and turf cover (**Fig. 3.3a–b**) and the fact that our sampling sites are located on oblique or exposed reef slopes where seagrass is largely absent, although other organic matter sources (e.g., phytoplankton, detritus) may also contribute. Future work (e.g., isotopic tracing) will be necessary to confirm nutrient provenance and disentangle the relative contributions of these potential sources.

Microbes enriched on fished reefs also showed enhanced energy-generating capacity (**Figs 3.5; Appendix B: S23**), with more complete carbohydrate utilisation pathways (M00004, M00308, M00008, M00631, M00632) converging on glyceraldehyde-3-phosphate for glycolysis[342]. Alongside enhanced acetyl-CoA synthesis (pyruvate oxidation M00307; β-oxidation M00087; **Fig. 3.5**), these pathways likely fuel the tricarboxylic acid cycle for ATP generation[342], suggesting fished-reef seawater microbial assemblages are

restructured for rapid energy harvest from organic substrates. This metabolic configuration supports their shift toward anabolic pathways, including the biosynthesis of lipids (**Fig. 3.5; Appendix B: S24**), amino acids such as serine, methionine, proline, and tryptophan (**Fig. 3.5; Appendix B: S25**), and cofactors like vitamin B12 (cobalamin; **Fig. 3.5; Appendix B: S26**). Through *de novo* synthesis of essential macromolecules, fished-reef microbial indicators may exhibit enhanced metabolic independence[300] providing a mechanistic explanation for the competitive advantage of fished-reef microbial indicators over other microbes (**Fig. 3.4g-i**). For example, the Entner–Doudoroff (ED) pathway, which is often more prevalent under elevated nutrient conditions, and the pentose phosphate pathway provide mechanisms by which dissolved organic carbon can be remineralised more quickly and less efficiently than through the Embden–Meyerhof–Parnas (glycolytic) pathway, which is more common in healthier coral-dominated reefs where nutrient levels are lower. This metabolic shift, known as the 'yield to power' switch[316], enables copiotrophic and opportunistic microbes to outcompete others by rapidly exploiting available resources[316,329]. Further, the ocean is globally depleted of vitamin B12 (cobalamin) despite it being a major cofactor required by most marine microbes[343,344]. Interestingly, many of the 114 pMAGs prevalent on fished reefs showed functional potential for cobalamin biosynthesis both via aerobic (cob genes; KEGG modules M00122 and M00925) and anaerobic (cbi genes: M00924) pathways (**Fig. 3.5**: Metabolism of cofactors and vitamins), further reinforcing their functional autonomy, environmental adaptability, and potential to outcompete other microbes.

A shift towards anabolic metabolism was proposed as a key mechanism driving 'microbialization' on degraded reefs, linked with fishing pressures, macroalgae overgrowth and changed reef water chemistry that select for microbial biomass accumulation and abundance shifts towards higher copiotrophic and potentially pathogenic seawater microbes[259,316,329,345,346]. This enrichment of seawater microbial copiotrophs is not limited to chronic macroalgal dominance, with recent evidence showing "microbialization" also occurs following other severe disturbances like mass bleaching episodes, where thermally stress bleached corals release labile dissolved organic matter (DOM) into the water column, significantly increasing bacterioplankton growth and abundances of copiotrophic and putatively pathogenic microbes[347]. Interestingly, fishing pressure in the offshore GBR is comparatively low due to its remoteness[271,312], hence our findings suggest that microbialization may represent a more universal ecological response, with microbial shifts similar to the DDAM (dissolved organic carbon, disease, algae, microorganism) model predictions also emerging under comparatively milder forms of disturbance. To test if there is greater microbial biomass in these fished reefs, future studies should collect microbial count data using readily obtainable field methods such as flow cytometry[333].

This picture of oligotrophic taxa in NTMRs and opportunistic microbes in fished reefs also suggests a restructuring of nutrient cycling and energy flow between reef zones. For example, despite positive correlations with dissolved nitrogen (**Fig. 3.3a-b**), microbial indicators of fished reefs lacked canonical transporter genes for nitrate (NRT1 and NRT2 family transporters), nitrite (NrtA/NrtB systems,

NirC, and the ABC transporters NrtC/D), and ammonium (including Rh family proteins SLC42A, and Amt/MEP family proteins) (**Appendix B: Fig. S29**). Thus, microbial indicators of fished reefs may be associated with higher NOx (**Fig. 3.3b; Appendix B: Fig. S30**) not due to direct uptake, but because they drive remineralisation of (particulate) organic matter to produce NOx (as well as ammonium; **Appendix B: Fig. S30**), consistent with the higher DIN:DIP ratio in fished reefs (9.9 ± 7.9) compared to NTMRs (7.2 ± 5.5; **Appendix B: Fig. S13**). Further research is warranted to confirm this hypothesis of enhanced microbial decomposition of POM into dissolved nitrogen in fished reefs, in addition to understanding how these taxonomic (**Fig. 3.2**) and functional (**Fig. 3.5**) reef bacterioplankton shifts between NTMRs and fished zones cascade through reef productivity, carbon cycling networks, and overall reef health.

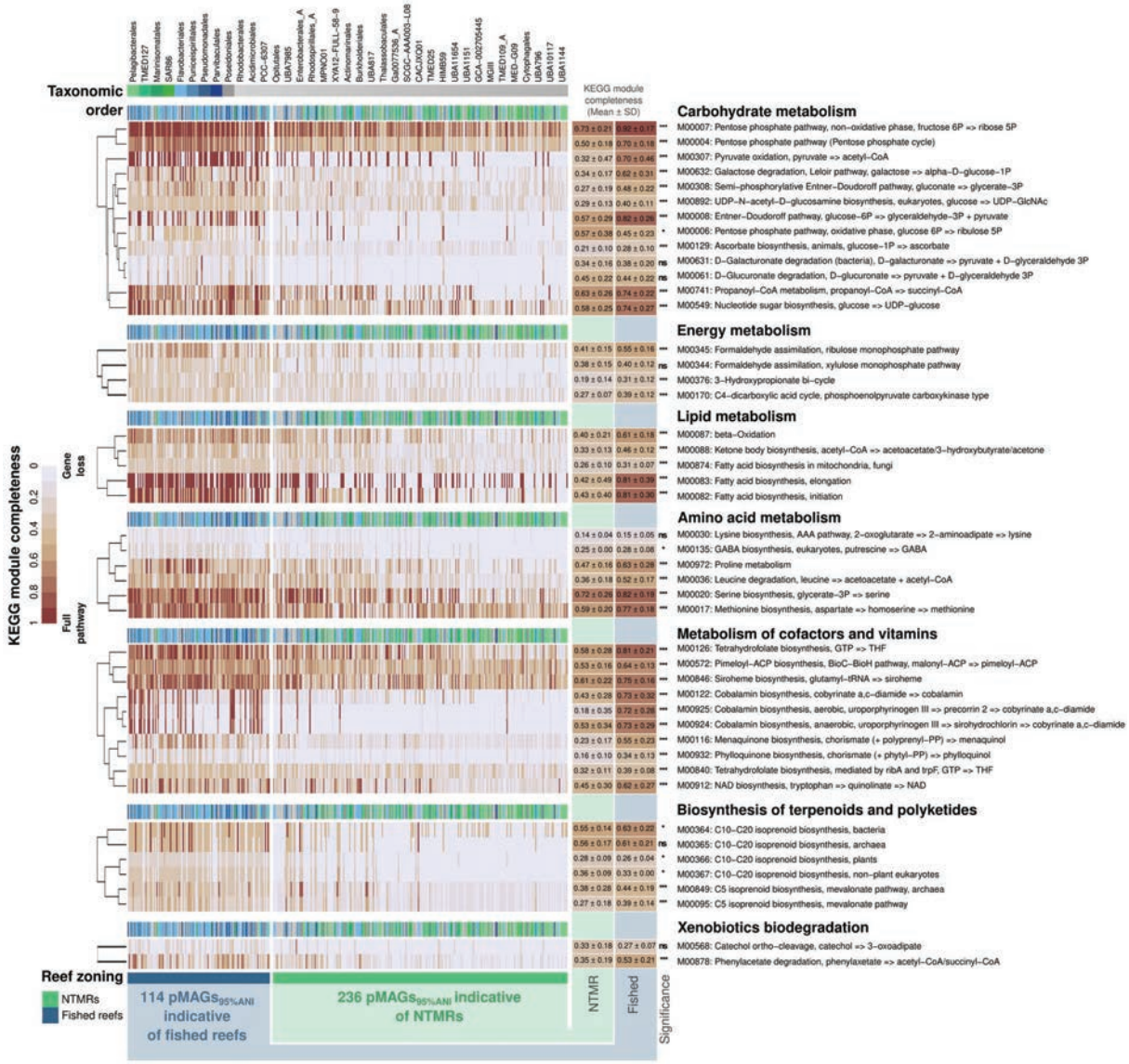

Figure 3.5: Differential metabolic potential of microbial indicator taxa in fished reefs and No-Take Marine Reserves (NTMRs). Heatmaps show the top 45 KEGG modules with the greatest differences in completeness scores between microbial indicator pMAGs from fished reefs and NTMRs. Rows represent KEGG modules (grouped by metabolic category), and columns represent indicator pMAGs clustered by reef zoning status. Module completeness scores (ranging from 0 to 1) are shown by color scale. Columns are annotated by the taxonomic order of each pMAG; only the four most abundant orders in each group are distinctly colored (green for NTMR-enriched, and blue for fished reef-enriched microbes; grey for others). Statistical differences in KEGG module completeness (mean ± SD) between groups were assessed using Wilcoxon rank-sum tests, with adjusted p-values indicated as: *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; "ns" = not significant.

3.4.5 Seawater microbes have a potential to predict continuous environmental variables

Having established seawater microbes as reliable predictors of reef zoning status (**Fig. 3.2**), we next evaluated their potential to predict continuous environmental variables structuring microbial

populations, offering a scalable, biologically integrated approach to track reef conditions beyond categorical (eg. NTMRs vs fished) classifications. Combining random forest models[301] and microbial niche modeling[307], we found that seawater microbes predict several environmental variables with high accuracy ($R^2 > 0.6$; *see methods*) including: seawater temperature, salinity, particulate (>0.7 µm in size) nutrients, and dissolved inorganic phosphorus (**Fig. 3.6a**, green boxplots). In contrast, prediction accuracies were low ($R^2 < 0.3$) to moderate ($0.3 < R^2 < 0.6$) for dissolved nitrogen and silicate, benthic cover variables, fish biomass, and all fish groups (**Fig. 3.6a**, orange and red boxplots) apart from corallivore fish, for which prediction accuracy was high (median $R^2 = 0.72$; **Fig. 3.6a**). This ranked assessment of microbial predictive utility (**Fig. 3.6a**) provides a roadmap for integrating seawater microbes into reef monitoring, with future validation across other reef systems essential to confirm their broader applicability.

Random forest predictions were validated by distinguishing microbial specialists (taxa with narrow environmental tolerance ranges that drove high prediction accuracy) from generalists, which exhibited broader niches and were thus associated with weaker predictive power. The top 50 microbial predictors for temperature (**Fig. 3.6b**, left), which showed a narrow thermal niche range (Q1–Q3: 27.38 ± 2.11℃ to 28.38 ± 1.73℃) with an environmental optimum (Q2) at 27.84 ± 1.88℃ (**Fig. 3.6c**, left), were mostly Flavobacteriales (46%, primarily Flavobacteriaceae and Schleiferiaceae; **Appendix B: Fig. S31**). Flavobacteriaceae have previously been identified as a predictor of seawater temperature in inshore GBR reefs[304], and temperature is well documented to be the main environmental driver structuring reef-associated bacterioplankton assemblages in the GBR[281,348], a pattern that extends to other Pacific reef systems[349] and to open-ocean microbial communities globally[350,351]. Accurate prediction of particulate organic carbon (median $R^2 = 0.74$; **Appendix B: Fig S32a**) and nitrogen (median $R^2 = 0.66$; see **Appendix B: Fig. S32b**) is likely due to the direct role of seawater microbes (picoplankton) in producing particulate organic matter (POM) in the offshore GBR. Seasonal increases in POM during the austral summer (2.3–3.4× higher; **Appendix B: Fig. S18**) are associated with elevated microbial biomass, particularly cyanobacteria that contribute directly to higher POM levels[281,352–354]. This likely explains why summer-enriched microbes were strong predictors of summer-elevated POC (~16% Flavobacteriales, including Arcticimaribacter, MED.G14, Croceivirga, and UBA10364) and particulate nitrogen (~22% Flavobacteriales, ~12% Enterobacterales, and ~6% Pseudomonadales) (**Appendix B: Fig. S32**).

Unlike particulates, most dissolved nutrients and specifically nitrogen (NH_4^+, NO_2^-, NO_3^-, TDN) were poorly predicted by microbial communities (**Fig. 3.6a**). This reflects the nitrogen limitation common on the GBR outer shelf[355], as evidenced by a DIN:DIP ratio of 8.55, which is well below the balanced Redfield ratio of 16:1[331]. Labile DIN is thus rapidly taken up by bacterioplankton, leaving nitrogen concentrations transient and unreliable as indicators of water quality on the offshore GBR reefs[354]. In contrast, phosphorus is rarely limiting and accumulates in winter (**Appendix B: Fig. S18**) due to reduced phytoplankton uptake[281,354]. Seasonal phosphorus dynamics in the GBR may explain the more accurate predictions of PO_4^{3-} ($R^2 = 0.69$; **Figs 3.6a; Appendix B: S33a**) and total dissolved phosphorus (TDP; R^2 =

0.79; **Figs 3.6a; Appendix B: S33b**) both by winter-dominated taxa (predicting high dissolved phosphorus, e.g. *Prochlorococcus*, predominantly observed at the highest measured TDP concentrations of 0.26-0.28 μM; **Appendix B: Fig S33b**) and summer-enriched taxa (e.g. Flavobacteriales, predicting lower PO_4^{3-} and TDP concentrations; **Appendix B: Fig. S33**).

Spatial decoupling and niche differentiation potentially explains why benthic cover variables were difficult to predict from seawater microbial communities. While seawater microbes reflect broader, reef-scale conditions[333], benthic cover (e.g., corals, algae) is often patchily distributed and shaped by microhabitats, contributing to spatial disconnect. Additionally, surface seawater microbial communities occupy a distinct ecological niche from those in the benthic boundary layer, where water chemistry and particle fluxes differ markedly[52,349,356]. As a result, interactions between benthic communities and seawater microbes are frequently mediated by dissolved organic matter or particles, which may dilute direct correlations and obscure mechanistic linkages. This disconnect between pelagic microbial signals and benthic cover highlights the importance of incorporating water chemistry data (**Fig. 3.3**) to contextualise microbial patterns, particularly in relation to nutrient and organic matter fluxes. It also points to future opportunities to improve benthic predictions by integrating spatially resolved ‘omics approaches (such as metatranscriptomics or single-cell sequencing) with benthic-proximal microbial sampling and process-based chemistry flux measurements. Explicitly incorporating hydrodynamic data including current movement, water residence time, and upwelling dynamics, would further resolve how water movement modulates the dispersal of seawater microbes, the flux of benthic-derived nutrients, and the spatial scaling of microbial-benthic coupling[357], ultimately refining our ability to predict reef-scale ecosystem states from microbial signatures. Such an integrative approach could better capture microbial sensitivity to localised benthic inputs (e.g., coral and algal exudates) and ultimately support the development of more context-aware molecular monitoring strategies.

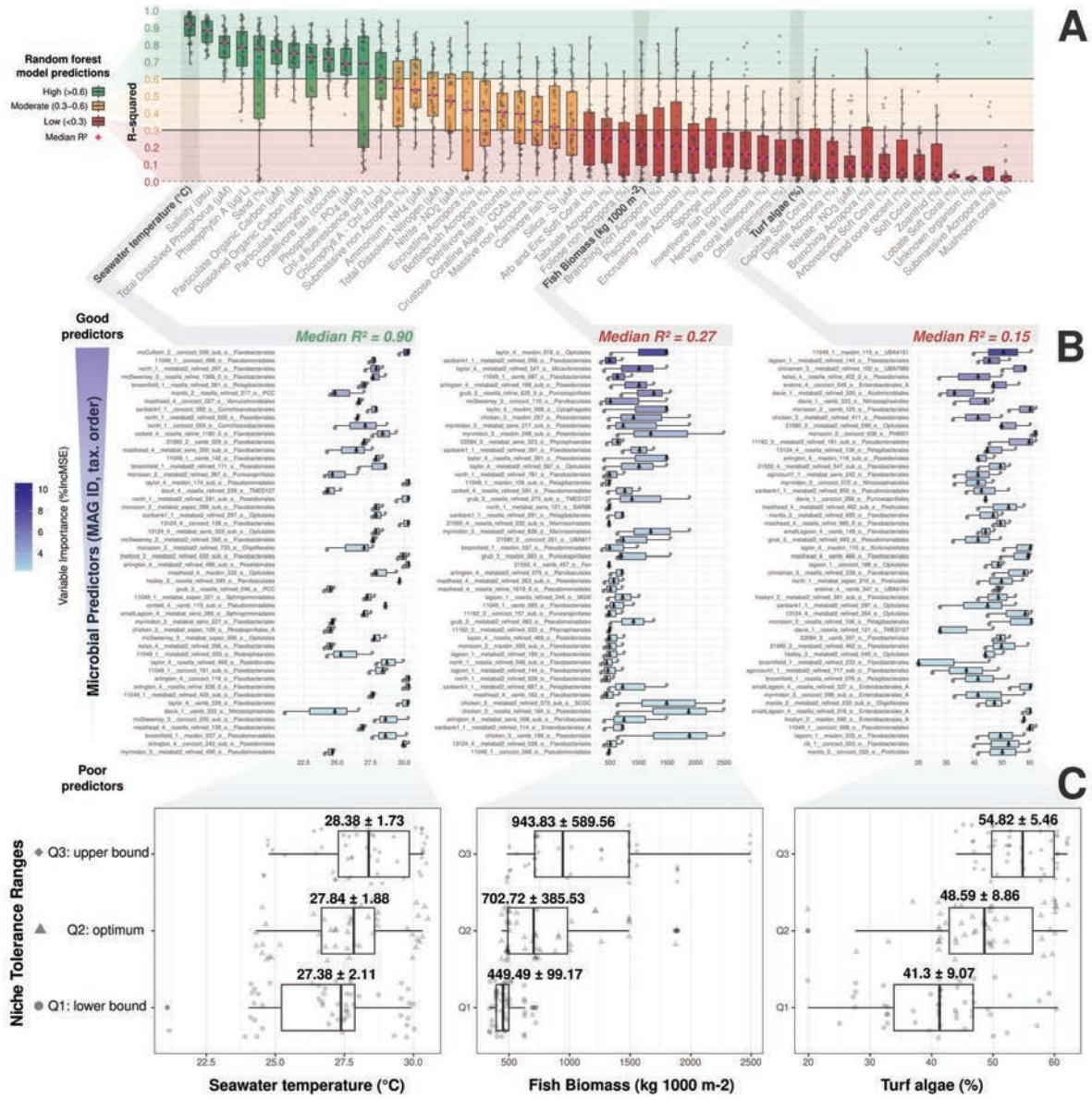

Figure 3.6: Microbial predictors of reef environmental variables, derived from random forest (RF) modeling and microbial niche analysis. (A) Boxplots show random forest model performance across environmental variables, based on R-squared (R^2) values from 50 stratified permutation tests per variable. Cross-validation used an 80/20 train/test data split stratified by the GBR sector. Gray points represent individual permutations, and pink diamonds indicate median R^2. Variables are ordered by decreasing median R^2, with boxplots colored by performance category. The dashed line marks null performance (R^2 = 0). (B-C) Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the top 50 microbial predictors—(B) per pMAG and (C) combined across all 50 pMAGs—for three environmental variables: seawater temperature (left), fish biomass (middle), and turf algae cover (right). Niche bound values are visualised using distinct point shapes. In (B), microbial predictors are additionally colored by random forest importance (%IncMSE).

3.5 Conclusions

We put forward a mechanistic explanation for the effects of reef zoning on seawater microbial communities, proposing that zoning in the offshore GBR may shape seawater microbial communities through ecological feedbacks tied to nutrient dynamics (summarised in **Fig. 3.7**). In a subset of NTMR reefs, lower nutrient concentrations may be indicative of efficient nutrient cycling cumulatively driven by high fish biomass, increased algae removal due to herbivory, and enhanced coral cover, which selects for oligotrophic (microbial streamlining theory) and cooperative (Black Queen Hypothesis) microbes (**Fig. 3.7**, NTMRs). In contrast, fished reefs with higher turf-to-coral ratios tend to exhibit more nutrient-rich conditions, favoring metabolically independent and competitive microbes with larger genomes (**Fig. 3.7**, fished reefs). These changes in water chemistry and benthic cover, and their cumulative flow-on effects on microbial community composition, enabled reef zoning to be predicted from seawater microbial data alone with 71% accuracy in the offshore GBR. This high classification accuracy highlights the value of seawater microbial markers as complementary indicators of zoning effectiveness, which is relevant because zoning non-compliance is an ongoing challenge, with 500–600 annual zoning offenses (e.g., illegal fishing) potentially undermining NTMR benefits like fish spillover and biodiversity protection[271,358,359]. Considering the global success of NTMRs, future research should explore whether similar microbial shifts occur in marine reserves of other reef systems beyond the GBR. Microbial monitoring could complement existing in-water assessments by expanding spatial coverage across a broader range of reef locations and environmental conditions, offering a scalable and sensitive addition to current monitoring frameworks, ultimately informing more holistic conservation and management strategies.

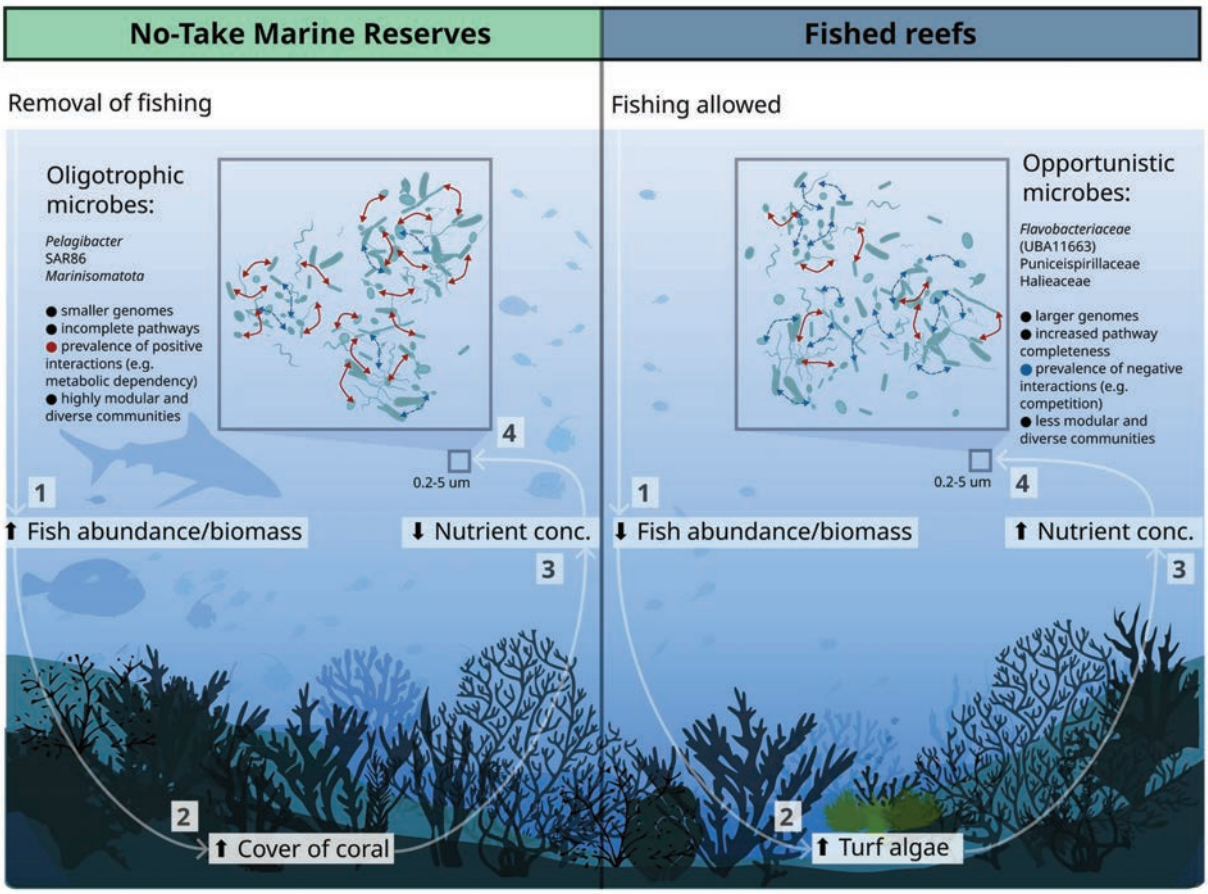

Figure 3.7: Conceptual model of the seawater microbial dynamics within No-Take Marine Reserves (NTMRs) and fished reefs in the Great Barrier Reef, in the context of physico-chemical, fish abundance, and benthic cover variables. (**A**) In some NTMR reefs, the removal of fishing pressure results in higher fish biomass, including herbivorous fish (**1**), which enhances grazing pressure on algae and promotes increased hard coral cover (e.g., *Acropora* spp., *Millepora*) and crustose coralline algae (CCA) abundance (**2**). Efficient nutrient cycling in such healthier reefs with elevated fish biomass and coral cover leads to lower nutrient availability (**3**). This reduction in nutrients imposes strong selective pressures on seawater microbial communities (**4**), enriching NTMRs with oligotrophic taxa (e.g., *Pelagibacterales*, SAR86, *Marinisomatota*) that exhibit streamlined genomes and incomplete biochemical pathways, as predicted by the microbial streamlining theory. These microbes form positively connected communities, likely reflecting metabolically co-dependent relationships where streamlined taxa rely on other bacterioplankton members to exchange missing metabolites, consistent with the Black Queen Hypothesis. (**B**) In fished reefs, lower fish biomass and reduced grazing pressure (**1**) result in higher turf-to-coral ratios (**2**) and higher nutrient concentrations (**3**), likely due to inefficient nutrient cycling and DOM release from turf algae, as predicted by the DDAM model. These conditions favor opportunistic microbial taxa (**4**), ultimately leading to distinct microbial community dynamics compared to NTMRs.

3.6 Declarations

3.6.1 Ethics approval and consent to participate

Samples were collected under the permit G12/35236-1 issued by the Great Barrier Reef Marine Park Authority.

3.6.2 Consent for publication

Not applicable.

3.6.3 Availability of data and material

Sequencing data and primary metagenomic assemblies have been uploaded to the European Nucleotide Archive (ENA) under project accession PRJEB82623 under the project name "Great Barrier Reef seawater microbiomes genome database". The 5,283 prokaryotic metagenome-assembled genomes (pMAGs) are accessible from Zenodo (DOI: 10.5281/zenodo.17109887).

The associated physico-chemical variables have been uploaded to the IMOS-AODN repository and are available from the Australian Institute of Marine Science (AIMS) (2022): Great Barrier Reef Microbial Genomics Database: Seawater Illumina Reads. https://doi.org/10.25845/Q4XH-YN10.

The benthic cover, fish abundance, and biomass data are managed by the AIMS Long-Term Monitoring Program (LTMP) and can be accessed via the AIMS data portal (https://apps.aims.gov.au/metadata/view/a17249ab-5316-4396-bb27-29f2d568f727).

Metagenomic analysis including metagenome hybrid assembly, binning, taxonomic annotation, and abundance estimation of pMAGs is described in Robbins et al. (2025).

All additional code including indicator analysis, microbial networks, integration of microbial and environmental data, and random forest machine learning is available at: https://github.com/mterzin/fishy_microbes (Terzin, 2025). The microbial niche analysis was performed following the protocol from Chaffron et al. (2021), and the specific code for this analysis is available from the authors of that study upon request.

3.6.4 Competing interests

The authors declare no competing interests.

3.6.5 Funding

This study forms part of the Australia's Integrated Marine Observing System (IMOS) Great Barrier Reef Microbial Genomic Database sub-facility (GBR-MGD), funded by the Queensland Research Infrastructure Co-investment Fund (RICF) by the Department of Environment and Science, Queensland. IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent. This study was also funded by an AIMS@JCU PhD Scholarship to MT. Additionally, MT acknowledges the EMBL Australia Short-Term Travel Grant, which facilitated collaborative research on microbial interaction network analysis with Dr. Samuel Chaffron (Nantes Université) and Dr. Flora Vincent (EMBL Heidelberg). The funders had no role in sampling design, data collection, processing and interpretation, preparation of the manuscript, or decision to publish.

3.6.6 *Authors' contributions*

NSW obtained funding for the project. NSW, DGB, PWL, RKG, and SJR conceived the sampling design. SCB collected seawater in the field (for metagenomics and physicochemical data) and processed all samples in the laboratory for metagenomic sequencing. LTMP data (benthic cover and fish abundance/biomass) were processed by MJE and DMC as part of the LTMP, with GLMMs run by MJE. Metagenomics analyses including hybrid assembly, binning, taxonomic annotation, and abundance estimation were performed by SJR, YKY, KED, and JZ, with the guidance of PH, PWL, and DGB. MT performed the functional annotation of metagenomes and carried out all subsequent analyses, including dataset integration and data visualisation, with assistance from PWL, KALC, YKY, DGB, SJR, SC, RKG, and MJE. MT wrote the original draft of the manuscript, and all authors made substantial contributions to its form. All authors critically reviewed the manuscript before submission.

3.6.7 Acknowledgements

The seawater samples analysed in this study for metagenomics and physico-chemical variables were collected across 48 reefs, from the sea country of various Indigenous groups who are Traditional Owners (TOs) of that land. We acknowledge the Gudang Yadhaigana TOs, custodians of the McSweeney, Monsoon, 11-049, and 11-162 reefs, which lie within their sea country estate. We pay our respects to the Kuuku Ya'u TOs of the Mantis and Lagoon reefs, and the Lama Lama TOs of the 13-124, Davie, and Corbett reefs in the western half of their territory. We also recognise the Cape Melville, Howicks, and Flinders Island TOs of the eastern half of Corbett and Sand bank #1 reefs, as well as the Eastern Kuku Yalanji TOs of St Crispin and Agincourt #1. We extend our respect to the Yirrgandji TOs of Hastings reef and the Gunggandji as TOs of Arlington, Thetford, and Moore reefs. We acknowledge the Gunggandji-Mandingalbay Yidinji TOs of McCulloch, Hedley, Peart, and Feather reefs, and the Mandubarra TOs of Farquaharson Reef. We honour the Girringun Aboriginal Corporation TUMRA for their connection to

Taylor Reef, and the Manbarra TOs of Rib, Kelso, Little Kelso, and John Brewer reefs. We also recognise the Wulgurukaba TOs of Myrmidon, Grub, and Helix reefs, the Bindal Traditional Owners of Knife, Fork, Centipede, Chicken, and Lynchs reefs, and the continuing connection of the Manbarra TOs to Roxburgh, and Fore and Aft reefs. Lastly, we acknowledge the PCCC TUMRA for their stewardship of North, Bloomfield, Eskine, Mast Head, Hoskyn, Fairfax, and Boult reefs. We pay our respects to their Elders, past, present, and emerging, and acknowledge their enduring connection to land and sea. Further, our desktop / lab research took place at the Australian Institute of Marine Science (AIMS) headquarters at Cape Ferguson, and we wish to acknowledge the Wulgurukaba and Bindal peoples as the Traditional Owners of that land. This research was also undertaken at the JCU Townsville Bebegu Yumba campus, and the authors acknowledge that the Australian Aboriginal and Torres Strait Islander peoples are the original inhabitants and traditional custodians of this continent and have unique cultural and spiritual relationships to the land and waters. We acknowledge the AIMS Water Quality team, especially Ulysse Bove, Keeley Glasson, and Daniel Moran for logistics, training, and processing of water chemistry samples. We acknowledge the AIMS-LTMP team and others involved in field collection and preparation of samples including Emmanuelle Botté, Johnston Davidson, Veronique Mocellin, and Josephine Nielsen. We thank the crew of the RV Solander and RV Cape Ferguson for their excellent logistical support in the field. We also acknowledge Gene Tyson for his support in facilitating the use of the NovaSeq at Microba Life Sciences Ltd. (Brisbane, QLD, Australia). We extend our gratitude to Murray Logan for his insightful discussions on the appropriate statistical handling of the data. AI tools (ChatGPT and DeepSeek) were used exclusively for proofreading (grammar, syntax, and clarity checks) and code assistance (debugging and statistical script optimization). No AI tools were used to generate original content, interpret data, or formulate conclusions. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2025648). MT extends his gratitude to the members of the Lê Cao Lab at Melbourne Integrative Genomics (MIG) for the supportive environment and valuable scientific discussions, particularly Vinicius Salazar, Saritha Kodikara, and Jiadong Mao.

Data Analysis | Integrating microbial and viral responses to reef zoning measures in the Great Barrier Reef

This Chapter is under preparation for submission.

4.1 Abstract:

Seawater viruses are increasingly recognised as key players in coral reef health through viral lysis and control of microbial cell abundances. Despite this recognition, seawater viromes are largely understudied in the Great Barrier Reef (GBR) marine park. As a part of the Great Barrier Reef Microbial Genomics Database infrastructure, we integrated viral and microbial (pMAG) community data to identify multi-omics temporal and geographic biomarker signatures of No-Take Marine Reserve (NTMR) versus fished reef zones. We successfully identified viral biomarkers distinguishing reef zoning status, particularly the Caudoviricetes that correlated with NTMRs or fished reefs, predicting zoning status with an average 72% accuracy (same as pMAGs). Integrating viral and microbial data however lowered zoning predictions (59% classification accuracy), likely due to the same virus associating with different microbial hosts. These results highlight viruses as promising (but complex) indicators of reef management impacts, and future studies should prioritise host-virus interaction mapping and time-series sampling to clarify viral roles in reef health across environmental gradients of the GBR.

4.2 Introduction

Seawater viruses are the most abundant biological entities in the ocean, lysing ~20% of microbial biomass daily[360]. On coral reefs, they regulate microbial community structure, shape host population dynamics, and drive nutrient cycling, thereby underpinning reef health and resilience[361]. While early conceptual models predicted that degraded, algae-dominated reefs would support higher viral loads due to increased microbial densities and intensified lytic infections, recent evidence points to a "more microbes, fewer viruses" pattern in degraded reefs[362–364]. Viral metagenomics studies revealed that the relative frequency of genes (but not virions) from temperate viruses (i.e. exhibiting both lytic and lysogenic infection strategies) increased with microbial density[362], suggesting a shift towards lysogeny when host abundance is high[364]. This aligns with the "Piggyback-the-Winner" hypothesis, in which viruses integrate into host genomes and replicate vertically[365], reducing lysis and ensuring viral persistence (i.e., by preventing superinfection of the same bacterial host cell by a closely-related phage).

Current consensus holds that this lysogenic phage integration into the host genome is facilitated by distinct molecular mechanisms operating both at low and high ends of the prokaryote cell density spectrum. Under low-density conditions (>10^4 cells mL^{-1}, e.g., in the deep ocean), nutrient limitation represses lytic genes and extends the lysogenic phase, consistent with the refugium hypothesis[366]. At high densities (>10^6 cells mL^{-1}, e.g., microbialized reefs), co-infections and altered host physiology favor integration into host genomes[364,366]. In contrast, intermediate microbial densities (10^5–10^6 cells mL^{-1}, e.g., healthy coral-dominated reefs) are thought to favour lytic infections due to higher viral-bacterial encounter rates and high intracellular ATP concentrations[367]. This "kill-the-winner" dynamic, analogous to the traditional Lotka–Volterra predator–prey model[368,369], reduces dominance of abundant microbes,

maintaining microbial biodiversity in both pelagic oceans[370] and coral reefs[367]. These dynamic infection strategies have profound implications for microbial community structure and reef biogeochemistry, with viral lysis removing up to half of bacterial standing stock each day including in healthy reef systems[371,372], where viruses act as a primary top-down control on microbialization and promote microbial diversity[367].

Understanding microbial interactions is also critical when interpreting microbial-based reef monitoring data. For example, a gammaproteobacterium *Vibrio cholerae* increased in abundance in phytoplankton bloom waters (suggesting *V. cholerae* could be a good indicator of bloom events), but was grazed by protozoan predators which kept *V. cholerae* populations under control, obscuring its value as an indicator[373]. Similarly for viruses, viral lysis may remove key microbial indicator taxa before sampling, underestimating their ecological role or abundance. Cross-domain analyses have shown that viruses can even outperform microbes and eukaryotes as predictors of environmental processes like carbon fluxes and export, and that 'classic' correlation-based approaches (e.g. regression) may miss key signals when species interactions are mutually exclusive (e.g. pathogenic phage-host dynamics), which can cancel out in bulk community datasets[374]. Given the potential for viruses to 'erase' microbial indicator signals in scenarios of viral lysis prior to field sampling, combining viral and microbial abundance data may offer a more accurate and holistic assessment of reef condition.

Here, we addressed this gap through two objectives: (1) identifying viral markers of reef zoning status (No-Take Marine Reserves vs. fished reefs), and (2) evaluating whether integrating viral and microbial (pMAG) abundances improves predictions compared to single-omics approaches. Despite their importance, few studies have characterised seawater viral communities in the Great Barrier Reef (GBR). Water-column viral assemblages from the GBR represented the most distinct communities identified in the Pacific Ocean Viromes dataset, highlighting the potential uniqueness of coral reef seawater viruses[375]. On the southern GBR shelf, virioplankton abundances correlated with bacterioplankton productivity and particulate nutrients, particularly in reef channels and inshore zones where vertical mixing enhances benthic-pelagic coupling[376]. By integrating microbial (pMAG) and viral omics layers generated within the Great Barrier Reef Microbial Genomics Database (GBR-MGD) infrastructure, the most comprehensive resource on GBR reef-associated seawater microbiomes to date[270], we provide the first multi-omic assessment of virus–microbe interactions across GBR management zones.

4.3 Materials and Methods

4.3.1 Description of the GBR-MGD dataset and selection of prokaryotic and viral omics layers

The Great Barrier Reef Microbial Genomics Database (GBR-MGD) generated by Australia's Integrated Marine Observing System (IMOS) represents a comprehensive resource containing: (1) 5,283 seawater prokaryotic (bacterial and archaeal) metagenome-assembled genomes (pMAGs); (2) 20

chromosome-level picoeukaryote genomes; (3) 808,585 viral genomes; and (4) 24,769 mobile elements (putative plasmids), collected alongside a range of contextual environmental parameters[270]. As detailed in Robbins et al. (2025)[270], a variety of tools were used to obtain each planktonic fraction. pMAGs were generated from hybrid (Illumina-Nanopore) and short-read-only assemblies, binned using the Aviary (https://github.com/rhysnewell/aviary; v0.3.3) pipeline, and dereplicated at 95% ANI in CoverM[285] (v0.6) to yield a final set of 876 high-quality, species-resolved genomes for analysis. For the picoeukaryote fraction, Whokaryote[377] (v1.1.2) and EukRep[378] (v0.6.6) were employed to identify putative eukaryotic contigs based on genomic characteristics (e.g. gene length, density, intergenic distance) and kmer frequencies. Viral genome identification relied on a suite of tools: GeNomad[379] (v1.5.018) with database v1.2, PPR-Meta[380] (v1.134) using a threshold of 0.9, VIBRANT[381] (v1.2.1), ViralVerify (https://github.com/ablab/viralVerify; v1.133), and VirSorter2[382] (v2.2.4). Finally, plasmids and other mobile elements were characterised using GeNomad, PPR-Meta, PlasForest[383] (v1.2), PlasX[384] (unversioned; applying a minimum score threshold of ≥ 0.9), and ViralVerify.

A standardised labelling scheme with the structure of IMOS__<TaxaGroup>__SAMPLENAME__CONTIG:<TaxaComponent> was applied to all sequences, where TaxaGroup indicates all taxa linked to a contig (E=eukaryote, M=plasmid, P=prokaryote, V=virus). Multiple letters indicate contigs linked to more than one group (e.g., a PV contig was found in a pMAG and simultaneously predicted to be a virus). This study utilised the dereplicated set of 876 pMAGs (IMOS_P) and viral contigs exclusively identified as viruses (IMOS_V).

4.3.2 Filtering of viral features

A total of 362,802 initial viral contigs (i.e., group "IMOS_V" within the GBR-MGD resource) were progressively filtered by (1) length (>3 kb); (2) prevalence (>75% of samples per reef zone); (3) relative abundance across all samples (>0.001%); and (4) taxonomic annotation (GENOMAD), yielding 6,024 high-confidence viral contigs for downstream indicator analysis. Specifically, we retained viral contigs >3 kb in length (289,137 contigs; 73,665 removed) as this aligns with IMG/VR and GOV2.0 database standards, with shorter fragments (< 3 kb) more likely to represent degraded DNA or assembly artifacts and often lack reliable viral signatures[385]. Subsequent prevalence filtering excluded contigs detected in fewer than 75% of either NTMR or fished reef samples (73,412 contigs remaining), in addition to rare and spurious viral contigs (with average relative abundance below 0.001% across samples) which were removed as they could represent artifacts of sequencing and assembly errors, yielding 13,603 virus contigs (including both annotated and unannotated viruses). The last filtering step involved removing the non-annotated viruses (i.e. missing GENOMAD annotations) to only focus on the annotated viral taxa as informative indicators, resulting in a final dataset of 6,024 viral contigs (>3kb; prevalence of >75% of samples within NTMRs and fished reefs; average rel. abund > 0.001%; with GENOMAD taxonomic annotations).

4.3.3 Identifying virus-microbes multi-omics signatures or reef zoning measures

Following centered log-ratio (CLR) transformation using the microbiome (version 1.24.0) R package (Lahti and Shetty 2017) for both microbiome (876 $pMAGs_{95\%ANI}$) and virome (6,024 contigs) datasets, we initially employed DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents)[386] in mixOmics[280] (v6.26.0) (R version 4.3.2; R Core Team, 2023) to correlate pMAGs and virus signatures discriminating No-Take Marine Reserves (NTMRs) and fished reefs (**Appendix C: Fig. S5-S11; Tables S2-S3**). Even though preliminary PLS regression and DIABLO diagnostic plot interpretation revealed strong cross-correlation (r~0.9) between microbial and viral components (**Appendix C: Fig. S6**), a design matrix of 0.1 was selected (i.e. prioritizing zoning discrimination over pMAGs-virus correlations) and applied using a 50 × fourfold cross-validation (Mahalanobis distance) to identify the optimal number of DIABLO components and features (per component). While DIABLO successfully identified microbial-viral multi-omics signatures that discriminate reef zoning, the model also captured significant batch effects originating from the time of sampling and geographic site proximity (**Appendix C: Fig. S5-S11**). This was inferred from the model's failure to stabilise, as DIABLO kept adding components (**Appendix C: Fig. S5; Table S2**) and incorporating more features (**Table S3**) without a clear performance plateau. The persistent grouping of samples by trip (i.e., sampling time and location) in the component score plots (**Figs. S7-S8**) further confirms these batch effects.

To address the spatiotemporal confounding factors and identify robust microbial and viral zoning indicators across the GBR, we first applied Multivariate INTegration Sparse Partial Least Squares Discriminant Analysis - MINT sPLS-DA[291,292,294] individually to the prokaryotic and viral datasets, as a P-integration approach[2]. This yielded a prediction accuracy of 71% for the prokaryotic model (retaining 350 and 180 pMAGs per component; **Appendix C: Figs S12-S15; Tables S3-S5**) and 72% for the viral model (retaining 240 and 110 viral contigs per component; **Appendix C: Figs. S16-S19; Table S6**). Aiming to extract universal multi-omics signatures (i.e., correlated pMAGs and viruses) of reef zoning that remain stable despite spatiotemporal sampling differences, we then implemented MINT BLOCK sPLS-DA as a hybrid method combining N-integration (integrative analysis of microbes and viruses from the same GBR-MGD samples) with P-integration (accounting for variability across distinct GBR sectors). Since current implementation lacks parameter tuning for MINT BLOCK sPLS-DA, the number of components and feature selection parameters (keepX) were derived from optimised MINT sPLS-DA runs on single-omics: pMAGs = [350, 180] and viruses = [240, 110] per component, and a 0.1 weighted design was again chosen to prioritise discrimination over pMAG-virus correlations[386]. MINT BLOCK sPLS-DA results were visualised as a consensus sample plot and a heatmap using mixOmics[280] and ggplot2[308]. Final composite plots were made in Inkscape 0.92.5.

4.3.4 Validating virus-host interactions

To provide ecological context and validate the biological relevance of the virus-microbe correlations identified in our multi-omics signatures, putative hosts for the viral contigs were predicted using the iPHoP[373] (Integrated Phage-Host Prediction) platform (v1.3.3). The query set focused exclusively on the high-confidence Great Barrier Reef Microbial Genomics Database (GBR-MGD[270]) viral contigs belonging to the IMOS_V group (i.e., consensus annotation was 'virus only'). The collection of 5,283 pMAGs from the GBR-MGD were incorporated into iPHoP's internal host database (version Aug_2023_pub_rw) for host prediction, and a combination of marker-based and sequence similarity approaches was used to assign taxonomic predictions and confidence scores (≥90% as cutoff). By incorporating pMAGs from the GBR-MGD into iPHoP, this analysis was specifically optimised to identify virus-host interactions within the GBR ecosystem, providing a robust validation of the correlations observed in our MINT BLOCK sPLS-DA model. iPHoP-predicted virus-host interactions were visualised in ggplot2[308].

4.3.5 Comparative performance of single- and multi-omics models

A Leave-One-Group-Out Cross-Validation[282] (LOGOCV) scheme was used to assess the predictive performance of the MINT BLOCK sPLS DA model. This process involved iteratively holding out all samples from a single GBR sector as an independent validation set and training the model on the six sectors, which was repeated until each of the seven sectors (CA, CB, CG, IN, PC, SW, TO) had served as the validation set once. For each iteration, the MINT BLOCK sPLS-DA model was trained on six sectors using the pre-tuned parameters: keepX = list(pMAGs = c(350, 180), Virus = c(240, 110)) for two components, and the previously defined small-weighted (0.1) design matrix to prioritise zoning discrimination. The trained model was then used to predict the reef zoning status (NTMRs or fished reefs) for each sample in the left-out seventh sector using the centroids distance metric. Prediction accuracy (1 – classification error) was calculated for each sector and for each omics block individually (pMAGs and viruses) per component, and for consensus predictions integrating both blocks (AveragedPredict).

The prediction accuracy scores from the MINT sPLS-DA (for pMAGs and viruses separately) and the MINT BLOCK sPLS-DA models (using AveragedPredict) were visualised and compared using boxplots in ggplot2[296] to evaluate if multi-omics integration improved model performance. Venn diagrams were used to visualise the overlap between the pMAGs and viruses selected on the first component of the MINT sPLS-DA models and those selected by the MINT BLOCK sPLS-DA model to assess the consistency of feature selection between the single and multi-omics approaches. This analysis revealed the proportion of discriminatory features that were uniquely identified by each method versus those that were consistently selected by both single- and multi-omics.

4.4 Results and Discussion

4.4.1 Seawater viruses accurately classify reef zoning status by tracking the dynamics of their prokaryotic hosts

Viral community composition strongly mirrored prokaryotic trends with both clustering primarily by season rather than reef zoning status (**Fig. S1**). For reef bacterioplankton, this trend was driven by seasonally abundant microbes such as *Synechococcus*, *Sphingomonas*, and *Luminiphilus* in summer and *Prochlorococcus*, SAR86, and *Pelagibacter* in winter (**Appendix C: Fig. S2**; above). Viral communities consisted primarily of tailed bacteriophages (class Caudoviricetes; **Appendix C: Fig. S2**, below), with seasonally abundant viral contigs (**Appendix C: Fig. S3**) associating with the dominant seasonal microbial hosts (**Appendix C: Fig. S4**), highlighting a tight ecological linkage between host and viral dynamics. This is expected as viruses are obligate intracellular parasites and thus exist in a dynamic state, being environmentally inert in their extracellular form (as virions) yet wholly dependent on host activity for replication[387]. Hence, viral abundances more strongly reflect host activity than direct environmental responses. Due to this tight virus-host coupling, both reef bacterioplankton (**Appendix C: Figs S12-S15; Tables S3-S5**) and viral communities (**Appendix C: Figs. S16-S19; Table S6**) achieved nearly identical classification accuracies of 71% and 72% (respectively) in predicting reef zoning status across 48 offshore reefs. Caudoviricetes (tailed bacteriophages) were the primary discriminators of reef zoning status consisting of 237 (98.75%) of the 240 viral biomarkers (**Appendix C: Fig. S17**), while 3 others (1.25%) belonged to giant viruses (Megaviricetes, Phycodnaviridae).

Integrating the viral and pMAG datasets (to identify the putative prokaryotic hosts of these zoning-specific viral biomarkers) showed that samples clustered based on reef zoning (NTMRs vs. fished reefs), with generally short pMAGs-Virus vectors in the MINT BLOCK sPLS-DA sample plot indicating strong agreement between the prokaryotic and viral omics blocks for most samples, suggesting a tightly coupled multi-omics signal (**Fig. 1a**). Selected features further revealed distinct abundance patterns of discriminant (and correlated) pMAGs and viruses across samples (**Fig. 1b**), and interestingly, many viruses indicative of a specific zoning status (e.g., fished reefs) were associated with prokaryotic hosts that were also discriminant for that same zoning status, as validated by iPHoP virus-host predictions (**Fig. 1c**). Specifically, viral indicators of NTMRs were clearly associated with NTMR-enriched pMAGs, including 86 Pelagibacteraceae pMAGs enriched in NTMRs (in addition to 169 Pelagibacteraceae non-discriminative of reef zoning), 87 members of the family GCA–002718135 (Alphaproteobacteria), and 50 NTMR-enriched pMAGs within the family TMED112 (genus SAR86) (**Fig. 1c**; viral indicators of NTMR reefs). In addition, viral indicators of fished reefs were associated with pMAGs enriched in fished reefs, predominantly classified within families Rhodobacteraceae (173 indicators), Flavobacteriaceae (71 indicators), and UA16 (2 indicators), with a single exception of one NTMR-enriched pMAG within Thalassobaculaceae associated with a single viral contig (IMOS__V__arlington_4__contig_13151_pilon__v) enriched in fished

reefs (**Fig. 1c**; viral indicators of fished reefs). This precise alignment confirms that the discriminative power of the viral community is executed through specific interactions with zoning-responsive prokaryotes as their putative hosts, whose abundance patterns they track. For example, some viral indicator signals may originate from viral genomic elements integrated within their host's DNA (e.g. prophages, endogenous viral elements, or auxiliary metabolic genes) that thereby mirror microbial abundances. Future research is needed however to better understand the mechanisms behind the observed pMAG-virus correlations.

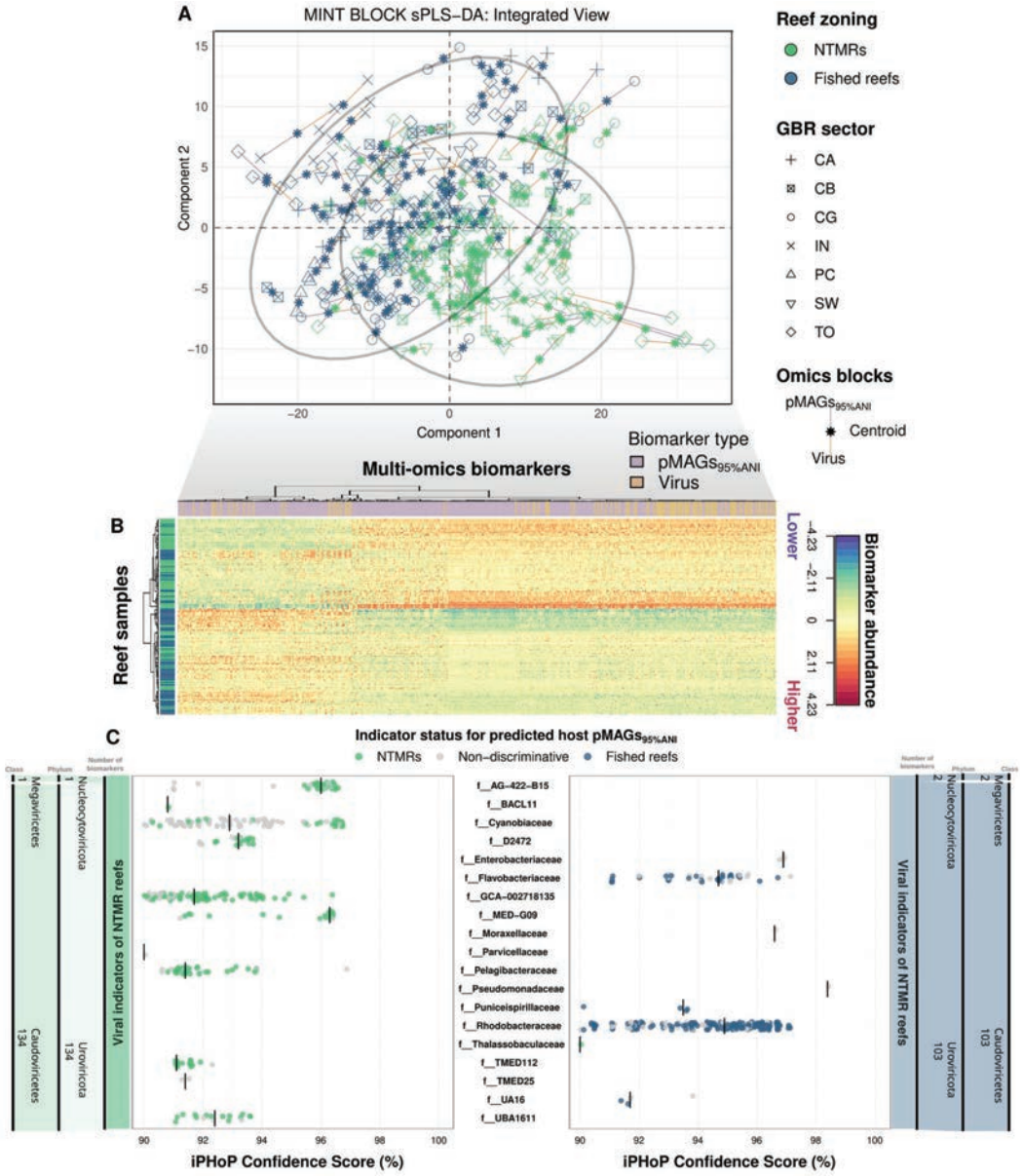

Figure 4.1: Multi-omics signatures (seawater microbes and viruses) that distinguish No-Take Marine Reserves (NTMRs) and fished reefs across Great Barrier Reef (GBR) sectors. Consensus sample plot from MINT BLOCK sPLS-DA on seawater pMAGs and virus data, used to discriminate reef zoning status (fished reefs in blue; NTMRs in green) across GBR sectors (shapes), and with ellipses denoting 95% confidence intervals. The asterisk for each sample represents its consensus centroid between both omics data blocks, with vectors from this centroid pointing to the sample's projection from each individual block (pMAGs in purple; Virus in orange). The length of a vector indicates the degree of disagreement between the blocks for that sample, with longer arrows representing a stronger block-specific signal. (B) Heatmap of variables selected by MINT BLOCK sPLS-DA (component 1), showing the differential abundance of discriminant seawater pMAGs (purple) and viruses (orange) across reef samples colored by reef zoning status. (C) Distribution of high-confidence iPHoP prediction scores (90-100%, x-axis) for viral contigs identified as indicators of NTMRs (left) or fished reefs (right), showing their predicted prokaryotic host pMAG95%ANI collapsed at family level (y-axis). Points represent individual virus-host predictions, colored by the indicator zoning status (NTMR, fished, non-discriminative) of the host pMAG95%ANI. Alluvial diagrams were used to show the taxonomy of indicator viral contigs.

4.4.2 Multi-omics integration trades prediction accuracy for ecological insight into reef viruses

While the integration of viral and microbial abundance data facilitated deeper ecological insights into virus-host dynamics in the context of reef zoning (**Fig. 1**), the integrated model counterintuitively reduced predictive performance to ~59% classification accuracy (**Fig. 2a**) compared to single-omics accuracies of ~71% (pMAGs; **Fig. 2b**) and ~72% (Virus; **Fig. 2c**). Further, while ~85% of microbial biomarkers remained consistent across single-omics and multi-omics analytical approaches (**Fig. 2d**), only ~55% of viral markers showed similar stability after integration (**Fig. 2e**). This apparent paradox likely stems from fundamental biological differences in how viruses and microbes respond to their environment, with lower stability of viral biomarkers reinforcing their indirect, host-mediated environmental responses. Temporal decoupling may also explain the observed accuracy drop of the multi-omics model, as viral lysis events that precede (and affect) microbial abundance changes can create mismatches in host-virus co-occurrence patterns at the time of sampling. While the generally short vectors between the consensus centroid and each block's projection indicate strong agreement between prokaryotic and viral signals for most sites (**Fig. 1a**), samples from Myrmidon, Farquaharson, and Chicken reefs show notable exceptions with longer vectors (**Appendix C: Fig. S20**) indicating disagreement between the microbial and viral datasets that suggests viral abundance may be temporally or contextually decoupled from its host at these locations (thus causing the observed accuracy drop of the multi-omics model). Further, our analysis revealed substantial variation in viral host range, with indicator viral contigs associating on average with 20 ± 32 distinct microbial hosts (range: 1-170 hosts per viral contig; see **Appendix C: Fig. S21**). The broad host range shows that some seawater viruses in the GBR are generalists (correlated with multiple microbial taxa) while others exhibit more specialised host associations. This extensive variation in host range, coupled with the likelihood that the same virus employs different infection strategies depending on local environmental conditions, seasonal dynamics, or unmeasured ecological factors, all contribute to the complex patterns observed in the MINT BLOCK sPLS-DA projections. The broad host ranges predicted by iPHoP may reflect both true ecological generalism among some viral lineages and the methodological detection of incidental sequence similarities or prophage-like integrations across distantly related hosts; future experimental validation (e.g., host-range assays, Hi-C sequencing, or viral tagging) is needed to distinguish between these possibilities and clarify the functional relevance of predicted virus-host interactions. Together, these factors explain why a smaller number of viral biomarkers are consistently selected between single- and multi-omics models (**Fig. 2**) and offer another plausible explanation for the drop in predictive accuracy of the multi-omics model (**Fig. 2a**) compared to single-omics models (**Fig. 2b-c**).

Despite lower classification accuracy, multi-omics integration should not be seen as a failure but as evidence that the model captures more complex biological interactions. Whereas single-omics models highlight discriminant features, the integrated model exposes the potential structure of virus–host

interaction networks, offering broader insight into how reef zoning shapes various components of seawater microbial communities. In particular, the identified "viral fingerprint" may act not only as a reflection of prokaryotic composition but also as a driver of bacterioplankton function under reef zoning (and potentially other human pressures), underscoring the potential of viruses as indicators of coral reef health.

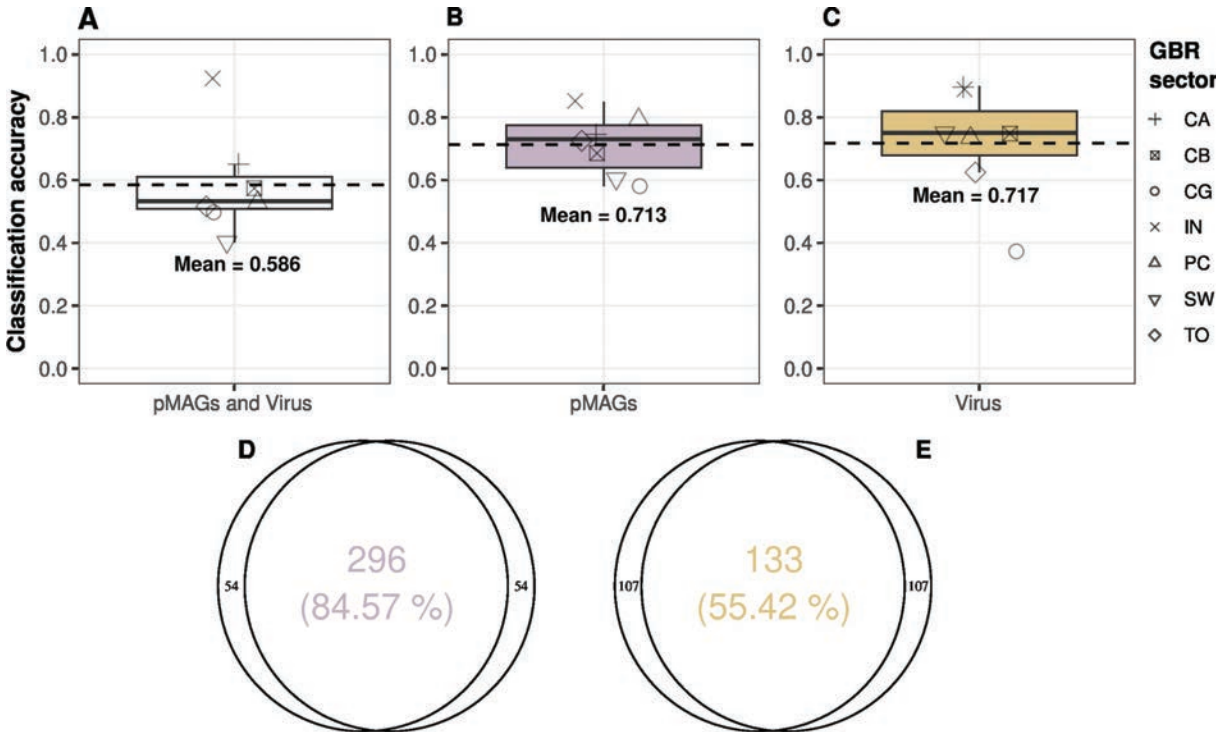

Figure 4.2: Comparative performance of single- and multi-omics models. Boxplots show the difference in model accuracy scores in discriminating between NTMRs and fished reefs, when integrating pMAGs and viral data to identify multi-omics signatures of NTMRs and fished reefs (MINT BLOCK sPLS-DA; A), and when performed on single-omics data (MINT sPLS-DA from pMAGs - B; and viruses - C). Dashed lines (A-C) show model classification accuracy averaged across sectors. Venn diagrams show agreement between zoning biomarkers selected in single-omics (MINT sPLS-DA, Component 1) and multi-omics (MINT BLOCK sPLS-DA, Component 1) approaches for (D) pMAGs and (E) viral contigs.

4.5 Conclusions

This study provides the first systematic assessment of seawater viral communities as potential indicators of reef zoning status across the GBR. While viral and microbial (pMAG) biomarkers independently predicted reef zoning with comparable accuracy (~72%), integrating viral and microbial abundance data to identify multi-omics zoning signatures revealed a trade-off with predictive performance declining to ~59% classification accuracy, yet revealing more holistic pMAGs-virus ecological interactions. This trade-off suggests that the choice between single-omics and multi-omics approaches is context-specific, dependent on whether the research goal is pure predictive power in reef monitoring, or a more nuanced understanding of microbe-to-virus associations. Overall these results confirm that viruses are useful, yet complex and context-dependent indicators of reef health. This complexity suggests that for

large-scale monitoring applications, more holistic (and simpler) metrics may be more effective than tracking specific viral indicator taxa. For instance, metrics like total viral abundance or virus-to-microbe ratios (VMRs) could provide robust and scalable indicators of reef health, as high VMRs strongly predict coral cover by reflecting elevated microbial turnover rates through viral lysis, effectively serving as a direct readout of carbon recycling to the dissolved organic matter (DOM) pool[364,366,367]. These integrated metrics may offer distinct advantages by bypassing the noise introduced by context-dependent virus-host interactions and temporal lags inherent in snapshot sampling. To fully leverage the power of viral indicators, future research should prioritise: (1) comprehensive host-virus mapping to clarify context-specific interaction networks, (2) time-series sampling to capture dynamic lytic/lysogenic phase transitions, and (3) expansion of reference databases to enhance viral annotation and functional interpretation. For example, a key mechanistic question arising from this work is whether viral infection strategies (specifically "Kill-the-Winner" lytic regulation versus "Piggyback-the-Winner" lysogeny) shift predictably across reef management zones, which would require testing beyond metagenomics (e.g., annotating viral metagenomes for lysogeny markers like integrase, excisionase, and repressor genes) towards experimental validation, such as induction assays to quantify prophages and direct measurements of viral production rates. Combining these approaches would clarify how environmental conditions mechanistically shape viral regulation, adding a causal, process-oriented layer to viral monitoring frameworks.

4.6 Declarations

4.6.1 Ethics approval and consent to participate

Samples were collected under the permit G12/35236-1 issued by the Great Barrier Reef Marine Park Authority.

4.6.2 Consent for publication

Not applicable.

4.6.3 Availability of data and material

Sequencing data and primary metagenomic assemblies have been uploaded to the European Nucleotide Archive (ENA) under project accession PRJEB82623 under the project name "Great Barrier Reef seawater microbiomes genome database". The 5,283 prokaryotic metagenome-assembled genomes (pMAGs) are accessible from Zenodo (DOI: 10.5281/zenodo.17109887).

Metagenomic analysis including for the prokaryotic and viral planktonic fractions is described in Robbins et al. (2025). All additional code including microbial community analysis, and indicator analysis using single- and multi-omics approaches is available at: https://github.com/mterzin/fishy_microbes_and_viruses (Terzin, 2025).

4.6.4 Competing interests

The authors declare no competing interests.

4.6.5 Funding

This study forms part of the Australia's Integrated Marine Observing System (IMOS) Great Barrier Reef Microbial Genomic Database sub-facility (GBR-MGD), funded by the Queensland Research Infrastructure Co-investment Fund (RICF) by the Department of Environment and Science, Queensland. IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent. This study was also funded by an AIMS@JCU PhD Scholarship to MT. The funders had no role in sampling design, data collection, processing and interpretation, preparation of the manuscript, or decision to publish.

4.6.6 Authors' contributions

NSW obtained funding for the project. NSW, DGB, PWL, RKG, and SJR conceived the sampling design. SCB collected seawater in the field (for metagenomics and physicochemical data) and processed all samples in the laboratory for metagenomic sequencing. Metagenomics analyses for pMAGs (including hybrid assembly, binning, taxonomic annotation, and abundance estimation) and viruses (including assembly, taxonomic annotation, and abundance estimation) were performed by SJR, YKY, KED, and JZ, with the guidance of PH, PWL, and DGB. MT performed all subsequent analyses including microbial community analysis, indicator analysis, multi-omics integration, and data visualisation, with assistance from PWL, KALC, YKY, DGB, SJR, SC, RKG, and MJE. MT wrote the original draft of the manuscript, and all authors made substantial contributions to its form.

4.6.7 Acknowledgements

The seawater samples analysed in this study for metagenomics and physico-chemical variables were collected across 48 reefs, from the sea country of various Indigenous groups who are Traditional Owners (TOs) of that land. We acknowledge the Gudang Yadhaigana TOs, custodians of the McSweeney, Monsoon, 11-049, and 11-162 reefs, which lie within their sea country estate. We pay our respects to the Kuuku Ya'u TOs of the Mantis and Lagoon reefs, and the Lama Lama TOs of the 13-124, Davie, and Corbett reefs in the western half of their territory. We also recognise the Cape Melville, Howicks, and Flinders Island TOs of the eastern half of Corbett and Sand bank #1 reefs, as well as the Eastern Kuku Yalanji TOs of St Crispin and Agincourt #1. We extend our respect to the Yirrgandji TOs of Hastings reef and the Gunggandji as TOs of Arlington, Thetford, and Moore reefs. We acknowledge the Gunggandji-Mandingalbay Yidinji TOs of McCulloch, Hedley, Peart, and Feather reefs, and the Mandubarra TOs of Farquaharson Reef. We honour the Girringun Aboriginal Corporation TUMRA for their connection to Taylor Reef, and the Manbarra TOs of Rib, Kelso, Little Kelso, and John Brewer reefs. We also recognise the Wulgurukaba TOs of Myrmidon, Grub, and Helix reefs, the Bindal Traditional Owners of Knife, Fork, Centipede, Chicken, and Lynchs reefs, and the continuing connection of the Manbarra TOs to Roxburgh, and Fore and Aft reefs. Lastly, we acknowledge the PCCC TUMRA for their stewardship of North, Bloomfield, Eskine, Mast Head, Hoskyn, Fairfax, and Boult reefs. We pay our respects to their Elders, past, present, and emerging, and acknowledge their enduring connection to land and sea. Further, our desktop / lab research took place at the Australian Institute of Marine Science (AIMS) headquarters at Cape Ferguson, and we wish to acknowledge the Wulgurukaba and Bindal peoples as the Traditional Owners of that land. This research was also undertaken at the JCU Townsville Bebegu Yumba campus, and the authors acknowledge that the Australian Aboriginal and Torres Strait Islander peoples are the original inhabitants and traditional custodians of this continent and have unique cultural and spiritual relationships to the land and waters. We acknowledge the AIMS Water Quality team, especially Ulysse Bove, Keeley Glasson, and Daniel Moran for logistics, training, and processing of water chemistry

samples. We acknowledge the AIMS-LTMP team and others involved in field collection and preparation of samples including Emmanuelle Botté, Johnston Davidson, Veronique Mocellin, and Josephine Nielsen. We thank the crew of the RV Solander and RV Cape Ferguson for their excellent logistical support in the field. We also acknowledge Gene Tyson for his support in facilitating the use of the NovaSeq at Microba Life Sciences Ltd. (Brisbane, QLD, Australia). We extend our gratitude to Murray Logan for his insightful discussions on the appropriate statistical handling of the data. AI tools (ChatGPT and DeepSeek) were used exclusively for proofreading (grammar, syntax, and clarity checks) and code assistance (debugging and statistical script optimization). No AI tools were used to generate original content, interpret data, or formulate conclusions. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2025648). MT extends his gratitude to the members of the Lê Cao Lab at Melbourne Integrative Genomics (MIG) for the supportive environment and valuable scientific discussions, particularly Vinicius Salazar, Saritha Kodikara, and Jiadong Mao.

5 Chapter 5

General Discussion | From discovery to application: integrating seawater microbial assays into coral reef monitoring programs

This PhD thesis directly contributes to the establishment of the Great Barrier Reef Microbial Genomics Database (GBR-MGD) by Australia's Integrated Marine Observing System (IMOS): a comprehensive infrastructure resource containing: (1) 5,283 seawater prokaryotic (bacterial and archaeal) metagenome-assembled genomes (pMAGs); (2) 20 chromosome-level picoeukaryote genomes; (3) 808,585 viral genomes; and (4) 24,769 mobile elements (putative plasmids), collected alongside a range of contextual environmental parameters[270]. Through integration of GBR-MGD microbial and environmental data (such as water chemistry, benthic cover, and fish abundance/biomass), my research provides a foundation for understanding the role of reef bacterioplankton in ecosystem functioning across offshore GBR reefs. As the main thesis focus, I explored the predictive (rather than purely descriptive) potential of seawater microorganisms for coral reef monitoring, emphasising the need to move beyond the 'indicator discovery' phase to develop actionable tools for reef management. In this chapter, I outline key priorities, opportunities, and challenges for integrating seawater microbial observations into reef monitoring programs, offering a forward-looking perspective on how these data can enhance real-time decision-making.

Box 1 | Thesis highlights

Chapter 1: This literature review shows that, after nearly two decades of research, the application of seawater microbes in reef monitoring is hindered by the lack of a unified framework that would move beyond descriptive indicator discovery and toward creating rapid assays that support (near) real-time management decisions. A comprehensive framework is proposed to move from descriptive to predictive reef monitoring, defined by five phases of research, catalysing a shift from fundamental research to applied approaches, thereby shifting from reactive to proactive reef management.

Chapter 2: By analysing seawater metagenomes from 48 offshore reefs across the Great Barrier Reef, this chapter integrates read-based microbial metagenomics with 17 physico-chemical environmental variables to compare the diagnostic potential of microbial taxonomic and functional information in reef monitoring. This work establishes a functional baseline for seawater microbiomes in the GBR, and demonstrates that microbial functional genes associate twice as stably to reef physico-chemical conditions than taxonomic data. While these results suggest microbial functional traits may better capture reef-scale environmental variability, taxonomic indicators inferred from cost-effective 16S rRNA amplicon sequencing remain the practical gold standard for monitoring, with functional meta-omics serving as a valuable complement when resources allow.

Chapter 3: We provide the first demonstration of an ecosystem-wide effect of fisheries management on the microbial communities of the surrounding seawater, showing that No-Take Marine Reserves (NTMRs) on the offshore Great Barrier Reef host distinct seawater microbial communities compared to fished reefs. Microbial communities reflected broader reef states, with NTMRs enriched in streamlined microbial oligotrophs (*Pelagibacter* and SAR86 MAGs) correlating with higher cover of hard coral, crustose coralline algae, and herbivore abundance under lower nutrient conditions, while fished reefs harbored opportunists (Flavobacteriales, especially UA16, and Pseudomonadales) responding cumulatively to elevated nitrite, nitrate, and turf algae cover. Microbial indicators accurately predict reef zoning (71% accuracy) and correlate with key environmental metrics, offering a novel tool for monitoring reef health and conservation outcomes.

Chapter 4: In Chapter 4, we pursued a key strategic question: can integrating viral data enhance the predictive model built on prokaryotic taxa? We identified viral indicators that, alongside prokaryotes, distinguish NTMRs from fished reefs, however this data integration did not substantially improve upon the 71% model accuracy achieved by prokaryotes alone. This result underscores that integrative omics analysis requires strategic consideration, and the added complexity of multi-omics approaches must be weighed against their marginal gains for specific monitoring goals.

5.2 Monitoring potential of reef seawater microbiomes: moving from description towards prediction

Coral reefs globally are experiencing alarming declines due to a combination of local disturbances, such as eutrophication and overfishing, and global pressures from climate change. As these ecosystems continue to degrade, the early identification of environmental stressors and declining reef health becomes critical for implementing effective mitigation and management strategies. Marine microorganisms, including bacteria and archaea, have been proposed as valuable tools for early diagnostics of reef health, owing to their short generation times and rapid responses to adverse environmental conditions and degraded reef health[263,264,388]. Previous studies have suggested that seawater microbes can outperform sediment and host-associated microbiomes (such as those of corals, sponges, and macroalgae), showing up to a fivefold increase in accuracy when predicting fluctuations in temperature and nutrient levels in the surrounding reef[304,349]. Specifically, opportunistic microbes in seawater have shown strong associations with indicators of reef degradation, including poor water quality, reduced coral cover, and increased macroalgae abundance. Nevertheless, seawater microbes are often criticised as being merely descriptive, with benthic indicators such as coral and algal cover still considered more relevant endpoint metrics for monitoring reef condition.

Seawater microbes have a fundamental role in the functioning of coral reef ecosystems, making them potentially valuable not only for describing current stress but also for forecasting future trajectories in reef health and functioning. Environmental changes can alter bacterioplankton functions (photosynthesis, nitrification, sulfate reduction, and broader nutrient cycling) and disturbance-triggered shifts in heterotrophic microbial activity are hypothesised to push keystone benthic organisms toward the limits of their resilience, leading to disruptions in biogeochemical cycling that cascade through marine food webs and affect entire reef ecosystems[263,389]. For example, the cumulative effects of nutrient eutrophication and elevated temperatures can trigger heterotrophic microbial activity in seawater, leading to harmful algal blooms and hypoxic conditions at reef scales, which can result in rapid coral mortality[390–392]. Despite their potential for prediction, the use of seawater microbial communities in forecasting coral declines and reef ecosystem health remains largely unexplored[393]. To address this knowledge gap, my thesis explores the potential of seawater microorganisms as predictive (rather than purely descriptive) tools for monitoring coral reef health. Based on the results presented, I was able to rate the predictive value of free-living seawater microbial communities according to: (1) scalability for rapid and cost-effective monitoring, evaluated through the comparison of read-based and genome-based metagenomic approaches; (2) the stability of microbiome-environment signatures across broad spatiotemporal scales; and (3) the role of microbial generalists versus specialists in predicting reef condition.

Chapter 1, published as Terzin et al. (2024)[267], highlights the challenge of lengthy bioinformatics processing pipelines, with meta-omics data typically requiring months (or even years) to analyse, thereby

limiting their utility for informing real-time reef management. The utility of microbial indicators is diminished if the information they provide refers to poor ecosystem health that occurred months or years earlier, during field sampling[394]. Additionally, the high costs of high-throughput sequencing further hinder the routine application of microbial data in reef monitoring. To address this in Chapter 2, published in *Microbiome*[281], we employed a read-based metagenomics approach (directly annotating raw sequencing reads against reference databases) as a faster and more cost-effective alternative to genome-centric strategies, which rely on read assembly, contig binning, and the reconstruction of metagenome-assembled genomes (MAGs)[395]. Read-based metagenomics reduced analysis time from months to weeks and was applied to a subset of short-read Illumina data, representing just 20% of the total sequencing cost of the hybrid Illumina–Nanopore dataset used in the GBR-MGD.

Apart from lowering sequencing costs and analysis time, read-based metagenomics also enabled the separation of taxonomic and functional signals, allowing us to directly assess whether microbial taxa or functional genes were more stable predictors of surrounding reef physico-chemical conditions. Using indicator stability scores derived from sPLS models, microbial functions showed approximately twice the stability of microbial taxa in their associations with 17 environmental variables across 48 reefs, supporting the idea that functional traits in pelagic microbial communities are more tightly coupled to environmental conditions than taxonomy. This finding aligns with the concept of functional redundancy, which posits that multiple microbial taxa can perform similar ecological roles, leading to greater consistency in functional profiles despite taxonomic turnover[235,396]. These results provide empirical support for the hypothesis that microbial functional traits offer enhanced predictive value for environmental monitoring than taxonomic composition alone, and I propose that microbial function should be a central focus in future efforts to develop microbial-based indicator assays of reef health.

In Chapter 3, the stability of microbiome–environment associations was further investigated using P-integration omics approaches, which combine data from the same omics layer across different samples, populations, or conditions to identify broader, more generalizable patterns[2,294]. This thesis is the first to have applied P-integration (specifically MINT-sPLS) on environmental microbiology datasets, which allowed us to identify microbial indicators that were consistent across GBR sectors and predictive of reef zoning status (NTMRs vs. fished reefs) with an average classification accuracy of 71%. These microbes not only discriminated between zoning categories (as categorical outcomes) but also exhibited stable associations with multiple continuous environmental metrics, underscoring the diagnostic potential of microbes for integrative reef monitoring across space and time in the GBR. Specifically, NTMRs were enriched in streamlined microbial oligotrophs (such as *Pelagibacter* and SAR86 MAGs) that were consistently associated with higher hard coral and crustose coralline algae cover, increased herbivore abundance, and lower concentrations of dissolved inorganic nutrients. In contrast, fished reefs harbored higher relative abundances of opportunistic taxa, including Flavobacteriales (particularly UA16) and Pseudomonadales, which showed cumulative responses to elevated nitrite, nitrate, and turf algae cover.

Building on these stable microbial-environment associations, random forest predictions and microbial niche modeling were then used to test whether seawater microbes could predict continuous environmental variables. High prediction accuracy was achieved for temperature, salinity, particulate nutrients, and dissolved inorganic phosphorus, while metrics such as dissolved nitrogen, silicate, benthic cover, and fish biomass were predicted with lower accuracy. The high prediction accuracy for temperature and particulate nutrients, was driven by specialised taxa like *Flavobacteriales,* well-documented to associate to elevated temperature and nutrients[281,304,348]. This contrasts with the poor performance for dissolved nitrogen (likely due to rapid microbial uptake[354]) and benthic cover metrics, where spatial decoupling between pelagic microbial communities and patchy reef benthos[349,356] obscures direct correlations. Collectively, this Chapter provides the first evidence that marine park management measures project down to microbial scales, opening the question of whether the documented ecosystem resilience in protected reefs is underpinned by these distinct and stable microbial mechanisms.

With the predictive power of seawater prokaryotes established in previous chapters, Chapter 4 expands the reef monitoring framework to include the most abundant biological entities in the ocean: viruses. Seawater viruses are recognised as key players in reef health, regulating microbial community structure through lysis and lysogeny and driving nutrient cycling via the viral shunt, yet virioplankton remains largely unexplored in the Great Barrier Reef (GBR). A crucial, yet understudied, aspect is that viral lysis can erase microbial indicator signals by removing key taxa before sampling, and therefore accounting for virus-host interactions could be vital for more robust microbial-based monitoring. This chapter had two primary objectives: first, to determine if seawater viral communities alone could accurately classify reef zoning status (No-Take Marine Reserves vs. fished reefs), and second, to evaluate whether integrating viral and microbial (pMAG) data could improve predictive accuracy and provide a more holistic assessment of reef condition. The analysis revealed that viral communities (dominated by tailed bacteriophages, Caudoviricetes) predicted reef zoning status with high accuracy (72%)—a result nearly identical to that achieved by prokaryotic biomarkers alone (71%). We hypothesise this is due to tight host-virus coupling, whereby the model selected for viral genomic elements already integrated within their host genomes, whose abundances they directly track. Counterintuitively, integrating the viral and prokaryotic (pMAG) datasets into a multi-omics NP integration model (MINT BLOCK sPLS-DA) reduced the overall classification accuracy to 59%. This trade-off was attributed to the inherent biological complexity of virus-host interactions. Given that viral indicators associated with numerous microbial hosts (1-170 hosts per virus) and their abundances are influenced by dynamic, context-dependent infection strategies, it is thus unrealistic to expect that multi-omics signatures (i.e., pMAG-virus correlations) will remain stable across a vast and spatiotemporally heterogeneous seascape of 48 reefs. Despite the decrease in pure predictive power, the integrated model provided superior insight into ecological relevance, showing that viral indicators of NTMRs were linked to putative NTMR-enriched oligotrophic hosts like Pelagibacteraceae and SAR86, while viruses indicative of fished reefs correlated with microbial indicators of fished reefs such as Rhodobacteraceae and Flavobacteriaceae. This precise

alignment confirms that the discriminative power of the viral community is executed through specific interactions with zoning-responsive prokaryotes. In conclusion, the decision to use a single-omics or multi-omics approach depends on the monitoring objective: for pure classification accuracy, viral indicators are equally effective as microbial ones; for a mechanistic understanding of the underlying virus-host interactions that drive ecosystem change, an integrated approach is essential. These results confirm that viruses are valuable yet complex components of the reef monitoring toolkit, and future efforts should prioritise time-series sampling and host-virus mapping to fully leverage their predictive potential.

Overall, this thesis provides an analysis framework to identify niche-specialised taxa and functions that show strong and consistent associations with environmental conditions across broad spatiotemporal scales, enabling the discovery of robust microbial indicators that can be contextualised with co-located water chemistry and benthic observations, potentially beyond the GBR.

Box 2 | Strategic Recommendation: Combining Gene- and Genome-Centric Metagenomics for Reef Monitoring

Gene-centric and MAG-centric approaches offer complementary advantages (Table 1), and both should be employed to maximise the value of microbial metagenomics data in reef monitoring. Based on my PhD thesis results, I propose the following strategic workflow:

1. **Start with gene-centric, read-based analysis as a rapid, exploratory tool to characterise broad microbial patterns** and identify taxa and functions responding to environmental stressors. These results can inform hypothesis-driven selection of candidate indicators while also allowing for early data integration, ensuring microbial and environmental datasets are formatted appropriately for rapid re-analysis once MAGs become available.

2. **Run MAG-centric analyses in parallel to assemble and bin microbial genomes from reef environments**. This is a necessary step to identify novel microbial taxa and understand their functional roles, especially relevant in reef ecosystems where functional diversity and ecological importance of microbial bacterioplankton is still poorly resolved. This is exemplified by the Tara Pacific expedition which highlights the extraordinary and largely uncharacterised microbial diversity (inferred from 16S rRNA amplicon sequencing) in coral reef plankton across the Pacific (Galand et al. 2023). Our results from the GBR also suggest reef environments may harbour unique or under-sampled microbial taxa that can be revealed through genome-resolved metagenomics, as the comparison of species-resolved GBR-MGD pMAGs (dereplicated at 95% ANI) to ~35,000 pMAGs in the Ocean Microbiomics Database revealed that 65% of GBR-MGD microbial genomes were novel (Robbins et al. 2025).

3. **In subsequent monitoring, shallow metagenomic or amplicon sequencing can be used** to map new data to the previously reconstructed MAGs and obtain rapid 'snapshot' assessments of ecosystem health based on previously identified indicator taxa and functions.

This framework should initially be deployed with established reef monitoring programs like the Long-Term Monitoring Program (LTMP) to ensure microbial and environmental data are collected in tandem. However, once robust microbial indicators of reef health are identified, microbial screening could provide a useful complementary toolkit for profiling reefs not covered by traditional surveys, in addition to assessing reef health following environmental disturbances (such as bleaching events, cyclones, outbreaks of crown-of-thorns starfish, or pollution incidents) where microbial communities can offer sensitive, early insights into ecosystem stress and recovery.

Table 5.1: An overview of advantages and disadvantages of gene-centric and MAG-centric analysis approaches in ecosystem monitoring purposes. Instances when a certain methodology performs better are marked in green.

Read-based metagenomics analysis	MAG-centric metagenomics analysis
Rapid profiling (weeks) of microbial taxonomic and functional composition once pipeline is established.	Long processing times (months), limiting rapid decision-making.
Captures signals from microbes difficult to assemble into MAGs (e.g. low-abundance taxa, high-strain heterogeneity), as well as marine protists[397] with large genomes and repetitive elements.	Smaller fraction of the community analysed, as MAG recovery limited to high-quality, complete genomes. We show long-read metagenomics can overcome these limitations[270].
Cannot detect novel taxa or genes not present in reference databases.	Enables discovery of novel microbial taxa by reconstructing (near) complete genomes, and their full genomic repertoire.
Link between taxonomy and function is weak because the unassembled reads are too short to infer which microbial populations they originate from.	Strong linkage of taxonomy and function via near-complete genome reconstruction. This allows us to investigate the role of individual microbial species in the environment.
Unassembled reads are too short to annotate microbes to species/strain level.	High-resolution taxonomic assignment (e.g., species or strain level).

5.3 Mining for seawater indicators from newly emerging large-scale meta-omics surveys on reef bacterioplankton

Pelagic microbiomes are now among the best-characterised microbial systems in terms of comprehensive molecular (meta-omics) profiling, due to advances in the field of metagenomics[398,399] and global sampling efforts of oceanic microbiomes such as the Global Ocean Sampling expeditions[399] (2003–2010), the Malaspina expedition[400] (2010–2011), the Tara Oceans[401,402] expedition (2009–2013), and ~72 long-term marine microbial observation stations catalogued by Buttigieg et al. (2018)[240]. For the first time, large-scale meta-omics surveys are also generating substantial data on reef-associated seawater microbes (**Fig. 5.1a**), alongside a comprehensive set of environmental parameters from regions such as the Pacific Ocean[349,389,403–406], the Florida Keys[407], and the Great Barrier Reef[270,281]. Most notably, the Tara Pacific

expedition is widely recognised as having generated the largest environmental dataset to date, collecting 58,000 samples across the Pacific Ocean, comprising approximately 102 terabytes of metabarcode, metagenomic, and metatranscriptomic data, along with over 5,000 metabolomic profiles and 3.8 million environmental data points[408]. While reef-associated bacterioplankton have been studied in smaller-scale surveys for approximately two decades in the context of reef health[264,265,267,388,389], this newly emerging global-scale data now allows for the study of reef bacterioplankton responses to environmental change on unprecedented spatio-temporal scales. However, a key question arises: can this data be integrated to establish baselines of seawater indicators for reef health that will be coherent across space and time?

This thesis is the first to have applied P-integration[2,294] on environmental microbiology datasets (Chapters 2, 3, 4), an integrative omics analysis that combines data from the same omics layer across different samples, populations, or conditions to identify broader, more generalizable patterns. P-integration allowed us to identify seawater microbial indicators of water chemistry (Chapter 2) and reef zoning (Chapters 3 and 4) that are stable across space and time in the Great Barrier Reef (GBR). Considering the newly emerging and large scale meta-omics data on reef bacterioplankton, future analysis efforts can explore the ubiquity of these patterns and if they extend beyond the GBR (**Fig 5.1a**). While P-integration is designed to deal with heterogeneous batch effects[294], it will nevertheless be essential to ensure the molecular data is standardised as much as possible[409], for example by reprocessing the raw meta-omics data from each study to homogenise normalisation techniques and account for batch effects (i.e. different laboratory techniques and data processing/sequencing platforms). Perhaps even more critical is the need for standardised seawater sampling methodologies, as physical parameters like pre-filtration can systematically exclude entire microbial groups (e.g., filamentous cyanobacteria including *Trichodesmium*), thereby confounding cross-study comparisons. The success of global ocean surveys like the Tara Oceans expedition, which developed best practices for data standardisation and interoperability, exemplified by the Marine Microbial Biodiversity, Bioinformatics and Biotechnology (M2B3) reporting standard for documenting sampling campaigns, stations, and events[410], serves as a powerful model for this standardised approach—with the same protocols later expanded upon in the coral reef-focused Tara Pacific project[404]. Additionally, non-standardised collection and processing methods for environmental data (like water chemistry and *in situ* benthic cover metrics) will also pose a major challenge for these meta-analysis efforts, although initial attempts could utilise metrics like coral and macroalgae cover as those variables have been collected in a standardised manner and have facilitated a global analysis previously[411].

Despite the inherent challenges of heterogeneous data (e.g. data sets of different formats, complexity, dimensionalities, information content, and scale), integrative meta-analysis of emerging global datasets on reef-associated plankton represents an important research frontier in the future. The application of artificial intelligence (AI) to globally collected reef water microbial data has recently been proposed to learn patterns and develop a simple reef water microbial health index[393]. However, many AI

models remain challenging to interpret and often function as "black boxes," making it difficult to extract ecological meaning from learned patterns. Therefore, AI will likely need to be integrated with other computational frameworks, including multi-omics data integration[292,294,386], ecological niche and associations distribution modelling[235,306,412], network analysis[306,374,413], multivariate statistics, and machine learning, to robustly identify and validate seawater microbial indicators that are ecologically informative and stable across large spatio-temporal gradients. Such integrative analysis (**Fig. 5.1b–e**) also has the potential to uncover patterns that remain invisible in single-omics studies. For example, a global multi-omics study on pelagic microbes reveals that oligotrophs like SAR11 rely on community turnover (tracked via metagenomics), while copiotrophs use transcriptional plasticity (tracked via metatranscriptomics) to respond to environmental changes[414]. Furthermore, activity (transcript abundance) is often decoupled from genomic abundance, a pattern demonstrated by picocyanobacteria that are transcriptionally hyperactive (relative to their abundances inferred from metagenomes), while heterotrophs like SAR11 show the inverse trend[415,416]. These newly-emerging large-scale meta-omics data will likely also generate novel, data-driven hypotheses and reveal ecological patterns and reef (bacterio)plankton functions that were previously unseen, simply because the necessary data did not exist until now.

Lastly, while these functional meta-omics data streams represent an important step towards establishing spatial baselines of reef seawater microbes, data in these large-scale initiatives were taken only once in space and time, and thus future sampling efforts should prioritise establishing a reef-associated long-term microbial observatory, which is still lacking, unlike the 72 long-term microbial observatories that were catalogued[240] for pelagic microbes. Such temporal data will be needed to provide invaluable insights into reef microbial baselines and community dynamics in relation to biogeochemical parameters, natural anomalies, and anthropogenic disturbances.

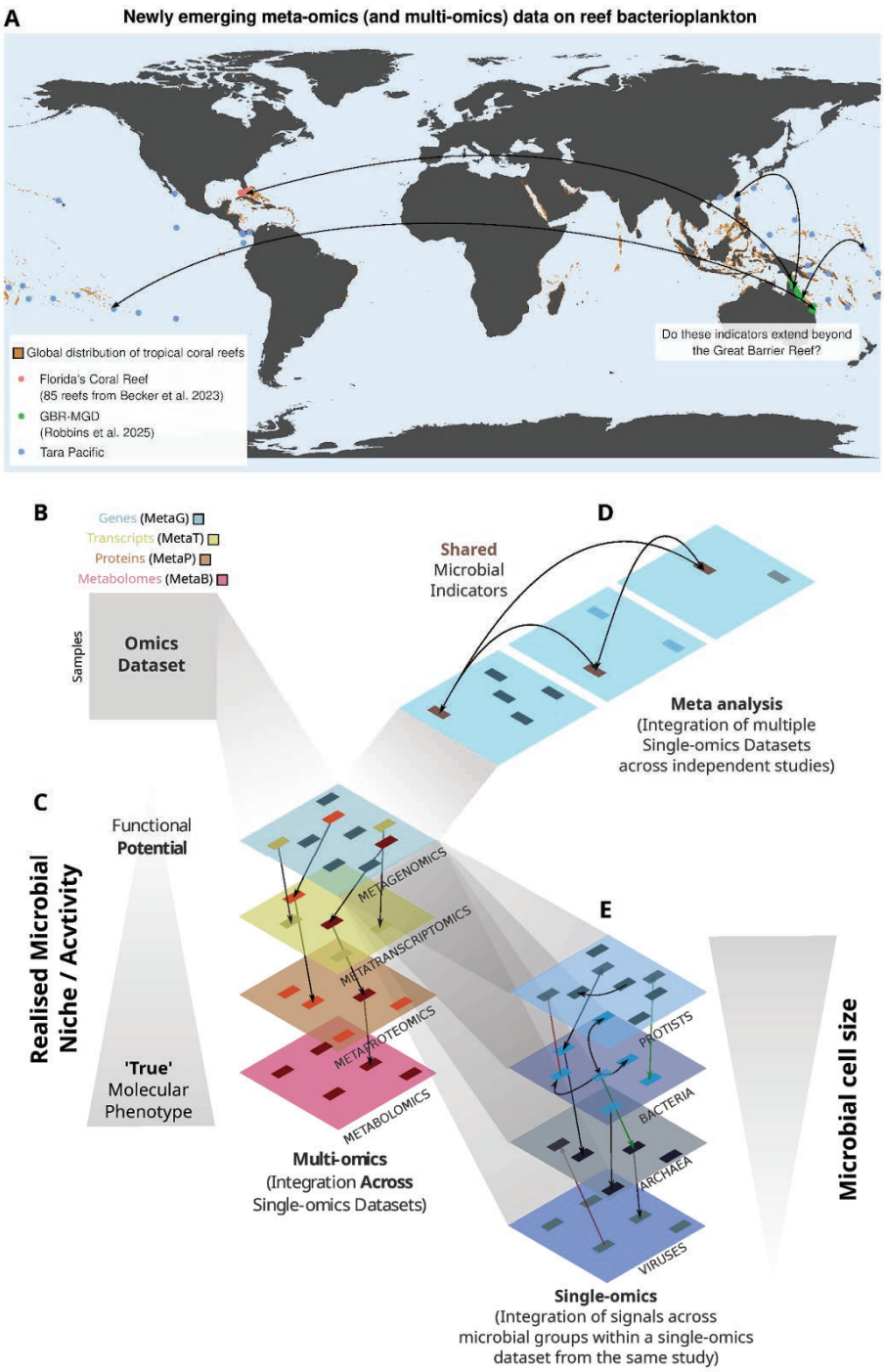

Figure 5.1: Data integration will be critical to identify novel microbial indicators of reef health from globally emerging data on reef bacterioplankton. (**A**) Global coral reef distribution and recent large-scale seawater microbial surveys. (**B**) Omics data can be represented as matrices of samples (rows) and features e.g genes, transcripts, proteins, metabolomes) in columns. Integration can occur (**C**) across different omics types measured on the same samples (N-integration) to for example connect genetic potential with molecular activity, or (**D**) within a single omics type to identify universal indicators via meta-analysis (P-integration). (**E**) Viral, prokaryotic and eukaryotic microbial 'layers' can potentially also be partitioned from a single-omics study (e.g., metagenomics) to investigate how environmental change affects microbe-to-microbe interactions.

5.4 Moving beyond sequencing to allow rapid decision-making

The generation of global[63,72,143,270,349,389,404,406] reef bacterioplankton data offers the opportunity for coordinated meta-analysis to identify spatiotemporally stable microbial indicators of reef health, an ambitious but achievable goal that will likely take years to realise. Although read-based metagenomics offers a comparatively faster and more accessible alternative to genome-resolved methods (Chapter 2), omics-based approaches still depend on laboratory protocols and sequencing turnaround times that can take weeks (if not months) from field sampling to data analysis and interpretation. While rapid 16S rRNA-based workflows with turnaround times as short as 10 days have been demonstrated in urgent outbreak contexts[417], and portable sequencers like the MinION enable near-real-time genomics in the field, such approaches are not yet the norm for routine monitoring. To truly enable real-time, field-deployable decision-making, we must begin to move beyond traditional sequencing workflows. Below, I outline key priorities and challenges that must be addressed to support the development of cost-effective, rapid microbial assays, and integrating microbial dynamics into predictive frameworks capable of supporting early intervention and management under varying disturbance pressures.

5.4.1 Identify problem and propose concrete scenarios where microbial data adds value to reef monitoring

Microbial-based reef monitoring, though promising, faces challenges similar to those in ecotoxicology, where microbial indicators are often critiqued as indirect proxies of broader ecosystem health. This is because measurements like coral cover or species diversity are typically easier to interpret and more directly aligned with management objectives. Moreover, many existing models can already accurately predict large-scale reef dynamics (e.g., coral bleaching predictions based on Degree Heating Weeks - DHW) without microbial inputs, raising the question of when microbial observations truly add value.

The key challenge for future research, therefore, is to identify specific microbial signals (whether functional or compositional) that provide unique or earlier insights that can complement conventional metrics. For example, while DHW can predict bleaching events effectively, this metric does not capture the increased vulnerability to coral disease that may follow post-bleaching[418–420]. Opportunistic microbes can persist in seawater after marine heatwaves[421], potentially increasing disease risk (as bleaching-induced thermal stress may have already compromised coral health) and compounding coral mortality. In such cases, microbial monitoring may be most valuable post-disturbance, potentially offering predictive insight into coral recovery or future disease outbreaks.

Similarly, the relevance of microbial indicators varies across geographies. In the GBR, herbicide and pesticide exposure from river runoff is a major concern[422], whereas heavy metal contamination is more prevalent in the Red Sea[423,424], and pathogen-driven coral diseases dominate in Florida[425]. Microbial tools could be tailored to each of these region-specific contexts, as well as depending on the research question,

which will impact the sampling strategy. For example, near-organismal seawater sampling can help identify health concerns for specific reef organisms, particularly when disease is suspected, whereas surface seawater sampling is likely more appropriate for assessing broader ecosystem functioning[393].

To conclude, microbial data will not replace conventional metrics but can complement them strategically when targeted at region- and context-specific issues, such as pathogen surveillance post-bleaching, or early detection of pollution-related functional disruptions. Furthermore, it offers a potential solution for monitoring reef ecosystems where safety concerns (e.g., dangerous animals, murky waters), difficult logistics (e.g., strong currents), or legally enforced restrictions (e.g., active shipping channels or mining sites) prohibit conventional in-water surveys like scuba diving. This sets the stage for developing practical microbial tools to translate such insights into reef management action.

5.4.2 Collect the most relevant microbial information to develop assays for rapid reef health assessments

Once concrete and context-specific scenarios are identified to complement and enhance current reef monitoring efforts with microbial observations, the development of rapid, field-ready microbial assays will be essential. Given the complexity of seawater microbial communities (comprising billions of cells and thousands of taxa per liter of seawater), some degree of coarse-graining will be required. Rather than aiming to resolve the full microbial diversity or all metabolic pathways, future assays should prioritise more basic yet ecologically informative microbial metrics, such as microbial abundance/counts which are in spreadsheet form and are thus easy to analyse[393], as well as microbial biomass, productivity, and enzymatic activity.

Microbial functions, rather than taxonomic profiles, should be prioritised in assay development due to their higher stability across spatial and temporal gradients and enhanced performance in predicting physico-chemical metrics (Chapter 2). Functional simplification can be achieved by grouping microbes based on key functional traits (e.g., anaerobic metabolism, sulfate reducers, nitrogen fixers, methanogenesis) or stress-response strategies (e.g. copiotrophic microbes rely on transcriptional changes to respond to changing environments, while oligotrophic microbes rely on community turnover[414]), a technique long used in Earth System Models to distill biological complexity[426–428]. In a less categorical fashion (i.e. the presence of aerobic versus anaerobic microbial functions which can distinguish environments with high versus low oxygen availability), microbial indicators can be expressed as a continuous function of relevant environmental drivers to quantify the extent of environmental impact and proportion of the microbial community responding to stress[429]. The development of such microbial ecotoxicology frameworks has commenced and sensitivity curves were derived from 16S rRNA gene amplicon data that can predict the fraction of a reef bacterioplankton community affected by copper exposure, with derived microbiome Hazard Concentrations (mHCx values) revealing thresholds 2-fold lower than eukaryotic-based guidelines[430,431]. These microbial ecotoxicology approaches can be extended to

quantify the loss of microbial functions under environmental change, an effort that holds significant value for identifying ecological thresholds beyond which microbial processes relevant to reef functioning may be lost[429], or to detect if disruptions to specific pathways could help prevent ecosystem tipping points. For instance, microbial functional shifts in response to eutrophication or climate change can drive oxic–anoxic regime shifts with cascading detrimental effects on ecosystem health[391].

Several microbial assay platforms (PCR screening, immunomagnetic separation, and colorimetric or enzymatic assays) are already well established in environmental science[237]. These have been used to detect microbial indicators such as (1) coliform bacteria for faecal contamination in public swimming waters[432,433], (2) antibiotic resistance genes as indicators of human impact[434], and (3) hydrocarbon-degrading taxa and genes for oil spill tracking[435]. These assays could be adapted for reef applications. Importantly, since DNA-based methods detect both dead and viable cells, activity-based assays targeting transcripts, ATP, or enzymatic/proteomic markers may offer greater ecological relevance. For example, a quantitative proteomic assay was developed to target nutrient stress (i.e. starvation) in *Prochlorococcus*[436], which could serve as a blueprint for reef-relevant indicators too. Assays targeting highly abundant taxa (such as *Prochlorococcus*) are particularly promising, as these microbes are more likely to be consistently detected across samples. Based on this thesis, the abundant SAR11 clade (i.e., *Pelagibacter*), comprising up to 25% of marine microbial cells[437,438], may serve as a robust microbial indicator for no-take marine reserves (NTMRs) in the Great Barrier Reef (Chapter 3). Another promising approach is immunomagnetic separation, in which magnetic beads coated with antibodies selectively bind to multiple target bacteria (e.g., *E. coli*) in a water sample. After magnetic isolation, the collective activity of captured bacteria is quantified by measuring ATP levels via bioluminescence in a luciferin–luciferase assay, returning results in hours[432].

Parallel to rapid and cost-effective field-ready assays, emerging web-based tools also offer promising platforms for real-time reef health surveillance. For example, the "*Vibrio* map viewer tool" is a near real-time model that uses daily updated remote sensing data to examine the environmental suitability for *Vibrio* growth in the context of climate change[439], though currently restricted to specific regions like the Baltic Sea, suggesting the model does not extrapolate well to other global ecosystems. The Ocean Gene Atlas is another web platform that provides a global interface for exploring plankton gene distributions interactively, using environmental variables of interest[440,441]. If adapted for coral reef ecosystems, such tools could greatly enhance reef management and forecasting. In the GBR, integration of microbial data into existing platforms like eReefs[442] could, in principle, enable near real-time detection of microbial community shifts under changing environmental conditions. However, this remains a distant goal that will require high-resolution temporal sampling and the routine collection of meta-omics data (e.g. metagenomes, metatranscriptomes, and metaproteomes) to support robust, ecosystem-specific models. Such integration would not only advance microbial-based monitoring but also improve

predictions of existing oceanographic models, as microbes contribute directly to key measured outputs such as chlorophyll-*a* concentrations, oxygen levels, and cyanobacterial blooms (e.g., *Trichodesmium*).

Ultimately, translating meta-omics data into standardised metrics with reduced complexity, whether microbial gene ratios, enzymatic activities, or indicator taxa abundances, will be key to developing rapid and cost-effective microbial-based reef monitoring tools. Though still a distant goal, success will depend on transdisciplinary collaboration among microbiologists, ecologists, modelers, and practitioners. Such partnerships will be instrumental in incorporating microbial insights into actionable reef management frameworks, enhancing the capacity of agencies to respond to environmental stressors through evidence-based interventions such as water quality improvement plans, temporal closures of reef sites for commercial and recreational activities, or bioremediation strategies.

5.5 Upscale locally and globally

Australia's monitoring programs for the Great Barrier Reef, specifically the Long Term Monitoring Program (LTMP[249]) and Marine Monitoring Program (MMP[443,444]), are uniquely positioned to integrate microbial indicator assays into their frameworks, as their long-standing, well-established operations ensure long-term, standardised surveys ideal for also tracking microbial trends. Given the practicality in collecting seawater in a non-destructive and easy manner alongside ongoing reef health surveys (reviewed in Chapter 1), the logical next step involves industry partnerships to support logistics and infrastructure investments to facilitate routine seawater collection and processing, with potential for upscaling to other Australian reef systems like Ningaloo and the Coral Sea. This upscaling effort presents a significant opportunity to move beyond top-down scientific approaches and support decolonised, community-led monitoring. Rather than just extracting data, a partnership model could empower Traditional Owner groups and interdependent coastal communities to drive their own monitoring activities, aligning scientific data collection with cultural values and priorities. In addition, citizen science initiatives such as UNESCO's Environmental DNA (eDNA) global expeditions[445] demonstrate how public participation can also scale data collection while raising conservation awareness. Microbial monitoring programs for coral reefs could adopt this model, potentially co-designed with communities from the outset to ensure it supports their sovereignty and environmental goals.

Beyond national programs, upscaling assays globally can be leveraged through international funding mechanisms like CORDAP (Coral Reef Research and Development Accelerator Platform) and the Global Fund for Coral Reefs, which prioritise scalable conservation technologies. The relevance of rapid and cost-effective microbial assays is particularly evident in reef nations with limited marine conservation budgets, where cheap tools could revolutionise reef monitoring. Strategic partnerships, supported by programs such as the GEF Small Grants Programme (focused on community-led marine conservation) or the World Bank's PROBLUE (for blue economy innovations), could facilitate capacity-building, shared protocols, and development accessible technologies (e.g., portable sequencers or simplified assay kits). By

aligning microbial monitoring with SDG 14 'Life Below Water' as both a scientific and equitable stewardship tool, we can bridge reef monitoring with community-engaged conservation, using microbial assays to validate local observations while providing early warnings of reef stress across diverse socioeconomic contexts.

The potential for this framework also extends beyond coral reefs to other keystone marine ecosystems. The principles of using seawater microbiomes for non-destructive, cost-effective health assessment could be readily adapted to monitor temperate kelp forests (e.g., the Great Southern Reef), polar systems (Antarctic and Arctic), or areas experiencing acute ecosystem disturbances. For instance, viral and microbial monitoring could provide crucial, rapid insights into marine disease outbreaks, like the current ecological shifts in South Australia, or the impacts of glacial melt and pollution in polar seas. By leveraging the same seawater sampling logistics, these vulnerable ecosystems could benefit from early-warning systems similar to those proposed for reefs.

5.6 Conclusions

This PhD thesis establishes seawater microbes as valuable tools to complement traditional reef monitoring, identifying functional genes (Chapter 2), NTMR-associated oligotrophic microbes vs. opportunists associated to fished reefs (Chapter 3), and viral indicators (Chapter 4) as novel markers that may provide additional layers of diagnostic information. To translate these findings into management tools, we propose three critical steps: (1) develop scalable function-based microbial assays (e.g., proteomics, ATP tests); (2) integrate assays into ongoing monitoring programs like AIMS-LTMP through industry partnerships; and (3) upscale these assays locally (through partnerships with Traditional Owners and citizen science initiatives) and globally, by leveraging funding initiatives (CORDAP, UNESCO eDNA) to ensure equitable tool implementation across reef nations.

With growing recognition of microbial role in ocean health, we now need regulatory frameworks to protect microbial functions that are ecologically relevant to reefs. This aligns with action calls to develop microbial-based solutions against environmental change[446] exemplified by the United Nations Plankton Manifesto's push for 'Plankton-Based Solutions' using cutting-edge technologies. As this manifesto gained traction at COP29 and beyond, our findings demonstrate how reef-associated bacterioplankton, particularly the functional markers and zoning indicators identified here, can operationalise this vision, transforming microbial insights into concrete monitoring tools for reef ecosystems. Lastly, the convergence of newly emerging global (multi-omics) sequencing data on reef bacterioplankton, novel lab methods (particularly in the field of single-cell biology), and computational approaches (including artificial intelligence) present a unique opportunity to bridge scales from molecules and cells to entire reef ecosystems, an ambitious goal which is likely to reveal previously unseen roles of seawater microbes central to reef health and functioning.

Box 3 | Outstanding Questions - Reef bacterioplankton based solutions in future reef monitoring and management

Site selection and prioritization: Can seawater microbial signatures (e.g., oligotrophic vs. copiotrophic ratios, loads of potentially pathogenic microbes, virus-to-microbe abundance ratios) predict optimal sites for coral outplanting or larval recruitment, complementing traditional metrics (e.g., water clarity, substrate stability)? Similarly, can seawater microbes inform which sites are too degraded for successful restoration, thereby guiding conservation triage efforts (the strategic allocation of limited resources to the areas where they will have the greatest impact)?

Optimising sampling design for microbial indicators: Should microbial monitoring employ stratified sampling that combines standardised depth-integrated sampling (for ecosystem context) with targeted benthic-proximal sampling (to capture interface-specific signals critical for coral health diagnostics)?

Assessing restoration efficacy: How do seawater microbiomes differ in degraded vs. restored reefs, and can these differences guide restoration thresholds, or assess restoration success?

Forecasting ecosystem trajectories: Microbial communities have shown utility in monitoring ecological recovery following water quality improvements (Pesce et al. 2017). Building on this, a key future direction is to test if specific reef-associated seawater microbes and functional genes (e.g., nitrogen fixation, sulfur cycling) can act as prognosticator biomarkers to track reef outcomes (i.e. recovery or decline). The ability to forecast ecosystem trajectory post-disturbance (e.g., by detecting the return of crucial biogeochemical functions or beneficial microbial taxa) has a potential to provide a critical early-window into reef dynamics, complementing and potentially preceding the measurement of traditional, lagging metrics like coral cover and fish biomass.

Monitoring intervention impacts: The rapid response and scalability of microbial assays make them ideal for monitoring reef interventions, such as coral outplanting, macroalgae removal, or water quality remediation. Where regulations require impact assessments (such as on the Great Barrier Reef), microbial prognosticators could provide a critical data layer, especially when traditional metrics are difficult to measure at scale within limited timeframes.

Detecting specific anthropogenic stressors: Can tailored microbial indicators be developed to detect and attribute the impact of specific industrial pressures (e.g., from mining, shipping, or tourism) on reef health, providing a sensitive tool for regulation and compliance?

6 References

1. Glasl, B. *et al.* Comparative genome-centric analysis reveals seasonal variation in the function of coral reef microbiomes. *The ISME Journal* **14**, 1435–1450 (2020).
2. Lê Cao, K.-A. & Welham, Z. M. *Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package*. (Chapman and Hall/CRC, Boca Raton, 2021). doi:10.1201/9781003026860.
3. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, 128 (2017).
4. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* **13**, e1005752 (2017).
5. Moberg, F. & Folke, C. Ecological goods and services of coral reef ecosystems. *Ecological Economics* **29**, 215–233 (1999).
6. Costanza, R. *et al.* Changes in the global value of ecosystem services. *Global Environmental Change* **26**, 152–158 (2014).
7. De'ath, G., Fabricius, K. E., Sweatman, H. & Puotinen, M. The 27-year decline of coral cover on the Great Barrier Reef and its causes. *Proc Natl Acad Sci U S A* **109**, 17995–17999 (2012).
8. Hoegh-Guldberg, O., Poloczanska, E. S., Skirving, W. & Dove, S. Coral Reef Ecosystems under Climate Change and Ocean Acidification. *Front. Mar. Sci.* **4**, 158 (2017).
9. Eddy, T. D. *et al.* Global decline in capacity of coral reefs to provide ecosystem services. *One Earth* **4**, 1278–1285 (2021).
10. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103**, 12115–12120 (2006).
11. Abreu, A. *et al.* Priorities for ocean microbiome research. *Nat Microbiol* **7**, 937–947 (2022).
12. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biology* **5**, e77 (2007).
13. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology* **9**, e1001177 (2011).
14. Duarte, C. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin* **24**, (2015).
15. Bork, P. *et al.* Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**, 873 (2015).
16. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
17. Buttigieg, P. L. *et al.* Marine microbes in 4D-using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Curr Opin Microbiol* **43**, 169–185 (2018).
18. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* **18**, 428–445 (2020).
19. Wild, C. *et al.* Coral mucus functions as an energy carrier and particle trap in the reef ecosystem. *Nature* **428**, 66–70 (2004).
20. Furnas, M., Mitchell, A., Skuza, M. & Brodie, J. In the other 90%: phytoplankton responses to enhanced nutrient availability in the Great Barrier Reef Lagoon. *Mar Pollut Bull* **51**, 253–265 (2005).
21. Garren, M. & Azam, F. New directions in coral reef microbial ecology. *Environ Microbiol* **14**, 833–844 (2012).
22. Vanwonterghem, I. & Webster, N. S. Coral Reef Microorganisms in a Changing Climate. *iScience* **23**, 100972 (2020).
23. Bourne, D. G., Morrow, K. M. & Webster, N. S. Insights into the Coral Microbiome: Underpinning the Health and Resilience of Reef Ecosystems. *Annu. Rev. Microbiol.* **70**, 317–340 (2016).
24. Pita, L., Rix, L., Slaby, B. M., Franke, A. & Hentschel, U. The sponge holobiont in a changing ocean: from microbes to ecosystems. *Microbiome* **6**, 46 (2018).
25. Engelberts, J. P. *et al.* Characterization of a sponge microbiome using an integrative genome-centric approach. *ISME J* **14**, 1100–1110 (2020).
26. Marangon, E., Laffy, P. W., Bourne, D. G. & Webster, N. S. Microbiome-mediated mechanisms contributing to the environmental tolerance of reef invertebrate species. *Mar Biol* **168**, 89 (2021).
27. Robbins, S. J. *et al.* A genomic view of the microbiome of coral reef demosponges. *The ISME Journal* **15**, 1641–1654 (2021).
28. Moeller, F. U. *et al.* Taurine as a key intermediate for host-symbiont interaction in the tropical sponge Ianthella basta. *ISME J* **17**, 1208–1223 (2023).

29. Bourne, D., Iida, Y., Uthicke, S. & Smith-Keune, C. Changes in coral-associated microbial communities during a bleaching event. *The ISME Journal* **2**, 350–363 (2008).

30. Littman, R. A., Willis, B. L. & Bourne, D. G. Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. *Environmental microbiology reports* **3 6**, 651–60 (2011).

31. McDevitt-Irwin, J. M., Baum, J. K., Garren, M. & Vega Thurber, R. L. Responses of Coral-Associated Bacterial Communities to Local and Global Stressors. *Front. Mar. Sci.* **4**, 262 (2017).

32. Fan, L., Liu, M., Simister, R., Webster, N. S. & Thomas, T. Marine microbial symbiosis heats up: the phylogenetic and functional response of a sponge holobiont to thermal stress. *The ISME Journal* **7**, 991–1002 (2013).

33. van Oppen, M. J. H. & Blackall, L. L. Coral microbiome dynamics, functions and design in a changing world. *Nature Reviews Microbiology* **17**, 557–567 (2019).

34. Glasl, B., Webster, N. S. & Bourne, D. G. Microbial indicators as a diagnostic tool for assessing water quality and climate stress in coral reef ecosystems. *Marine Biology* **164**, 91 (2017).

35. Roitman, S., Joseph Pollock, F. & Medina, M. Coral Microbiomes as Bioindicators of Reef Health. in *Population Genomics: Marine Organisms* (eds Oleksiak, M. F. & Rajora, O. P.) 39–57 (Springer International Publishing, Cham, 2018). doi:10.1007/13836_2018_29.

36. Dinsdale, E. A. *et al.* Microbial Ecology of Four Coral Atolls in the Northern Line Islands. *PLOS ONE* **3**, e1584 (2008).

37. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology* **25**, 56–66 (2015).

38. Peixoto, R. S., Rosado, P. M., Leite, D. C. de A., Rosado, A. S. & Bourne, D. G. Beneficial Microorganisms for Corals (BMC): Proposed Mechanisms for Coral Health and Resilience. *Frontiers in Microbiology* **8**, (2017).

39. Glasl, B., Bourne, D., Frade, P. & Webster, N. Establishing microbial baselines to identify indicators of coral reef health. *Microbiology Australia* **39**, (2018).

40. Glasl, B. *et al.* Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* **7**, 94 (2019).

41. Semenza, J. C. *et al.* Environmental Suitability of Vibrio Infections in a Warming Climate: An Early Warning System. *Environ Health Perspect* **125**, 107004 (2017).

42. Leite, D. C. A. *et al.* Coral Bacterial-Core Abundance and Network Complexity as Proxies for Anthropogenic Pollution. *Front Microbiol* **9**, 833 (2018).

43. Frade, P. R. *et al.* Spatial patterns of microbial communities across surface waters of the Great Barrier Reef. *Communications Biology* **3**, 442 (2020).

44. Hughes, T. P. *et al.* Climate Change, Human Impacts, and the Resilience of Coral Reefs. *Science* **301**, 929–933 (2003).

45. Glasl, B., Smith, C. E., Bourne, D. G. & Webster, N. S. Disentangling the effect of host-genotype and environment on the microbiome of the coral Acropora tenuis. *PeerJ* **7**, e6377 (2019).

46. Glasl, B., Smith, C. E., Bourne, D. G. & Webster, N. S. Exploring the diversity-stability paradigm using sponge microbial communities. *Scientific Reports* **8**, 8425 (2018).

47. Simister, R. *et al.* Thermal stress responses in the bacterial biosphere of the Great Barrier Reef sponge, Rhopaloeides odorabile. *Environ Microbiol* **14**, 3232–3246 (2012).

48. Luter, H. M., Gibb, K. & Webster, N. S. Eutrophication has no short-term effect on the Cymbastela stipitata holobiont. *Frontiers in Microbiology* **5**, (2014).

49. Pineda, M.-C. *et al.* Effects of suspended sediments on the sponge holobiont with implications for dredging management. *Scientific Reports* **7**, 4925 (2017).

50. Galand, P. E. *et al.* Diversity of the Pacific Ocean coral reef microbiome. *Nat Commun* **14**, 3039 (2023).

51. Kelly, L. W. *et al.* Diel population and functional synchrony of microbial communities on coral reefs. *Nature Communications* **10**, (2019).

52. Weber, L. & Apprill, A. Diel, daily, and spatial variation of coral reef seawater microbial communities. *PLOS ONE* **15**, e0229442 (2020).

53. Haas, A. F. *et al.* Global microbialization of coral reefs. *Nat Microbiol* **1**, 16042 (2016).

54. Raj, K. D. *et al.* Low oxygen levels caused by Noctiluca scintillans bloom kills corals in Gulf of Mannar, India. *Sci Rep* **10**, 22133 (2020).

55. Bush, T. *et al.* Oxic-anoxic regime shifts mediated by feedbacks between biogeochemical processes and microbial community dynamics. *Nature Communications* **8**, 789 (2017).

56. Johnson, M. D. *et al.* Rapid ecosystem-scale consequences of acute deoxygenation on a Caribbean coral reef. *Nature Communications* **12**, 4522 (2021).

57. Barott, K. L. & Rohwer, F. L. Unseen players shape benthic competition on coral reefs. *Trends in Microbiology* **20**, 621–628 (2012).

58. Haas, A. F. *et al.* Influence of coral and algal exudates on microbially mediated reef metabolism. *PeerJ* **1**, e108 (2013).

59. Weber, L. *et al.* Microbial signatures of protected and impacted Northern Caribbean reefs: changes from Cuba to the Florida Keys. *Environ Microbiol* **22**, 499–519 (2020).

60. Wambua, S. *et al.* Cross-Sectional Variations in Structure and Function of Coral Reef Microbiome With Local Anthropogenic Impacts on the Kenyan Coast of the Indian Ocean. *Frontiers in Microbiology* **12**, (2021).

61. Angly, F. E. *et al.* Marine microbial communities of the Great Barrier Reef lagoon are influenced by riverine floodwaters and seasonal weather events. *PeerJ* **4**, e1511 (2016).

62. Gorsky, G. *et al.* Expanding Tara Oceans Protocols for Underway, Ecosystemic Sampling of the Ocean-Atmosphere Interface During Tara Pacific Expedition (2016–2018). *Frontiers in Marine Science* **6**, (2019).

63. Planes, S. *et al.* The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLOS Biology* **17**, e3000483 (2019).

64. Lombard, F. *et al.* Open science resources from the Tara Pacific expedition across coral reef and surface ocean ecosystems. *Sci Data* **10**, 324 (2023).

65. Belser, C. *et al.* Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition. Preprint at https://doi.org/10.48550/arXiv.2207.02475 (2022).

66. Salazar, G. & Sunagawa, S. Marine microbial diversity. *Current Biology* **27**, R489–R494 (2017).

67. Lombard, F. *et al.* Globally Consistent Quantitative Observations of Planktonic Ecosystems. *Front. Mar. Sci.* **6**, 196 (2019).

68. Cordier, T. *et al.* Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology* **30**, 2937–2958 (2021).

69. Beale, D. J. *et al.* Omics-based ecosurveillance for the assessment of ecosystem function, health, and resilience. *Emerg Top Life Sci* **6**, 185–199 (2022).

70. Djemiel, C. *et al.* Potential of Meta-Omics to Provide Modern Microbial Indicators for Monitoring Soil Quality and Securing Food Production. *Frontiers in Microbiology* **13**, (2022).

71. Parkinson, J. E. *et al.* Molecular tools for coral reef restoration: Beyond biomarker discovery. *Conservation Letters* **13**, e12687 (2020).

72. Belser, C. *et al.* Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition. *Sci Data* **10**, 326 (2023).

73. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).

74. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623-1635.e11 (2019).

75. Faure, E., Ayata, S.-D. & Bittner, L. Towards omics-based predictions of planktonic functional composition from environmental data. *Nat Commun* **12**, 4361 (2021).

76. Frémont, P. *et al.* Restructuring of plankton genomic biogeography in the surface ocean under climate change. *Nature Climate Change* **12**, 1–9 (2022).

77. Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun Biol* **3**, 1–10 (2020).

78. Chaffron, S. *et al.* Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv* **7**, eabg1921 (2021).

79. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).

80. Tagliabue, A. 'Oceans are hugely complex': modelling marine microbes is key to climate forecasts. *Nature* **623**, 250–252 (2023).

81. Planes, S. & Allemand, D. Insights and achievements from the Tara Pacific expedition. *Nat Commun* **14**, 3131 (2023).

82. Nelson, C. E., Wegley Kelly, L. & Haas, A. F. Microbial Interactions with Dissolved Organic Matter Are Central to Coral Reef Ecosystem Function and Resilience. *Annu. Rev. Mar. Sci.* **15**, 431–460 (2023).

83. Webster, N. S., Wagner, M. & Negri, A. P. Microbial conservation in the Anthropocene. *Environ Microbiol* **20**, 1925–1928 (2018).

84. Correa-Garcia, S., Constant, P. & Yergeau, E. The forecasting power of the microbiome. *Trends in Microbiology* https://doi.org/10.1016/j.tim.2022.11.013 (2022) doi:10.1016/j.tim.2022.11.013.

85. Doyle, S. M. *et al.* Microbial Community Dynamics Provide Evidence for Hypoxia during a Coral Reef Mortality Event. *Applied and Environmental Microbiology* **88**, e00347-22 (2022).

86. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nat Ecol Evol* **2**, 936–943 (2018).

87. Alker, A. P., Smith, G. W. & Kim, K. Characterization of Aspergillus sydowii (Thom et Church), a fungal pathogen of Caribbean sea fan corals. *Hydrobiologia* **460**, 105–111 (2001).

88. Bourne, D. G. *et al.* Microbial disease and the coral holobiont. *Trends Microbiol* **17**, 554–562 (2009).

89. Sutherland, K. P., Shaban, S., Joyner, J. L., Porter, J. W. & Lipp, E. K. Human Pathogen Shown to Cause Disease in the Threatened Eklhorn Coral Acropora palmata. *PLoS One* **6**, e23468 (2011).

90. Bourne, D. G., Smith, H. A. & Page, C. A. Diseases of scleractinian corals. in *Invertebrate Pathology* (eds Rowley, A. F., Coates, C. J. & Whitten, M. W.) 0 (Oxford University Press, 2022). doi:10.1093/oso/9780198853756.003.0004.

91. Munn, C. B. The Role of Vibrios in Diseases of Corals. *Microbiol Spectr* **3**, (2015).

92. Amin, A. K. M. R. *et al.* The First Temporal and Spatial Assessment of Vibrio Diversity of the Surrounding Seawater of Coral Reefs in Ishigaki, Japan. *Frontiers in Microbiology* **7**, (2016).

93. Nelson, C. E. *et al.* Coral and macroalgal exudates vary in neutral sugar composition and differentially enrich reef bacterioplankton lineages. *ISME J* **7**, 962–979 (2013).

94. Cárdenas, A. *et al.* Excess labile carbon promotes the expression of virulence factors in coral reef bacterioplankton. *ISME J* **12**, 59–76 (2018).

95. Smith, J. E. *et al.* Indirect effects of algae on coral: algae-mediated, microbe-induced coral mortality. *Ecol Lett* **9**, 835–845 (2006).

96. Brown, A. & Carpenter, R. Water-flow mediated oxygen dynamics within massive Porites-algal turf interactions. *Mar Ecol Prog Ser* **490**, 1–10 (2013).

97. Jorissen, H., Skinner, C., Osinga, R., de Beer, D. & Nugues, M. M. Evidence for water-mediated mechanisms in coral-algal interactions. *Proc Biol Sci* **283**, (2016).

98. Walsh, K. *et al.* Aura-biomes are present in the water layer above coral reef benthic macro-organisms. *PeerJ* **5**, e3666 (2017).

99. Harvell, C. D. *et al.* Climate warming and disease risks for terrestrial and marine biota. *Science* **296**, 2158–2162 (2002).

100. Harvell, C. *et al.* Coral Disease, Environmental Drivers, and the Balance Between Coral and Microbial Associates. *Oceanography (Washington D.C.)* **20**, 58–81 (2007).

101. Altizer, S., Ostfeld, R. S., Johnson, P. T. J., Kutz, S. & Harvell, C. D. Climate Change and Infectious Diseases: From Evidence to a Predictive Framework. *Science* **341**, 514–519 (2013).

102. Maynard, J. *et al.* Improving marine disease surveillance through sea temperature monitoring, outlooks and projections. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150208 (2016).

103. Cavicchioli, R. *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**, 569–586 (2019).

104. Semenza, J. C. *et al.* Climate Change Impact Assessment of Food- and Waterborne Diseases. *Crit Rev Environ Sci Technol* **42**, 857–890 (2012).

105. Tout, J. *et al.* Increased seawater temperature increases the abundance and alters the structure of natural Vibrio populations associated with the coral Pocillopora damicornis. *Frontiers in Microbiology* **6**, (2015).

106. Doni, L. *et al.* Large-scale impact of the 2016 Marine Heatwave on the plankton-associated microbial communities of the Great Barrier Reef (Australia). *Mar Pollut Bull* **188**, 114685 (2023).

107. Selig, E. R. *et al.* Analyzing the relationship between ocean temperature anomalies and coral disease outbreaks at broad spatial scales. in *Coastal and Estuarine Studies* (eds Phinney, J. T., Hoegh-Guldberg, O., Kleypas, J., Skirving, W. & Strong, A.) vol. 61 111–128 (American Geophysical Union, Washington, D. C., 2006).

108. Brandt, M. E. & McManus, J. W. Disease incidence is related to bleaching extent in reef-building corals. *Ecology* **90**, 2859–2867 (2009).

109. Miller, J. *et al.* Coral disease following massive bleaching in 2005 causes 60% decline in coral cover on reefs in the US Virgin Islands. *Coral Reefs* **28**, 925–937 (2009).

110. Randall, C. J. & van Woesik, R. Contemporary white-band disease in Caribbean corals driven by climate change. *Nature Clim Change* **5**, 375–379 (2015).

111. Precht, W. F., Gintert, B. E., Robbart, M. L., Fura, R. & van Woesik, R. Unprecedented Disease-Related Coral Mortality in Southeastern Florida. *Sci Rep* **6**, 31374 (2016).

112. Brodnicke, O. B. *et al.* Unravelling the links between heat stress, bleaching and disease: fate of tabular corals following a combined disease and bleaching event. *Coral Reefs* **38**, 591–603 (2019).

113. Jackson, J. B. C. *et al.* Historical Overfishing and the Recent Collapse of Coastal Ecosystems. *Science* **293**, 629–637 (2001).

114. McDole, T. *et al.* Assessing Coral Reefs on a Pacific-Wide Scale Using the Microbialization Score. *PLOS ONE* **7**, e43233 (2012).

115. Silveira, C. B. *et al.* Biophysical and physiological processes causing oxygen loss from coral reefs. *eLife* **8**, e49114 (2019).

116. Webster, N. & Gorsuch, H. *Monitoring Additional Values within the Reef 2050 Integrated Monitoring and Reporting Program: Final Report of the Microbes Expert Group*. https://elibrary.gbrmpa.gov.au/jspui/handle/11017/3594 (2020).

117. Bushon, R. n., Brady, A. m., Likirdopulos, C. a. & Cireddu, J. v. Rapid detection of Escherichia coli and enterococci in recreational water using an immunomagnetic separation/adenosine triphosphate technique. *Journal of Applied Microbiology* **106**, 432–441 (2009).

118. Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. Pepper Mild Mottle Virus as an Indicator of Fecal Pollution. *Appl Environ Microbiol* **75**, 7261–7267 (2009).

119. Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* **9**, 2490–2502 (2015).

120. Laroche, O. *et al.* A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin* **127**, 97–107 (2018).

121. Walsh, D. A. *et al.* Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**, 578–582 (2009).

122. Kimes, N. E. *et al.* Temperature regulation of virulence factors in the pathogen Vibrio coralliilyticus. *The ISME Journal* **6**, 835–846 (2012).

123. Hoegh-Guldberg, O., Poloczanska, E. S., Skirving, W. & Dove, S. Coral Reef Ecosystems under Climate Change and Ocean Acidification. *Frontiers in Marine Science* **4**, (2017).

124. Hughes, T. P. *et al.* Coral reefs in the Anthropocene. *Nature* **546**, 82–90 (2017).

125. Souter, D. *et al.* Status of Coral Reefs of the World: 2020. (2020).

126. Cooper, T. F., Gilmour, J. P. & Fabricius, K. E. Bioindicators of changes in water quality on coral reefs: review and recommendations for monitoring programmes. *Coral Reefs* **28**, 589–606 (2009).

127. Fabricius, K. E. *et al.* A bioindicator system for water quality on inshore coral reefs of the Great Barrier Reef. *Marine Pollution Bulletin* **65**, 320–332 (2012).

128. Apprill, A. *et al.* Toward a New Era of Coral Reef Monitoring. *Environmental Science & Technology* **57**, 5117 (2023).

129. Becker, C. C., Weber, L., Llopiz, J. K., Mooney, T. A. & Apprill, A. Microorganisms uniquely capture and predict stony coral tissue loss disease and hurricane disturbance impacts on US Virgin Island reefs. *Environ Microbiol* **26**, e16610 (2024).

130. Terzin, M. *et al.* The road forward to incorporate seawater microbes in predictive reef monitoring. *Environmental Microbiome* **19**, 5 (2024).

131. Chen, J., McIlroy, S. E., Archana, A., Baker, D. M. & Panagiotou, G. A pollution gradient contributes to the taxonomic, functional, and resistome diversity of microbial communities in marine sediments. *Microbiome* **7**, 104 (2019).

132. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

133. Song, W. *et al.* Functional Traits Resolve Mechanisms Governing the Assembly and Distribution of Nitrogen-Cycling Microbial Communities in the Global Ocean. *mBio* **13**, e03832-21 (2022).

134. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

135. Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences* **105**, 11512–11519 (2008).

136. Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences* **108**, 14288–14293 (2011).

137. Banerjee, S. *et al.* Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* **97**, 188–198 (2016).

138. Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nat Ecol Evol* **1**, 1–12 (2016).

139. Jurburg, S. D., Salles, J. F., Jurburg, S. D. & Salles, J. F. Functional Redundancy and Ecosystem Function — The Soil Microbiota as a Case Study. in *Biodiversity in Ecosystems - Linking Structure and Function* (IntechOpen, 2015). doi:10.5772/58981.

140. Moya, A. & Ferrer, M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in Microbiology* **24**, 402–413 (2016).

141. Fassarella, M. *et al.* Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health. *Gut* **70**, 595–605 (2021).

142. Haggerty, J. M. & Dinsdale, E. A. Distinct biogeographical patterns of marine bacterial taxonomy and functional genes. *Global Ecology and Biogeography* **26**, 177–190 (2017).

143. Becker, C. C. *et al.* Microorganisms and dissolved metabolites distinguish Florida's Coral Reef habitats. *PNAS Nexus* **2**, pgad287 (2023).

144. Lê Cao, K.-A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* **7**, Article 35 (2008).

145. Lê Cao, K.-A., Martin, P. G., Robert-Granié, C. & Besse, P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10**, 34 (2009).

146. Jameson, B. D. *et al.* Network analysis of 16S rRNA sequences suggests microbial keystone taxa contribute to marine N2O cycling. *Commun Biol* **6**, 1–14 (2023).

147. Priest, T. *et al.* Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities. *ISME J* **17**, 1612–1625 (2023).

148. R Core Team. RStudio Desktop. RStudio (2023).

149. Hijmans, R. J. *et al.* raster: Geographic Data Analysis and Modeling. (2023).

150. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

151. Dunnington, D., Thorne, B. & Hernangómez, D. ggspatial: Spatial Data Framework for ggplot2. (2023).

152. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439 (2018).

153. Pebesma, E. & Bivand, R. *Spatial Data Science: With Applications in R*. (Chapman and Hall/CRC, New York, 2023). doi:10.1201/9780429459016.

154. Barneche, D. R. *et al.* dataaimsr: An R Client for the Australian Institute of Marine Science Data Platform API which provides easy access to AIMS Data Platform. *Journal of Open Source Software* **6**, 3282 (2021).

155. Slowikowski, K. *et al.* ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2024).

156. Great Barrier Reef Marine Park Authority. *Marine Monitoring Program Annual Report Quality Assurance and Quality Control Manual 2020-21*. https://elibrary.gbrmpa.gov.au/jspui/handle/11017/3932 (2022).

157. IMOS. Underway sensors: Enhanced measurements from Ships of Opportunity (SOOP): RV Cape Ferguson | AIMS Data Repository | aims.gov.au. https://apps.aims.gov.au/metadata/view/da560e78-1a4e-43dc-aa4b-c99c3c4ab700 (2015).

158. Ryle, V. D., Mueller, H. R. & Gentien, P. *Automated Analysis of Nutrients in Tropical Sea Waters*. 24 (1981).

159. Parsons, T. R., Maita, Y. & Lalli, C. M. *A Manual of Chemical and Biological Methods for Seawater Analysis*. (Pergamon Press, 1984). doi:10.25607/OBP-1830.

160. Bran & Luebbe. *Directory of Autoanalyser Methods*. (1997).

161. Valderrama, J. C. The simultaneous analysis of total nitrogen and total phosphorus in natural waters. *Marine Chemistry* **10**, 109–122 (1981).

162. Menzel, D. W. & Corwin, N. The Measurement of Total Phosphorus in Seawater Based on the Liberation of Organically Bound Fractions by Persulfate OXIDATION1. *Limnology and Oceanography* **10**, 280–282 (1965).

163. Strickland, J. D. H. & Parsons, T. R. A Practical Handbook of Seawater Analysis, 2nd edition. https://doi.org/10.25607/OBP-1791 (1972) doi:10.25607/OBP-1791.

164. Botté, E. S. *et al.* Changes in the metabolic potential of the sponge microbiome under ocean acidification. *Nat Commun* **10**, 4134 (2019).

165. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).

166. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217 (2013).

167. Lahti, L. & Shetty, S. microbiome R package. https://doi.org/10.18129/B9.bioc.microbiome (2017) doi:10.18129/B9.bioc.microbiome.

168. Martinez, A. pairwiseAdonis: Pairwise multilevel comparison using adonis. R package version 0.4. *R package version 0.4* (2020).

169. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).

170. Wickham, H. *Ggplot2*. (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-24277-4.

171. Oksanen, J. *et al.* The vegan package. *Community ecology package* **10**, 719 (2007).

172. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473 (2010).

173. Lê Cao, K.-A. L. & Welham, Z. M. *Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package*. (Chapman and Hall/CRC, New York, 2021). doi:10.1201/9781003026860.

174. Loreau, M. Does functional redundancy exist? https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/j.0030-1299.2004.12685.x (2004).

175. Biggs, C. *et al.* Does functional redundancy affect ecological stability and resilience? A review and meta-analysis. *Ecosphere* **11**, e03184 (2020).

176. Shade, A. Microbiome rescue: directing resilience of environmental microbial communities. *Current Opinion in Microbiology* **72**, 102263 (2023).

177. Flensborg, L. C., Maureaud, A. A., Bravo, D. N. & Lindegren, M. An indicator-based approach for assessing marine ecosystem resilience. *ICES Journal of Marine Science* **80**, 1487–1499 (2023).

178. Chambers, J. C., Allen, C. R. & Cushman, S. A. Operationalizing Ecological Resilience Concepts for Managing Species and Ecosystems at Risk. *Front. Ecol. Evol.* **7**, (2019).

179. Brodie, J., De'ath, G., Devlin, M., Furnas, M. & Wright, M. Spatial and temporal patterns of near-surface chlorophyll a in the Great Barrier Reef lagoon. *Marine and Freshwater Research* **58**, 342–353 (2007).

180. De'ath, G. & Fabricius, K. E. *Water Quality of the Great Barrier Reef: Distributions, Effects on Reef Biota and Trigger Values for the Protection of Ecosystem Health*. https://elibrary.gbrmpa.gov.au/jspui/handle/11017/416 (2008).

181. Furnas, M. J. & Mitchell, A. W. Phytoplankton dynamics in the central Great Barrier Reef—I. Seasonal changes in biomass and community structure and their relation to intrusive activity. *Continental Shelf Research* **6**, 363–384 (1986).

182. Benthuysen, J., Tonin, H., Brinkman, R., Herzfeld, M. & Steinberg, C. Intrusive upwelling in the Central Great Barrier Reef. *Journal of Geophysical Research: Oceans* **121**, (2016).

183. Charpy, L., Casareto, B. E., Langlade, M. J. & Suzuki, Y. Cyanobacteria in Coral Reef Ecosystems: A Review. *Journal of Marine Sciences* **2012**, e259571 (2012).

184. Bahadori, M. *et al.* The origin of suspended particulate matter in the Great Barrier Reef. *Nat Commun* **14**, 5629 (2023).

185. McNally, S. P., Parsons, R. J., Santoro, A. E. & Apprill, A. Multifaceted impacts of the stony coral Porites astreoides on picoplankton abundance and community composition. *Limnology and Oceanography* **62**, 217–234 (2017).

186. Meunier, V. *et al.* Bleaching forces coral's heterotrophy on diazotrophs and Synechococcus. *The ISME Journal* **13**, 2882–2886 (2019).

187. Crosbie, N. & Furnas, M. Abundance, distribution and flow-cytometric characterization of picophytoprokaryote populations in central (17degreesS) and southern (20degreesS) shelf waters of the Great Barrier Reef. *Journal of Plankton Research - J PLANKTON RES* **23**, 809–828 (2001).

188. Fabricius, K. E. Effects of terrestrial runoff on the ecology of corals and coral reefs: review and synthesis. *Mar Pollut Bull* **50**, 125–146 (2005).

189. Dave, U. C. & Kadeppagari, R.-K. Alanine dehydrogenase and its applications – A review. *Critical Reviews in Biotechnology* **39**, 648–664 (2019).

190. Hudek, L., Premachandra, D., Webster, W. A. J. & Bräu, L. Role of Phosphate Transport System Component PstB1 in Phosphate Internalization by Nostoc punctiforme. *Appl Environ Microbiol* **82**, 6344–6356 (2016).

191. Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H. & Amann, R. High Rate of Uptake of Organic Nitrogen Compounds by Prochlorococcus Cyanobacteria as a Key to Their Dominance in Oligotrophic Oceanic Waters. *Appl Environ Microbiol* **69**, 1299–1304 (2003).

192. Martiny, A., Coleman, M. & Chisholm, S. Phosphate acquisition genes in Prochlorococcus ecotypes. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12552–7 (2006).

193. Bouman, H. A. *et al.* Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. *Science* **312**, 918–921 (2006).

194. Rocap, G. *et al.* Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).

195. Sohm, J. A. *et al.* Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* **10**, 333–345 (2016).

196. Fogg, G. E. The Ecological Significance of Extracellular Products of Phytoplankton Photosynthesis. https://www.degruyter.com/document/doi/10.1515/botm.1983.26.1.3/html?lang=en (1983).

197. Moran, M. A. *et al.* The Ocean's labile DOC supply chain. *Limnology and Oceanography* **67**, 1007–1021 (2022).

198. Moran, M. A. *et al.* Microbial metabolites in the marine carbon cycle. *Nat Microbiol* **7**, 508–523 (2022).

199. He, W., Chen, M., Schlautman, M. A. & Hur, J. Dynamic exchanges between DOM and POM pools in coastal and inland aquatic ecosystems: A review. *Science of The Total Environment* **551–552**, 415–428 (2016).

200. Vardi, A. *et al.* Host–virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proceedings of the National Academy of Sciences* **109**, 19327–19332 (2012).

201. Bidle, K. D. The Molecular Ecophysiology of Programmed Cell Death in Marine Phytoplankton. *Annual Review of Marine Science* **7**, 341–375 (2015).

202. Steinberg, D. K. & Landry, M. R. Zooplankton and the Ocean Carbon Cycle. *Annual Review of Marine Science* **9**, 413–444 (2017).

203. De Corte, D. *et al.* Zooplankton-derived dissolved organic matter composition and its bioavailability to natural prokaryotic communities. *Limnology and Oceanography* **68**, 336–347 (2023).

204. Carlson, C. A. & Hansell, D. A. Chapter 3 - DOM Sources, Sinks, Reactivity, and Budgets. in *Biogeochemistry of Marine Dissolved Organic Matter (Second Edition)* (eds Hansell, D. A. & Carlson, C. A.) 65–126 (Academic Press, Boston, 2015). doi:10.1016/B978-0-12-405940-5.00003-0.

205. Enke, T. N. *et al.* Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. *Current Biology* **29**, 1528-1535.e6 (2019).

206. Mentges, A., Feenders, C., Deutsch, C., Blasius, B. & Dittmar, T. Long-term stability of marine dissolved organic carbon emerges from a neutral network of compounds and microbes. *Sci Rep* **9**, 17780 (2019).

207. Azam, F. *et al.* The Ecological Role of Water-Column Microbes in the Sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).

208. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nat Rev Microbiol* **5**, 782–791 (2007).

209. Kirchman, D. L. Carbon Pumps in the Oceans. in *Microbes: The Unseen Agents of Climate Change* (ed. Kirchman, D. L.) 0 (Oxford University Press, 2024). doi:10.1093/oso/9780197688564.003.0004.

210. Cardini, U., Bednarz, V., Foster, R. & Wild, C. Benthic N2 fixation in coral reefs and the potential effects of human-induced environmental change. *Ecology and Evolution* **4**, (2014).

211. Pujalte, M. J., Lucena, T., Ruvira, M. A., Arahal, D. R. & Macián, M. C. The Family Rhodobacteraceae. in *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria* (eds Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 439–512 (Springer, Berlin, Heidelberg, 2014). doi:10.1007/978-3-642-30197-1_377.

212. Baldani, J. I. *et al.* The Family Rhodospirillaceae. in *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria* (eds Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 533–618 (Springer, Berlin, Heidelberg, 2014). doi:10.1007/978-3-642-30197-1_300.

213. Gavriilidou, A. *et al.* Comparative genomic analysis of Flavobacteriaceae: insights into carbohydrate metabolism, gliding motility and secondary metabolite biosynthesis. *BMC Genomics* **21**, 569 (2020).

214. Je, W. The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains. *Quarterly reviews of biophysics* **25**, (1992).

215. Ohnishi, T. Iron–sulfur clusters/semiquinones in Complex I. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1364**, 186–206 (1998).

216. Yagi, T. & Matsuno-Yagi, A. The proton-translocating NADH-quinone oxidoreductase in the respiratory chain: the secret unlocked. *Biochemistry* **42**, 2266–2274 (2003).

217. Hederstedt, L. Diversity of Cytochrome c Oxidase Assembly Proteins in Bacteria. *Microorganisms* **10**, 926 (2022).

218. Rinta-Kanto, J. M., Sun, S., Sharma, S., Kiene, R. P. & Moran, M. A. Bacterial community transcription patterns during a marine phytoplankton bloom. *Environmental Microbiology* **14**, 228–239 (2012).

219. Teeling, H. *et al.* Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–611 (2012).

220. Gregor, R. *et al.* Widespread B-vitamin auxotrophy in marine particle-associated bacteria. 2023.10.16.562604 Preprint at https://doi.org/10.1101/2023.10.16.562604 (2023).

221. Sivaraman, J. *et al.* Crystal structure of Escherichia coli PdxA, an enzyme involved in the pyridoxal phosphate biosynthesis pathway. *J Biol Chem* **278**, 43682–43690 (2003).

222. John, R. A. Pyridoxal phosphate-dependent enzymes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1248**, 81–96 (1995).

223. Hayashi, H. Pyridoxal Enzymes: Mechanistic Diversity and Uniformity. *The Journal of Biochemistry* **118**, 463–473 (1995).

224. Eliot, A. C. & Kirsch, J. F. Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu Rev Biochem* **73**, 383–415 (2004).

225. Jörnvall, H. *et al.* Short-chain dehydrogenases/reductases (SDR). *Biochemistry* **34**, 6003–6013 (1995).

226. Oppermann, U. *et al.* Short-chain dehydrogenases/reductases (SDR): the 2002 update. *Chemico-Biological Interactions* **143–144**, 247–253 (2003).

227. Herrmann, K. M. & Weaver, L. M. THE SHIKIMATE PATHWAY. *Annu Rev Plant Physiol Plant Mol Biol* **50**, 473–503 (1999).

228. Ramakrishnan, V. Ribosome Structure and the Mechanism of Translation. *Cell* **108**, 557–572 (2002).

229. Vollmer, W., Blanot, D. & De Pedro, M. A. Peptidoglycan structure and architecture. *FEMS Microbiology Reviews* **32**, 149–167 (2008).

230. Silhavy, T. J., Kahne, D. & Walker, S. The Bacterial Cell Envelope. *Cold Spring Harb Perspect Biol* **2**, a000414 (2010).

231. Park, J. T. Identification of a Dedicated Recycling Pathway for Anhydro-N-Acetylmuramic Acid andN-Acetylglucosamine Derived from Escherichia coli Cell Wall Murein. *Journal of Bacteriology* **183**, 3842–3847 (2001).

232. Riemann, L. & Azam, F. Widespread N-Acetyl-d-Glucosamine Uptake among Pelagic Marine Bacteria and Its Ecological Implications. https://journals.asm.org/doi/10.1128/aem.68.11.5554-5562.2002 (2002).

233. Uehara, T. *et al.* Recycling of the Anhydro-N-Acetylmuramic Acid Derived from Cell Wall Murein Involves a Two-Step Conversion to N-Acetylglucosamine-Phosphate. *Journal of Bacteriology* **187**, 3643–3649 (2005).

234. Uehara, T., Suefuji, K., Jaeger, T., Mayer, C. & Park, J. T. MurQ Etherase Is Required by Escherichia coli in Order To Metabolize Anhydro-N-Acetylmuramic Acid Obtained either from the Environment or from Its Own Cell Wall. *Journal of Bacteriology* **188**, 1660–1662 (2006).

235. Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun Biol* **3**, 1–10 (2020).

236. Faure, E., Ayata, S.-D. & Bittner, L. Towards omics-based predictions of planktonic functional composition from environmental data. *Nat Commun* **12**, 4361 (2021).

237. Viegas, C. A. Chapter Four - Microbial bioassays in environmental toxicity testing. in *Advances in Applied Microbiology* (eds Gadd, G. M. & Sariaslani, S.) vol. 115 115–158 (Academic Press, 2021).

238. Durrieu, C. & Tran-Minh, C. Optical algal biosensor using alkaline phosphatase for determination of heavy metals. *Ecotoxicol Environ Saf* **51**, 206–209 (2002).

239. Knapik, K., Bagi, A., Krolicka, A. & Baussant, T. Metatranscriptomic Analysis of Oil-Exposed Seawater Bacterial Communities Archived by an Environmental Sample Processor (ESP). *Microorganisms* **8**, 744 (2020).

240. Buttigieg, P. L. *et al.* Marine microbes in 4D-using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Curr Opin Microbiol* **43**, 169–185 (2018).

241. Martin-Platero, A. M. *et al.* High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat Commun* **9**, 266 (2018).

242. Laber, C. *et al.* Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology* **3**, (2018).

243. Mora, C. *et al.* Coral Reefs and the Global Network of Marine Protected Areas. *Science* **312**, 1750–1751 (2006).

244. Graham, N. A. J. & McClanahan, T. R. The Last Call for Marine Wilderness? *BioScience* **63**, 397–402 (2013).

245. Edgar, G. J. *et al.* Global conservation outcomes depend on marine protected areas with five key features. *Nature* **506**, 216–220 (2014).

246. Sala, E. & Giakoumi, S. No-take marine reserves are the most effective protected areas in the ocean. *ICES Journal of Marine Science* **75**, 1166–1168 (2018).

247. Caldwell, I. R. *et al.* Protection efforts have resulted in ~10% of existing fish biomass on coral reefs. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2308605121 (2024).

248. Great Barrier Reef Marine Park Authority. *Great Barrier Reef Marine Park Zoning Plan 2003*. (Great Barrier Reef Marine Park Authority, 2004).

249. Emslie, M. J. *et al.* Decades of monitoring have informed the stewardship and ecological understanding of Australia's Great Barrier Reef. *Biological Conservation* **252**, 108854 (2020).

250. Fernandes, L. *et al.* Establishing Representative No-Take Areas in the Great Barrier Reef: Large-Scale Implementation of Theory on Marine Protected Areas. *Conservation Biology* **19**, 1733–1744 (2005).

251. Babcock, R. C. *et al.* Decadal trends in marine reserves reveal differential rates of change in direct and indirect effects. *Proc Natl Acad Sci U S A* **107**, 18256–18261 (2010).

252. Russ, G. R. *et al.* Rapid increase in fish numbers follows creation of world's largest marine reserve network. *Current Biology* **18**, R514–R515 (2008).

253. Emslie, M. J. *et al.* Expectations and Outcomes of Reserve Network Performance following Re-zoning of the Great Barrier Reef Marine Park. *Current Biology* **25**, 983–992 (2015).

254. Bode, M. *et al.* Marine reserves contribute half of the larval supply to a coral reef fishery. *Science Advances* **11**, eadt0216 (2025).

255. Mellin, C., Aaron MacNeil, M., Cheal, A. J., Emslie, M. J. & Julian Caley, M. Marine protected areas increase resilience among coral reef communities. *Ecology Letters* **19**, 629–637 (2016).

256. Mellin, C., Bradshaw, C. J. A., Fordham, D. A. & Caley, M. J. Strong but opposing β-diversity–stability relationships in coral reef fish communities. *Proc Biol Sci* **281**, 20131993 (2014).

257. Kroon, F. J., Barneche, D. R. & Emslie, M. J. Fish predators control outbreaks of Crown-of-Thorns Starfish. *Nat Commun* **12**, 6986 (2021).

258. Lamb, J. B., Williamson, D. H., Russ, G. R. & Willis, B. L. Protected areas mitigate diseases of reef-building corals by reducing damage from fishing. *Ecology* **96**, 2555–2567 (2015).

259. Sandin, S. A. *et al.* Baselines and Degradation of Coral Reefs in the Northern Line Islands. *PLOS ONE* **3**, e1548 (2008).

260. Selig, E. R. & Bruno, J. F. A Global Analysis of the Effectiveness of Marine Protected Areas in Preventing Coral Loss. *PLOS ONE* **5**, e9278 (2010).

261. Benedetti-Cecchi, L. *et al.* Marine protected areas promote stability of reef fish communities under climate warming. *Nat Commun* **15**, 1822 (2024).

262. Mumby, P. J. & Harborne, A. R. Marine Reserves Enhance the Recovery of Corals on Caribbean Reefs. *PLOS ONE* **5**, e8657 (2010).

263. Vanwonterghem, I. & Webster, N. S. Coral Reef Microorganisms in a Changing Climate. *iScience* **23**, 100972 (2020).

264. Apprill, A. *et al.* Toward a New Era of Coral Reef Monitoring. *Environ Sci Technol* **57**, 5117–5124 (2023).

265. Glasl, B., Webster, N. S. & Bourne, D. G. Microbial indicators as a diagnostic tool for assessing water quality and climate stress in coral reef ecosystems. *Mar Biol* **164**, 91 (2017).

266. Nelson, C. E., Wegley Kelly, L. & Haas, A. F. Microbial Interactions with Dissolved Organic Matter Are Central to Coral Reef Ecosystem Function and Resilience. *Annu. Rev. Mar. Sci.* **15**, 431–460 (2023).

267. Terzin, M. *et al.* The road forward to incorporate seawater microbes in predictive reef monitoring. *Environmental Microbiome* **19**, 5 (2024).

268. Bruce, T. *et al.* Abrolhos Bank Reef Health Evaluated by Means of Water Quality, Microbial Diversity, Benthic Cover, and Fish Biomass Data. *PLoS ONE* **7**, e36687 (2012).

269. Fakhraldeen, S. A. *et al.* Shotgun metagenomics reveals the interplay between microbiome diversity and environmental gradients in the first marine protected area in the northern Arabian Gulf. *Front. Microbiol.* **15**, (2025).

270. Robbins, S. J. *et al.* The planktonic microbiome of the Great Barrier Reef. 2025.05.13.653689 Preprint at https://doi.org/10.1101/2025.05.13.653689 (2025).

271. Great Barrier Reef Marine Park Authority. *Great Barrier Reef Outlook Report 2019*. https://elibrary.gbrmpa.gov.au/handle/11017/3474 (2019).

272. Great Barrier Reef Marine Park Authority. *Marine Monitoring Program Annual Report Quality Assurance and Quality Control Manual 2020-21*. https://elibrary.gbrmpa.gov.au/jspui/handle/11017/3932 (2022).

273. Bran & Luebbe. *Directory of Autoanalyser Methods*. (1997).

274. Parsons, T. R., Maita, Y. & Lalli, C. M. *A Manual of Chemical and Biological Methods for Seawater Analysis.* (Pergamon Press, 1984).

275. Ryle, V. D., Mueller, H. R. & Gentien, P. *Automated Analysis of Nutrients in Tropical Sea Waters*. 24 (1981).

276. Valderrama, J. C. The simultaneous analysis of total nitrogen and total phosphorus in natural waters. *Marine Chemistry* **10**, 109–122 (1981).

277. Menzel, D. W. & Corwin, N. The Measurement of Total Phosphorus in Seawater Based on the Liberation of Organically Bound Fractions by Persulfate Oxidation. *Limnology and Oceanography* **10**, 280–282 (1965).

278. Strickland, J. D. H. & Parsons, T. R. A Practical Handbook of Seawater Analysis, 2nd edition. https://doi.org/10.25607/OBP-1791 (1972) doi:10.25607/OBP-1791.

279. Australian Institute of Marine Science. Underway Sensors: Enhanced Measurements from Ships of Opportunity (SOOP). Australian Institute of Marine Science https://doi.org/10.25845/9VR7-9G80 (2015).

280. IMOS. Underway sensors: Enhanced measurements from Ships of Opportunity (SOOP): RV Cape Ferguson | AIMS Data Repository | aims.gov.au. https://apps.aims.gov.au/metadata/view/da560e78-1a4e-43dc-aa4b-c99c3c4ab700 (2015).

281. Terzin, M. *et al.* Gene content of seawater microbes is a strong predictor of water chemistry across the Great Barrier Reef. *Microbiome* **13**, 11 (2025).

282. Froese, R. & Pauly, D. Froese, R. and D. Pauly. Editors. 2024.FishBase. World Wide Web electronic publication. www.fishbase.org, (10/2024). https://fishbase.mnhn.fr/ (2024).

283. Jonker, M., Bray, P., Johns, K. & Osborne, K. *Surveys of Benthic Reef Communities Using Underwater Digital Photography and Counts of Juvenile Corals Long Term Monitoring of the Great Barrier Reef Standard Operational Procedure Number 10*. (2020). doi:10.25845/jjzj-0v14.

284. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

285. Aroney, S. T. N. *et al.* CoverM: Read alignment statistics for metagenomics. Preprint at https://doi.org/10.48550/arXiv.2501.11217 (2025).

286. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

287. Lahti, L. & Shetty, S. microbiome R package. https://doi.org/10.18129/B9.bioc.microbiome (2017) doi:10.18129/B9.bioc.microbiome.

288. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

289. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

290. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

291. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).

292. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* **13**, e1005752 (2017).

293. González-Barrios, F. J. *et al.* Emergent patterns of reef fish diversity correlate with coral assemblage shifts along the Great Barrier Reef. *Nat Commun* **16**, 303 (2025).

294. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, 128 (2017).

295. Lê Cao, K.-A., Martin, P. G., Robert-Granié, C. & Besse, P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10**, 34 (2009).

296. Lê Cao, K.-A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* **7**, Article-35 (2008).

297. Brooks, M. E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* **9**, 378–400 (2017).

298. Hartig, F., Lohse, L. & leite, M. de S. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. (2024).

299. Herren, C. M. & McMahon, K. D. Cohesion: a method for quantifying the connectivity of microbial communities. *ISME J* **11**, 2426–2438 (2017).

300. Veseli, I. *et al.* Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *eLife* **12**, (2024).
301. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
302. Liaw, A. & Wiener, M. Classification and Regression by randomForest. (2002).
303. Adhikary, S., Tiwari, S. P., Banerjee, S., Dwivedi, A. D. & Rahman, S. M. Global marine phytoplankton dynamics analysis with machine learning and reanalyzed remote sensing. *PeerJ* **12**, e17361 (2024).
304. Glasl, B. *et al.* Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* **7**, 94 (2019).
305. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
306. Chaffron, S. *et al.* Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv* **7**, eabg1921 (2021).
307. Cristóbal, E., Ayuso, S. V., Justel, A. & Toro, M. Robust optima and tolerance ranges of biological indicators: a new method to identify sentinels of global warming. *Ecol Res* **29**, 55–68 (2014).
308. Wickham, H. *Ggplot2*. (Springer International Publishing, Cham, 2016).
309. Ceccarelli, D. *et al.* Patterns in the chaos: Scale and the spatiotemporal dynamics of coral reef fish assemblages on the Great Barrier Reef. *Ecosphere* **16**, (2025).
310. Australian Institute of Marine Science. Introduction to circulation and upwelling and why it is important. *eAtlas* https://eatlas.org.au/ne-aus-seascape-connectivity/circulation-upwelling (2018).
311. CSIRO. eReefs hydrodynamic models. *CSIRO eReefs* https://research.csiro.au/ereefs/models/models-about/models-hydrodynamics/ (2023).
312. Castro-Sanguino, C. *et al.* Detecting conservation benefits of marine reserves on remote reefs of the northern GBR. *PLOS ONE* **12**, e0186146 (2017).
313. Osborne, K. *et al.* Delayed coral recovery in a warming ocean. *Glob Chang Biol* **23**, 3869–3881 (2017).
314. Benthuysen, J. A., Emslie, M. J., Currey-Randall, L. M., Cheal, A. J. & Heupel, M. R. Oceanographic influences on reef fish assemblages along the Great Barrier Reef. *Progress in Oceanography* **208**, 102901 (2022).
315. Mumby, P. J. & Steneck, R. S. Coral reef management and conservation in light of rapidly evolving ecological paradigms. *Trends in Ecology & Evolution* **23**, 555–563 (2008).
316. Haas, A. F. *et al.* Global microbialization of coral reefs. *Nat Microbiol* **1**, 16042 (2016).
317. Emslie, M. J., Cheal, A. J. & Johns, K. A. Retention of Habitat Complexity Minimizes Disassembly of Reef Fish Communities following Disturbance: A Large-Scale Natural Experiment. *PLOS ONE* **9**, e105384 (2014).
318. Emslie, M. J., Pratchett, M. S., Cheal, A. J. & Osborne, K. Great Barrier Reef butterflyfish community structure: the role of shelf position and benthic community type. *Coral Reefs* **29**, 705–715 (2010).
319. Gratwicke, B. & Speight, M. R. The relationship between fish species richness, abundance and habitat complexity in a range of shallow tropical marine habitats. *Journal of Fish Biology* **66**, 650–667 (2005).
320. Pratchett, M. S., Berumen, M. L., Marnane, M. J., Eagle, J. V. & Pratchett, D. J. Habitat associations of juvenile versus adult butterflyfishes. *Coral Reefs* **27**, 541–551 (2008).
321. Harder, T., Tebben, J., Möller, M. & Schupp, P. J. Chemical Ecology of Marine Invertebrate Larval Settlement. in *Chemical Ecology* (CRC Press, 2018).
322. Randall, C. J. *et al.* Sexual production of corals for reef restoration in the Anthropocene. *Marine Ecology Progress Series* **635**, 203–232 (2020).
323. Houlbrèque, F. & Ferrier-Pagès, C. Heterotrophy in Tropical Scleractinian Corals. *Biological Reviews* **84**, 1–17 (2009).
324. Zehr, J. & Ward, B. Zehr JP, Ward BB.. Nitrogen cycling in the ocean: New perspectives on processes and paradigms. Appl Environ Microbiol 68: 1015-1024. *Applied and environmental microbiology* **68**, 1015–24 (2002).
325. Zhou, L., Tan, Y. & Huang, L. Coral reef ecological pump for gathering and retaining nutrients and exporting carbon: a review and perspectives. *Acta Oceanol. Sin.* **42**, 1–15 (2023).
326. González-Díaz, P. *et al.* Status of Cuban coral reefs. *Bulletin of Marine Science* **94**, 229–247 (2018).
327. Denux, M. *et al.* Coral Reef Water Microbial Communities of Jardines de la Reina, Cuba. *Microorganisms* **12**, 1822 (2024).
328. Robinson, J. P. W. *et al.* Habitat and fishing control grazing potential on coral reefs. *Functional Ecology* **34**, 240–251 (2020).

329. Barott, K. L. & Rohwer, F. L. Unseen players shape benthic competition on coral reefs. *Trends in Microbiology* **20**, 621–628 (2012).

330. Weber, L. *et al.* Extracellular Reef Metabolites Across the Protected Jardines de la Reina, Cuba Reef System. *Front. Mar. Sci.* **7**, (2020).

331. Redfield, A. C. On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton. in *James Johnstone Memorial Volume* 176–192 (University Press of Liverpool, Liverpool, 1934).

332. Schiettekatte, N. M. D. *et al.* The role of fish feces for nutrient cycling on coral reefs. *Oikos* **2023**, e09914 (2023).

333. Apprill, A. & Salerno, J. L. Reef water microorganisms as diagnostic indicators for coral reef ecosystem management and sustainability. *Cell Reports Sustainability* 100403 (2025) doi:10.1016/j.crsus.2025.100403.

334. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J* **8**, 1553–1565 (2014).

335. Martinez-Gutierrez, C. A. & Aylward, F. O. Strong Purifying Selection Is Associated with Genome Streamlining in Epipelagic Marinimicrobia. *Genome Biology and Evolution* **11**, 2887–2894 (2019).

336. Roda-Garcia, J. J., Haro-Moreno, J. M., Rodriguez-Valera, F., Almagro-Moreno, S. & López-Pérez, M. Single-amplified genomes reveal most streamlined free-living marine bacteria. *Environmental Microbiology* **25**, 1136–1154 (2023).

337. Giordano, N. *et al.* Genome-scale community modelling reveals conserved metabolic cross-feedings in epipelagic bacterioplankton communities. *Nat Commun* **15**, 2721 (2024).

338. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* **3**, 10.1128/mbio.00036-12 (2012).

339. Leite, D. C. A. *et al.* Coral Bacterial-Core Abundance and Network Complexity as Proxies for Anthropogenic Pollution. *Front. Microbiol.* **9**, (2018).

340. Hernandez, D. J., David, A. S., Menges, E. S., Searcy, C. A. & Afkhami, M. E. Environmental stress destabilizes microbial networks. *The ISME Journal* **15**, 1722–1734 (2021).

341. Kajihara, K. T. & Hynson, N. A. Networks as tools for defining emergent properties of microbiomes and their stability. *Microbiome* **12**, 184 (2024).

342. Madigan, M., Bender, K., Buckley, D., Sattley, W. & Stahl, D. *Brock Biology of Microorganisms 15/e*. (2017).

343. Sañudo-Wilhelmy, S. A., Gómez-Consarnau, L., Suffridge, C. & Webb, E. A. The role of B vitamins in marine biogeochemistry. *Ann Rev Mar Sci* **6**, 339–367 (2014).

344. Sañudo-Wilhelmy, S. A. *et al.* Multiple B-vitamin depletion in large areas of the coastal ocean. *Proceedings of the National Academy of Sciences* **109**, 14041–14045 (2012).

345. Haas, A. F. *et al.* Effects of Coral Reef Benthic Primary Producers on Dissolved Organic Carbon and Microbial Activity. *PLOS ONE* **6**, e27973 (2011).

346. Zaneveld, J. R. *et al.* Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales. *Nat Commun* **7**, 11833 (2016).

347. Sparagon, W. J. *et al.* Coral thermal stress and bleaching enrich and restructure reef microbial communities via altered organic matter exudation. *Commun Biol* **7**, 160 (2024).

348. Frade, P. R. *et al.* Spatial patterns of microbial communities across surface waters of the Great Barrier Reef. *Communications Biology* **3**, 442 (2020).

349. Galand, P. E. *et al.* Diversity of the Pacific Ocean coral reef microbiome. *Nat Commun* **14**, 3039 (2023).

350. Salazar, V. W., Verbruggen, H., Marcelino, V. R. & Cao, K.-A. L. Global picoplankton biogeography revealed by metagenomic and climatic data integration. 2024.11.23.624595 Preprint at https://doi.org/10.1101/2024.11.23.624595 (2024).

351. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

352. Bahadori, M. *et al.* The origin of suspended particulate matter in the Great Barrier Reef. *Nat Commun* **14**, 5629 (2023).

353. Charpy, L., Casareto, B. E., Langlade, M. J. & Suzuki, Y. Cyanobacteria in Coral Reef Ecosystems: A Review. *Journal of Marine Sciences* **2012**, e259571 (2012).

354. Furnas, M., Mitchell, A., Skuza, M. & Brodie, J. In the other 90%: phytoplankton responses to enhanced nutrient availability in the Great Barrier Reef Lagoon. *Mar Pollut Bull* **51**, 253–265 (2005).

355. Furnas, M., Alongi, D., Mckinnon, A., Trott, L. & Skuza, M. Regional-scale nitrogen and phosphorus budgets for the northern (14°S) and central (17°S) Great Barrier Reef shelf ecosystem. *Continental Shelf Research - CONT SHELF RES* **31**, (2011).

356. Walsh, K. *et al.* Aura-biomes are present in the water layer above coral reef benthic macro-organisms. *PeerJ* **5**, e3666 (2017).

357. Garcia, B. M. *et al.* Habitat and Hydrodynamics Influence Coral Reef and Seagrass Microbial and Exometabolite Dynamics. 2026.01.12.699023 Preprint at https://doi.org/10.64898/2026.01.12.699023 (2026).

358. Harrison, H. B., Bode, M., Williamson, D. H., Berumen, M. L. & Jones, G. P. A connectivity portfolio effect stabilizes marine reserve performance. *Proceedings of the National Academy of Sciences* **117**, 25595–25600 (2020).

359. Williamson, D. H., Ceccarelli, D. M., Evans, R. D., Hill, J. K. & Russ, G. R. Derelict Fishing Line Provides a Useful Proxy for Estimating Levels of Non-Compliance with No-Take Marine Reserves. *PLOS ONE* **9**, e114395 (2014).

360. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801–812 (2007).

361. Thurber, R. V., Payet, J. P., Thurber, A. R. & Correa, A. M. S. Virus-host interactions and their roles in coral reef health and disease. *Nat Rev Microbiol* **15**, 205–216 (2017).

362. Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).

363. Varona, N. S. *et al.* Host-specific viral predation network on coral reefs. *The ISME Journal* **18**, wrae240 (2024).

364. Baer, J. & Rohwer, F. Coral Reef Microbialization and Viralization Shape Ecosystem Health, Stability, and Resilience. in *Coral Reef Microbiome* (eds Peixoto, R. S. & Voolstra, C. R.) 145–165 (Springer Nature Switzerland, Cham, 2025). doi:10.1007/978-3-031-76692-3_11.

365. Silveira, C. B. & Rohwer, F. L. Piggyback-the-Winner in host-associated microbial communities. *npj Biofilms Microbiomes* **2**, 16010 (2016).

366. Silveira, C. B., Luque, A. & Rohwer, F. The landscape of lysogeny across microbial community density, diversity and energetics. *Environ Microbiol* **23**, 4098–4111 (2021).

367. Silveira, C. B. *et al.* Viral predation pressure on coral reefs. *BMC Biol* **21**, 77 (2023).

368. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* **45**, 1320–1328 (2000).

369. Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A. & Giske, J. A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proceedings of the National Academy of Sciences* **111**, 7813–7818 (2014).

370. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* **28**, 127–181 (2004).

371. Payet, J. P., McMinds, R., Burkepile, D. E. & Vega Thurber, R. L. Unprecedented evidence for high viral abundance and lytic activity in coral reef waters of the South Pacific Ocean. *Front. Microbiol.* **5**, (2014).

372. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754–766 (2018).

373. Worden, A. Z. *et al.* Trophic regulation of Vibrio cholerae in coastal marine waters. *Environmental Microbiology* **8**, 21–29 (2006).

374. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).

375. Hurwitz, B. L. & Sullivan, M. B. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLOS ONE* **8**, e57355 (2013).

376. Alongi, D. M. *et al.* Phytoplankton, bacterioplankton and virioplankton structure and function across the southern Great Barrier Reef shelf. *Journal of Marine Systems* **142**, 25–39 (2015).

377. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microbial Genomics* **8**, 000823 (2022).

378. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* **28**, 569–580 (2018).

379. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* **42**, 1303–1312 (2024).

380. Fang, Z. *et al.* PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8**, giz066 (2019).

381. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).

382. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, (2021).

383. Pradier, L., Tissot, T., Fiston-Lavier, A.-S. & Bedhomme, S. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinformatics* **22**, 349 (2021).

384. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across human gut metagenomes revealed by PlasX and MobMess. *Nat Microbiol* **9**, 830–847 (2024).

385. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research* **47**, D678–D686 (2019).

386. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).

387. Koonin, E. V., Dolja, V. V., Krupovic, M. & Kuhn, J. H. Viruses Defined by the Position of the Virosphere within the Replicator Space. *Microbiol Mol Biol Rev* **85**, e0019320 (2021).

388. Roitman, S., Joseph Pollock, F. & Medina, M. Coral Microbiomes as Bioindicators of Reef Health. in *Population Genomics: Marine Organisms* (eds Oleksiak, M. F. & Rajora, O. P.) 39–57 (Springer International Publishing, Cham, 2018).

389. Nelson, C. E., Wegley Kelly, L. & Haas, A. F. Microbial Interactions with Dissolved Organic Matter Are Central to Coral Reef Ecosystem Function and Resilience. *Annu. Rev. Mar. Sci.* **15**, 431–460 (2023).

390. Bauman, A. G., Burt, J. A., Feary, D. A., Marquis, E. & Usseglio, P. Tropical harmful algal blooms: An emerging threat to coral reef communities? *Marine Pollution Bulletin* **60**, 2117–2122 (2010).

391. Bush, T. *et al.* Oxic-anoxic regime shifts mediated by feedbacks between biogeochemical processes and microbial community dynamics. *Nature Communications* **8**, 789 (2017).

392. Johnson, M. D. *et al.* Rapid ecosystem-scale consequences of acute deoxygenation on a Caribbean coral reef. *Nature Communications* **12**, 4522 (2021).

393. Apprill, A. & Salerno, J. L. Reef water microorganisms as diagnostic indicators for coral reef ecosystem management and sustainability. *Cell Reports Sustainability* **0**, (2025).

394. Bourlat, S. J. *et al.* Genomics in marine monitoring: New opportunities for assessing marine health status. *Marine Pollution Bulletin* **74**, 19–31 (2013).

395. Bengtsson-Palme, J. Chapter 3 - Strategies for Taxonomic and Functional Annotation of Metagenomes. in *Metagenomics* (ed. Nagarajan, M.) 55–79 (Academic Press, 2018). doi:10.1016/B978-0-08-102268-9.00003-3.

396. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

397. Tee, H. S., Waite, D., Lear, G. & Handley, K. M. Microbial river-to-sea continuum: gradients in benthic and planktonic diversity, osmoregulation and nutrient cycling. *Microbiome* **9**, 190 (2021).

398. Tringe, S. G. *et al.* Comparative Metagenomics of Microbial Communities. *Science* **308**, 554–557 (2005).

399. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biology* **5**, e77 (2007).

400. Duarte, C. M. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin* **24**, 11–14 (2015).

401. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol* **9**, e1001177 (2011).

402. Bork, P. *et al.* Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**, 873 (2015).

403. Planes, S. *et al.* The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLoS Biol* **17**, e3000483 (2019).

404. Gorsky, G. *et al.* Expanding Tara Oceans Protocols for Underway, Ecosystemic Sampling of the Ocean-Atmosphere Interface During Tara Pacific Expedition (2016–2018). *Frontiers in Marine Science* **6**, (2019).

405. Belser, C. *et al.* Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition. *Sci Data* **10**, 326 (2023).

406. Lombard, F. *et al.* Open science resources from the Tara Pacific expedition across coral reef and surface ocean ecosystems. *Sci Data* **10**, 324 (2023).

407. Becker, C. C. *et al.* Microorganisms and dissolved metabolites distinguish Florida's Coral Reef habitats. *PNAS Nexus* **2**, pgad287 (2023).

408. Planes, S. & Allemand, D. Insights and achievements from the Tara Pacific expedition. *Nat Commun* **14**, 3131 (2023).

409. Voolstra, C. R. *et al.* Standardized Methods to Assess the Impacts of Thermal Stress on Coral Reef Marine Life. *Annual Review of Marine Science* **17**, 193–226 (2025).

410. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 150023 (2015).

411. Souter, D. *et al.* Status of Coral Reefs of the World: 2020. (2020).

412. Gaudin, M., Eveillard, D. & Chaffron, S. Ecological associations distribution modelling of marine plankton at a global scale. *Philosophical Transactions of the Royal Society B: Biological Sciences* **379**, 20230169 (2024).

413. Kodikara, S. & Lê Cao, K.-A. Microbial network inference for longitudinal microbiome studies with LUPINE. *Microbiome* **13**, 64 (2025).

414. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).

415. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences* **105**, 3805–3810 (2008).

416. Dupont, C. L. *et al.* Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME J* **9**, 1076–1092 (2015).

417. Becker, C. C., Brandt, M., Miller, C. A. & Apprill, A. Microbial bioindicators of Stony Coral Tissue Loss Disease identified in corals and overlying waters using a rapid field-based sequencing approach. *Environ Microbiol* **24**, 1166–1182 (2022).

418. Selig, E. R. *et al.* Analyzing the relationship between ocean temperature anomalies and coral disease outbreaks at broad spatial scales. in *Coastal and Estuarine Studies* (eds Phinney, J. T., Hoegh-Guldberg, O., Kleypas, J., Skirving, W. & Strong, A.) vol. 61 111–128 (American Geophysical Union, Washington, D. C., 2006).

419. Precht, W. F., Gintert, B. E., Robbart, M. L., Fura, R. & van Woesik, R. Unprecedented Disease-Related Coral Mortality in Southeastern Florida. *Sci Rep* **6**, 31374 (2016).

420. Brodnicke, O. B. *et al.* Unravelling the links between heat stress, bleaching and disease: fate of tabular corals following a combined disease and bleaching event. *Coral Reefs* **38**, 591–603 (2019).

421. Doni, L. *et al.* Large-scale impact of the 2016 Marine Heatwave on the plankton-associated microbial communities of the Great Barrier Reef (Australia). *Mar Pollut Bull* **188**, 114685 (2023).

422. Lewis, S. E. *et al.* Herbicides: a new threat to the Great Barrier Reef. *Environ Pollut* **157**, 2470–2484 (2009).

423. Badr, N. B. E., El-Fiky, A. A., Mostafa, A. R. & Al-Mur, B. A. Metal pollution records in core sediments of some Red Sea coastal areas, Kingdom of Saudi Arabia. *Environ Monit Assess* **155**, 509–526 (2009).

424. Haynes, D. & Johnson, J. E. Organochlorine, Heavy Metal and Polyaromatic Hydrocarbon Pollutant Concentrations in the Great Barrier Reef (Australia) Environment: a Review. *Marine Pollution Bulletin* **41**, 267–278 (2000).

425. Becker, C. C., Weber, L., Llopiz, J. K., Mooney, T. A. & Apprill, A. Microorganisms uniquely capture and predict stony coral tissue loss disease and hurricane disturbance impacts on US Virgin Island reefs. *Environ Microbiol* **26**, e16610 (2024).

426. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).

427. Malik, A. A. *et al.* Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *ISME J* **14**, 1–9 (2020).

428. Lennon, J. T. *et al.* Priorities, opportunities, and challenges for integrating microorganisms into Earth system models for climate change prediction. *mBio* **15**, e00455-24 (2024).

429. Webster, N. S., Wagner, M. & Negri, A. P. Microbial conservation in the Anthropocene. *Environ Microbiol* **20**, 1925–1928 (2018).

430. Thomas, M. C. *et al.* Protecting the invisible: Establishing guideline values for copper toxicity to marine microbiomes. *Science of The Total Environment* **904**, 166658 (2023).

431. Thomas, M. C. *et al.* Development of a quantitative PMA-16S rRNA gene sequencing workflow for absolute abundance measurements of seawater microbial communities. Preprint at https://doi.org/10.21203/rs.3.rs-5451626/v1 (2024).

432. Bushon, R. n., Brady, A. m., Likirdopulos, C. a. & Cireddu, J. v. Rapid detection of Escherichia coli and enterococci in recreational water using an immunomagnetic separation/adenosine triphosphate technique. *Journal of Applied Microbiology* **106**, 432–441 (2009).

433. Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. Pepper Mild Mottle Virus as an Indicator of Fecal Pollution. *Appl Environ Microbiol* **75**, 7261–7267 (2009).

434. Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* **9**, 2490–2502 (2015).

435. Laroche, O. *et al.* A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin* **127**, 97–107 (2018).

436. Saito, M. A. *et al.* Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**, 1173–1177 (2014).

437. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).

438. Giovannoni, S. J. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annu. Rev. Mar. Sci.* **9**, 231–255 (2017).

439. Semenza, J. C. *et al.* Environmental Suitability of Vibrio Infections in a Warming Climate: An Early Warning System. *Environ Health Perspect* **125**, 107004 (2017).

440. Villar, E. *et al.* The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Research* **46**, W289–W295 (2018).

441. Vernette, C. *et al.* The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Research* **50**, W516–W526 (2022).

442. Steven, A. D. L. *et al.* eReefs: An operational information system for managing the Great Barrier Reef. *Journal of Operational Oceanography* **12**, S12–S28 (2019).

443. Great Barrier Reef Marine Park Authority. *Great Barrier Reef Marine Monitoring Program Quality Assurance and Quality Control Manual 2022-23*. https://hdl.handle.net/11017/4064 (2024).

444. Waterhouse, J. *et al.* Marine Monitoring Program: Annual report for inshore water quality monitoring 2019-20. https://hdl.handle.net/11017/3826 (2021).

445. Suominen, S. *et al. Engaging Communities to Safeguard Ocean Life: UNESCO Environmental DNA Expeditions*. (2024). doi:10.58337/CBXU3518.

446. Peixoto, R. *et al.* Microbial solutions must be deployed against climate catastrophe. *Nat Microbiol* **9**, 3084–3085 (2024).

7 Appendices

8 Appendix A – Supplementary Material for Chapter 2

Table S1. Statistics on Illumina sequencing before and after quality filtering.

Sample_ID	Raw_counts	After_Trimmomatic	Percentage_retained
11-049-1_S89_R1	11391661	9162136	80.43
11-049-1_S89_R2	11391661	9162136	80.43
11-049-2_S90_R1	9506774	7178113	75.51
11-049-2_S90_R2	9506774	7178113	75.51
11-049-3_S91_R1	19387690	15648432	80.71
11-049-3_S91_R2	19387690	15648432	80.71
11-049-4_S92_R1	19484249	15795415	81.07
11-049-4_S92_R2	19484249	15795415	81.07
11-162-1_S81_R1	16545320	12753615	77.08
11-162-1_S81_R2	16545320	12753615	77.08
11-162-2_S82_R1	14615572	10176877	69.63
11-162-2_S82_R2	14615572	10176877	69.63
11-162-3_S83_R1	19143379	14946146	78.07
11-162-3_S83_R2	19143379	14946146	78.07
11-162-4_S84_R1	22970532	17895321	77.91
11-162-4_S84_R2	22970532	17895321	77.91
13-124-1_S9_R1	19915667	15869540	79.68
13-124-1_S9_R2	19915667	15869540	79.68
13-124-2_S10_R1	22747351	16339989	71.83
13-124-2_S10_R2	22747351	16339989	71.83
13-124-3_S11_R1	16804263	13278069	79.02
13-124-3_S11_R2	16804263	13278069	79.02
13-124-4_S12_R1	21950573	17036503	77.61
13-124-4_S12_R2	21950573	17036503	77.61
21-550-1_S69_R1	20100727	16218922	80.69
21-550-1_S69_R2	20100727	16218922	80.69
21-550-2_S70_R1	22553087	18460291	81.85
21-550-2_S70_R2	22553087	18460291	81.85
21-550-3_S71_R1	21563329	16412859	76.11
21-550-3_S71_R2	21563329	16412859	76.11
21-550-4_S72_R1	26678391	20829016	78.07
21-550-4_S72_R2	26678391	20829016	78.07
21-580-1_S57_R1	21996050	17534464	79.72
21-580-1_S57_R2	21996050	17534464	79.72
21-580-2_S58_R1	17510809	14002231	79.96

21-580-2_S58_R2	17510809	14002231	79.96
21-580-3_S59_R1	21164806	16686850	78.84
21-580-3_S59_R2	21164806	16686850	78.84
21-580-4_S60_R1	22786331	18154819	79.67
21-580-4_S60_R2	22786331	18154819	79.67
22-084-1_S41_R1	21554612	17317715	80.34
22-084-1_S41_R2	21554612	17317715	80.34
22-084-2_S42_R1	23120185	18544566	80.21
22-084-2_S42_R2	23120185	18544566	80.21
22-084-3_S43_R1	20633510	16549809	80.21
22-084-3_S43_R2	20633510	16549809	80.21
22-084-4_S44_R1	18036072	13208697	73.23
22-084-4_S44_R2	18036072	13208697	73.23
Agincourt1-1_S33_R1	18208575	14678322	80.61
Agincourt1-1_S33_R2	18208575	14678322	80.61
Agincourt1-2_S34_R1	14998998	11433509	76.23
Agincourt1-2_S34_R2	14998998	11433509	76.23
Agincourt1-3_S35_R1	15827149	13030639	82.33
Agincourt1-3_S35_R2	15827149	13030639	82.33
Agincourt1-4_S36_R1	16259240	13367259	82.21
Agincourt1-4_S36_R2	16259240	13367259	82.21
Arlington-1_S37_R1	12936835	10580761	81.79
Arlington-1_S37_R2	12936835	10580761	81.79
Arlington-2_S38_R1	12186989	9966004	81.78
Arlington-2_S38_R2	12186989	9966004	81.78
Arlington-3_S39_R1	12372043	9706692	78.46
Arlington-3_S39_R2	12372043	9706692	78.46
Arlington-4_S40_R1	20072805	16473992	82.07
Arlington-4_S40_R2	20072805	16473992	82.07
Boult-1_S25_R1	21121009	16974878	80.37
Boult-1_S25_R2	21121009	16974878	80.37
Boult-2_S26_R1	22444634	18484731	82.36
Boult-2_S26_R2	22444634	18484731	82.36
Boult-3_S27_R1	16841533	13434588	79.77
Boult-3_S27_R2	16841533	13434588	79.77
Boult-4_S28_R1	18856352	15363196	81.47
Boult-4_S28_R2	18856352	15363196	81.47
Broomfield-1_S49_R1	18982594	15580233	82.08
Broomfield-1_S49_R2	18982594	15580233	82.08
Broomfield-2_S50_R1	48579	3767	7.75
Broomfield-2_S50_R2	48579	3767	7.75

Broomfield-3_S51_R1	19790260	16241262	82.07
Broomfield-3_S51_R2	19790260	16241262	82.07
Broomfield-4_S52_R1	24245268	19962172	82.33
Broomfield-4_S52_R2	24245268	19962172	82.33
Broomfield-rpt-2_S115_R1	16677916	13077884	78.41
Broomfield-rpt-2_S115_R2	16677916	13077884	78.41
Centipede-1_S57_R1	14292259	10763055	75.31
Centipede-1_S57_R2	14292259	10763055	75.31
Centipede-2_S58_R1	15073864	12048345	79.93
Centipede-2_S58_R2	15073864	12048345	79.93
Centipede-3_S59_R1	13760054	11090867	80.6
Centipede-3_S59_R2	13760054	11090867	80.6
Centipede-4_S60_R1	14408385	11474288	79.64
Centipede-4_S60_R2	14408385	11474288	79.64
Chicken-1_S69_R1	13205982	10405605	78.79
Chicken-1_S69_R2	13205982	10405605	78.79
Chicken-2_S70_R1	17214103	13766095	79.97
Chicken-2_S70_R2	17214103	13766095	79.97
Chicken-3_S71_R1	13245279	10339532	78.06
Chicken-3_S71_R2	13245279	10339532	78.06
Chicken-4_S72_R1	15847423	12473351	78.71
Chicken-4_S72_R2	15847423	12473351	78.71
Chinaman-1_S65_R1	19371128	15913263	82.15
Chinaman-1_S65_R2	19371128	15913263	82.15
Chinaman-2_S66_R1	21106100	16757447	79.4
Chinaman-2_S66_R2	21106100	16757447	79.4
Chinaman-3_S67_R1	19379451	15128925	78.07
Chinaman-3_S67_R2	19379451	15128925	78.07
Chinaman-4_S68_R1	21248990	16928948	79.67
Chinaman-4_S68_R2	21248990	16928948	79.67
Corbett-1_S17_R1	14621387	11474069	78.47
Corbett-1_S17_R2	14621387	11474069	78.47
Corbett-2_S18_R1	20871896	16470187	78.91
Corbett-2_S18_R2	20871896	16470187	78.91
Corbett-3_S19_R1	22095113	16567974	74.98
Corbett-3_S19_R2	22095113	16567974	74.98
Corbett-4_S20_R1	19415512	15233388	78.46
Corbett-4_S20_R2	19415512	15233388	78.46
Davie-1_S1_R1	22015462	17908923	81.35
Davie-1_S1_R2	22015462	17908923	81.35

Davie-2_S2_R1	19501345	15511415	79.54
Davie-2_S2_R2	19501345	15511415	79.54
Davie-3_S3_R1	20188921	16126387	79.88
Davie-3_S3_R2	20188921	16126387	79.88
Davie-4_S4_R1	12946457	10219690	78.94
Davie-4_S4_R2	12946457	10219690	78.94
Erskine-1_S61_R1	14284277	10848583	75.95
Erskine-1_S61_R2	14284277	10848583	75.95
Erskine-2_S62_R1	13439806	10151986	75.54
Erskine-2_S62_R2	13439806	10151986	75.54
Erskine-3_S63_R1	15193999	11527881	75.87
Erskine-3_S63_R2	15193999	11527881	75.87
Erskine-4_S64_R1	16105531	12841271	79.73
Erskine-4_S64_R2	16105531	12841271	79.73
Fairfax-1_S33_R1	27456473	21098696	76.84
Fairfax-1_S33_R2	27456473	21098696	76.84
Fairfax-2_S34_R1	23861979	18895715	79.19
Fairfax-2_S34_R2	23861979	18895715	79.19
Fairfax-3_S35_R1	22929934	17993982	78.47
Fairfax-3_S35_R2	22929934	17993982	78.47
Fairfax-4_S36_R1	22162467	18334455	82.73
Fairfax-4_S36_R2	22162467	18334455	82.73
Farquaharson-1_S1_R1	14635499	11660269	79.67
Farquaharson-1_S1_R2	14635499	11660269	79.67
Farquaharson-2_S2_R1	15345357	11844876	77.19
Farquaharson-2_S2_R2	15345357	11844876	77.19
Farquaharson-3_S3_R1	12975351	10519862	81.08
Farquaharson-3_S3_R2	12975351	10519862	81.08
Farquaharson-4_S4_R1	13121583	10593460	80.73
Farquaharson-4_S4_R2	13121583	10593460	80.73
Feather-1_S5_R1	10834100	8679881	80.12
Feather-1_S5_R2	10834100	8679881	80.12
Feather-2_S6_R1	17316450	13332799	76.99
Feather-2_S6_R2	17316450	13332799	76.99
Feather-3_S7_R1	12855246	10216539	79.47
Feather-3_S7_R2	12855246	10216539	79.47
Feather-4_S8_R1	11193091	8782905	78.47
Feather-4_S8_R2	11193091	8782905	78.47
Fork-1_S49_R1	13227928	10508991	79.45
Fork-1_S49_R2	13227928	10508991	79.45

Fork-2_S50_R1	14791708	11564354	78.18
Fork-2_S50_R2	14791708	11564354	78.18
Fork-3_S51_R1	12033145	9399583	78.11
Fork-3_S51_R2	12033145	9399583	78.11
Fork-4_S52_R1	11840621	9059843	76.51
Fork-4_S52_R2	11840621	9059843	76.51
Grub-1_S65_R1	16345885	12830110	78.49
Grub-1_S65_R2	16345885	12830110	78.49
Grub-2_S66_R1	12862409	9968730	77.5
Grub-2_S66_R2	12862409	9968730	77.5
Grub-3_S67_R1	16660304	12962238	77.8
Grub-3_S67_R2	16660304	12962238	77.8
Grub-4_S68_R1	15196158	11898219	78.3
Grub-4_S68_R2	15196158	11898219	78.3
Hastings-1_S41_R1	15694329	12439223	79.26
Hastings-1_S41_R2	15694329	12439223	79.26
Hastings-2_S42_R1	14203851	10859760	76.46
Hastings-2_S42_R2	14203851	10859760	76.46
Hastings-3_S43_R1	17815673	14399848	80.83
Hastings-3_S43_R2	17815673	14399848	80.83
Hastings-4_S44_R1	16618775	13580106	81.72
Hastings-4_S44_R2	16618775	13580106	81.72
Hedley-1_S21_R1	13054511	10574148	81
Hedley-1_S21_R2	13054511	10574148	81
Hedley-2_S22_R1	14054961	11098597	78.97
Hedley-2_S22_R2	14054961	11098597	78.97
Hedley-3_S23_R1	13165314	10729411	81.5
Hedley-3_S23_R2	13165314	10729411	81.5
Helix-1_S61_R1	16564330	13202851	79.71
Helix-1_S61_R2	16564330	13202851	79.71
Helix-2_S62_R1	13562187	10585635	78.05
Helix-2_S62_R2	13562187	10585635	78.05
Helix-3_S63_R1	19197802	15335565	79.88
Helix-3_S63_R2	19197802	15335565	79.88
Helix-4_S64_R1	13754412	10555498	76.74
Helix-4_S64_R2	13754412	10555498	76.74
Hoskyn-1_S29_R1	18943463	15251668	80.51
Hoskyn-1_S29_R2	18943463	15251668	80.51
Hoskyn-2_S30_R1	20376534	16323841	80.11
Hoskyn-2_S30_R2	20376534	16323841	80.11

Hoskyn-3_S31_R1	23809747	19714232	82.8
Hoskyn-3_S31_R2	23809747	19714232	82.8
Hoskyn-4_S32_R1	21466806	17025580	79.31
Hoskyn-4_S32_R2	21466806	17025580	79.31
JohnBrewer-1_S93_R1	20229504	15505862	76.65
JohnBrewer-1_S93_R2	20229504	15505862	76.65
JohnBrewer-2_S94_R1	17866582	13087586	73.25
JohnBrewer-2_S94_R2	17866582	13087586	73.25
JohnBrewer-3_S97_R1	15091104	12222232	80.99
JohnBrewer-3_S97_R2	15091104	12222232	80.99
JohnBrewer-4_S98_R1	18972654	15490727	81.65
JohnBrewer-4_S98_R2	18972654	15490727	81.65
Kelso-1_S85_R1	18183883	14573742	80.15
Kelso-1_S85_R2	18183883	14573742	80.15
Kelso-2_S86_R1	15760712	12396041	78.65
Kelso-2_S86_R2	15760712	12396041	78.65
Kelso-3_S87_R1	16438019	13053054	79.41
Kelso-3_S87_R2	16438019	13053054	79.41
Kelso-4_S88_R1	14210790	11115237	78.22
Kelso-4_S88_R2	14210790	11115237	78.22
Knife-1_S45_R1	10158905	7604403	74.85
Knife-1_S45_R2	10158905	7604403	74.85
Knife-2_S46_R1	10123508	7867437	77.71
Knife-2_S46_R2	10123508	7867437	77.71
Knife-3_S47_R1	13637941	10614685	77.83
Knife-3_S47_R2	13637941	10614685	77.83
Knife-4_S48_R1	12792248	9071031	70.91
Knife-4_S48_R2	12792248	9071031	70.91
Lagoon-1_S13_R1	23078370	18229727	78.99
Lagoon-1_S13_R2	23078370	18229727	78.99
Lagoon-2_S14_R1	19727931	15915193	80.67
Lagoon-2_S14_R2	19727931	15915193	80.67
Lagoon-3_S15_R1	20953079	17713070	84.54
Lagoon-3_S15_R2	20953079	17713070	84.54
Lagoon-4_S16_R1	22291127	17777877	79.75
Lagoon-4_S16_R2	22291127	17777877	79.75
LittleKelso-1_S81_R1	14780869	11862978	80.26
LittleKelso-1_S81_R2	14780869	11862978	80.26
LittleKelso-2_S82_R1	16355172	12732746	77.85
LittleKelso-2_S82_R2	16355172	12732746	77.85

LittleKelso-3_S83_R1	17771440	14193671	79.87
LittleKelso-3_S83_R2	17771440	14193671	79.87
LittleKelso-4_S84_R1	14134872	10782383	76.28
LittleKelso-4_S84_R2	14134872	10782383	76.28
Lynchs-1_S99_R1	17440400	13917717	79.8
Lynchs-1_S99_R2	17440400	13917717	79.8
Lynchs-2_S100_R1	13916361	10377547	74.57
Lynchs-2_S100_R2	13916361	10377547	74.57
Lynchs-3_S101_R1	18150036	14281813	78.69
Lynchs-3_S101_R2	18150036	14281813	78.69
Lynchs-4_S102_R1	16006063	12847768	80.27
Lynchs-4_S102_R2	16006063	12847768	80.27
Lynchs-PF-1_S107_R1	14641024	11420516	78
Lynchs-PF-1_S107_R2	14641024	11420516	78
Lynchs-PF-2_S108_R1	12044789	9288669	77.12
Lynchs-PF-2_S108_R2	12044789	9288669	77.12
Lynchs-PF-3_S109_R1	15864692	12141196	76.53
Lynchs-PF-3_S109_R2	15864692	12141196	76.53
Lynchs-PF-4_S110_R1	16330916	11912270	72.94
Lynchs-PF-4_S110_R2	16330916	11912270	72.94
Mantis-1_S85_R1	18993466	14558201	76.65
Mantis-1_S85_R2	18993466	14558201	76.65
Mantis-2_S86_R1	21619816	15635305	72.32
Mantis-2_S86_R2	21619816	15635305	72.32
Mantis-3_S87_R1	18180195	14425519	79.35
Mantis-3_S87_R2	18180195	14425519	79.35
Mantis-4_S88_R1	19588794	15831087	80.82
Mantis-4_S88_R2	19588794	15831087	80.82
Masthead-1_S53_R1	24458492	19381421	79.24
Masthead-1_S53_R2	24458492	19381421	79.24
Masthead-2_S54_R1	17157643	13676920	79.71
Masthead-2_S54_R2	17157643	13676920	79.71
Masthead-3_S55_R1	21267856	16708121	78.56
Masthead-3_S55_R2	21267856	16708121	78.56
Masthead-4_S56_R1	25186255	19727992	78.33
Masthead-4_S56_R2	25186255	19727992	78.33
McCulloch-1_S17_R1	16259643	13239204	81.42
McCulloch-1_S17_R2	16259643	13239204	81.42
McCulloch-2_S18_R1	15248329	12469021	81.77
McCulloch-2_S18_R2	15248329	12469021	81.77

McCulloch-3_S19_R1	21195708	17380964	82
McCulloch-3_S19_R2	21195708	17380964	82
McCulloch-4_S20_R1	12440154	10216023	82.12
McCulloch-4_S20_R2	12440154	10216023	82.12
McSweeney-1_S5_R1	21741600	17088561	78.6
McSweeney-1_S5_R2	21741600	17088561	78.6
McSweeney-2_S6_R1	19967227	15684275	78.55
McSweeney-2_S6_R2	19967227	15684275	78.55
McSweeney-3_S7_R1	27085493	22004592	81.24
McSweeney-3_S7_R2	27085493	22004592	81.24
McSweeney-4_S8_R1	23783727	18680695	78.54
McSweeney-4_S8_R2	23783727	18680695	78.54
Monsoon-1_S21_R1	20358282	14560732	71.52
Monsoon-1_S21_R2	20358282	14560732	71.52
Monsoon-2_S22_R1	23634715	18519565	78.36
Monsoon-2_S22_R2	23634715	18519565	78.36
Monsoon-3_S23_R1	19236392	15353922	79.82
Monsoon-3_S23_R2	19236392	15353922	79.82
Monsoon-4_S24_R1	20673438	16903268	81.76
Monsoon-4_S24_R2	20673438	16903268	81.76
Moore-1_S25_R1	18190924	14597050	80.24
Moore-1_S25_R2	18190924	14597050	80.24
Moore-2_S26_R1	14319030	11477735	80.16
Moore-2_S26_R2	14319030	11477735	80.16
Moore-3_S27_R1	12845546	10118946	78.77
Moore-3_S27_R2	12845546	10118946	78.77
Moore-4_S28_R1	15775687	12604975	79.9
Moore-4_S28_R2	15775687	12604975	79.9
Myrmidon-1_S53_R1	9585650	7487447	78.11
Myrmidon-1_S53_R2	9585650	7487447	78.11
Myrmidon-2_S54_R1	11445140	8866528	77.47
Myrmidon-2_S54_R2	11445140	8866528	77.47
Myrmidon-3_S55_R1	11835347	9243782	78.1
Myrmidon-3_S55_R2	11835347	9243782	78.1
Myrmidon-4_S56_R1	12884881	10408827	80.78
Myrmidon-4_S56_R2	12884881	10408827	80.78
Myrmidon-PF-1_S111_R1	15752444	12008135	76.23
Myrmidon-PF-1_S111_R2	15752444	12008135	76.23
Myrmidon-PF-2_S112_R1	15638842	11648815	74.49
Myrmidon-PF-2_S112_R2	15638842	11648815	74.49

Myrmidon-PF-3_S113_R1	15632059	10240030	65.51
Myrmidon-PF-3_S113_R2	15632059	10240030	65.51
Myrmidon-PF-4_S114_R1	14295862	10984119	76.83
Myrmidon-PF-4_S114_R2	14295862	10984119	76.83
Neg-control-1_S101_R1	136397	42076	30.85
Neg-control-1_S101_R2	136397	42076	30.85
Neg-control-2_S24_R1	230229	108061	46.94
Neg-control-2_S24_R2	230229	108061	46.94
Neg-control-3_S116_R1	154621	73562	47.58
Neg-control-3_S116_R2	154621	73562	47.58
North-1_S37_R1	15893048	12712728	79.99
North-1_S37_R2	15893048	12712728	79.99
North-2_S38_R1	19790656	16189274	81.8
North-2_S38_R2	19790656	16189274	81.8
North-3_S39_R1	24246873	19981797	82.41
North-3_S39_R2	24246873	19981797	82.41
North-4_S40_R1	19528038	16037916	82.13
North-4_S40_R2	19528038	16037916	82.13
Peart-1_S13_R1	12498052	10061259	80.5
Peart-1_S13_R2	12498052	10061259	80.5
Peart-2_S14_R1	13788722	11231822	81.46
Peart-2_S14_R2	13788722	11231822	81.46
Peart-3_S15_R1	11397354	9065275	79.54
Peart-3_S15_R2	11397354	9065275	79.54
Peart-4_S16_R1	12814561	9894545	77.21
Peart-4_S16_R2	12814561	9894545	77.21
Rib-1_S73_R1	16349857	12907883	78.95
Rib-1_S73_R2	16349857	12907883	78.95
Rib-2_S74_R1	18617365	14801568	79.5
Rib-2_S74_R2	18617365	14801568	79.5
Rib-3_S75_R1	19502285	15618797	80.09
Rib-3_S75_R2	19502285	15618797	80.09
Rib-4_S76_R1	17561081	13809062	78.63
Rib-4_S76_R2	17561081	13809062	78.63
Rib-PF-1_S103_R1	14358252	11178583	77.85
Rib-PF-1_S103_R2	14358252	11178583	77.85
Rib-PF-2_S104_R1	14560465	10962656	75.29
Rib-PF-2_S104_R2	14560465	10962656	75.29
Rib-PF-3_S105_R1	13061479	9981761	76.42
Rib-PF-3_S105_R2	13061479	9981761	76.42

Rib-PF-4_S106_R1	14388545	10997289	76.43
Rib-PF-4_S106_R2	14388545	10997289	76.43
Roxburgh-1_S89_R1	15711737	12658785	80.57
Roxburgh-1_S89_R2	15711737	12658785	80.57
Roxburgh-2_S90_R1	17574836	13869737	78.92
Roxburgh-2_S90_R2	17574836	13869737	78.92
Roxburgh-3_S91_R1	14481635	11658659	80.51
Roxburgh-3_S91_R2	14481635	11658659	80.51
Roxburgh-4_S92_R1	16622610	13266811	79.81
Roxburgh-4_S92_R2	16622610	13266811	79.81
Sanbank1-1_S77_R1	21458501	17368212	80.94
Sanbank1-1_S77_R2	21458501	17368212	80.94
Sanbank1-2_S78_R1	18263182	15192787	83.19
Sanbank1-2_S78_R2	18263182	15192787	83.19
Sanbank1-3_S79_R1	23266034	18571860	79.82
Sanbank1-3_S79_R2	23266034	18571860	79.82
Sanbank1-4_S80_R1	19199687	15386064	80.14
Sanbank1-4_S80_R2	19199687	15386064	80.14
SmallLagoon-1_S45_R1	20241764	16190538	79.99
SmallLagoon-1_S45_R2	20241764	16190538	79.99
SmallLagoon-2_S46_R1	22897812	19044958	83.17
SmallLagoon-2_S46_R2	22897812	19044958	83.17
SmallLagoon-3_S47_R1	19426244	15977442	82.25
SmallLagoon-3_S47_R2	19426244	15977442	82.25
SmallLagoon-4_S48_R1	16892626	13129683	77.72
SmallLagoon-4_S48_R2	16892626	13129683	77.72
St-Crispin-1_S73_R1	20269787	16730735	82.54
St-Crispin-1_S73_R2	20269787	16730735	82.54
St-Crispin-2_S74_R1	22821633	18780890	82.29
St-Crispin-2_S74_R2	22821633	18780890	82.29
St-Crispin-3_S75_R1	22609122	17726032	78.4
St-Crispin-3_S75_R2	22609122	17726032	78.4
St-Crispin-4_S76_R1	15583537	12224682	78.45
St-Crispin-4_S76_R2	15583537	12224682	78.45
Taylor-1_S9_R1	19551092	15557593	79.57
Taylor-1_S9_R2	19551092	15557593	79.57
Taylor-2_S10_R1	13338249	10343864	77.55
Taylor-2_S10_R2	13338249	10343864	77.55
Taylor-3_S11_R1	14349624	11147360	77.68
Taylor-3_S11_R2	14349624	11147360	77.68

Taylor-4_S12_R1	11147073	8838527	79.29
Taylor-4_S12_R2	11147073	8838527	79.29
Thetford-1_S29_R1	19374041	16034796	82.76
Thetford-1_S29_R2	19374041	16034796	82.76
Thetford-2_S30_R1	16355562	12740151	77.89
Thetford-2_S30_R2	16355562	12740151	77.89
Thetford-3_S31_R1	18511046	15217895	82.21
Thetford-3_S31_R2	18511046	15217895	82.21
Thetford-4_S32_R1	15617475	12342274	79.03
Thetford-4_S32_R2	15617475	12342274	79.03
Average	17046637	13485411	78.84
SD	4490520	3662725	7.23

Table S2. Final Illumina sequencing counts after 5 additional filtering steps in R, i.e. after removing (1) non-annotated reads; taxa annotated as (2) eukaryotic or (3) viral; (4) prokaryotic reads annotated to the Domain level only (Bacteria or Archaea); and (5) rare/spurious reads (relative abundance < 0.0001%). These values are only reported for Forward reads (R1 samples).

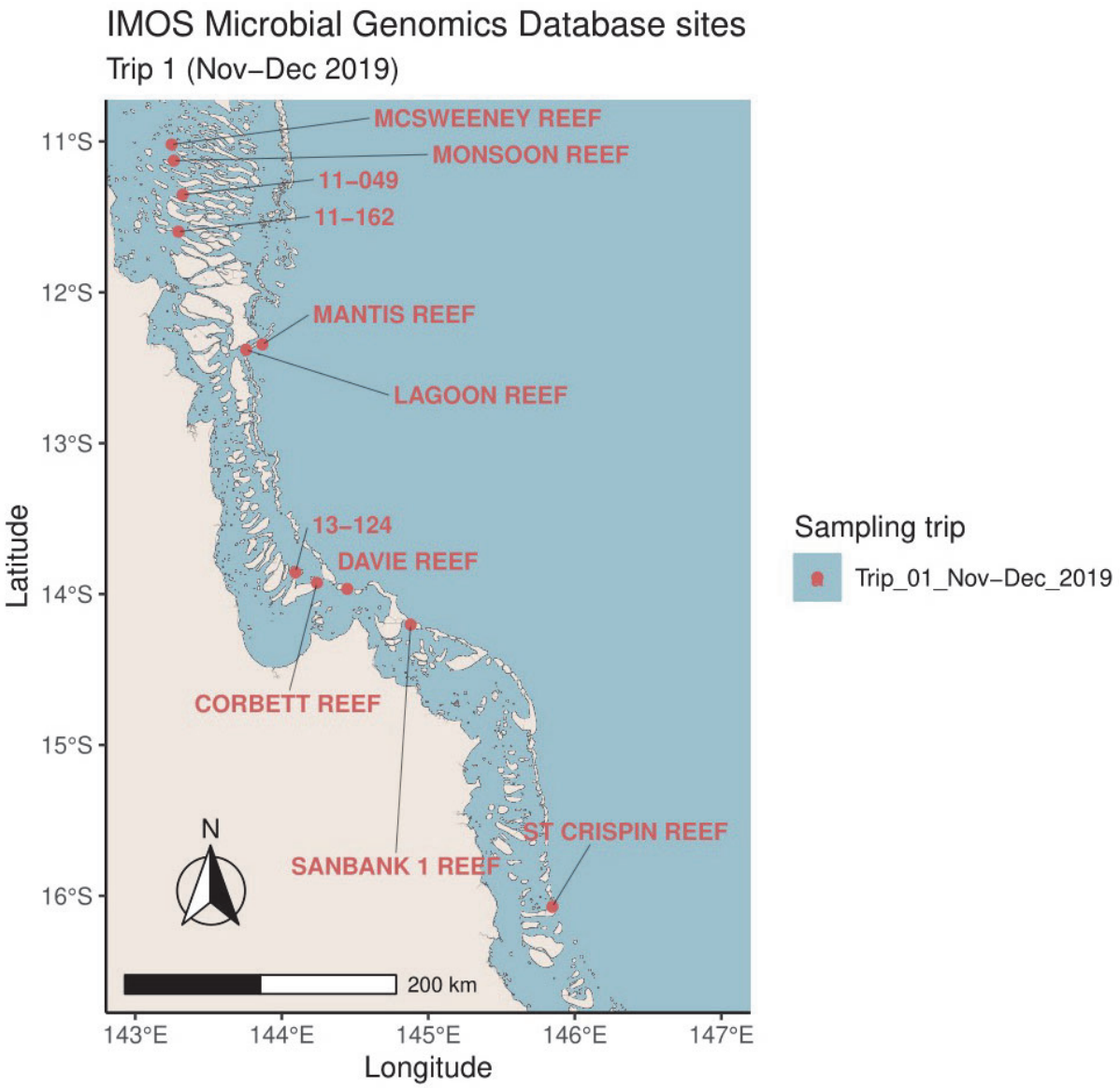

Figure S1: Reef sites pertinent to the Trip 1 sampling event, conducted in November and December 2019, in the Cape Grenville and Princess Charlotte bay sectors of the northern GBR.

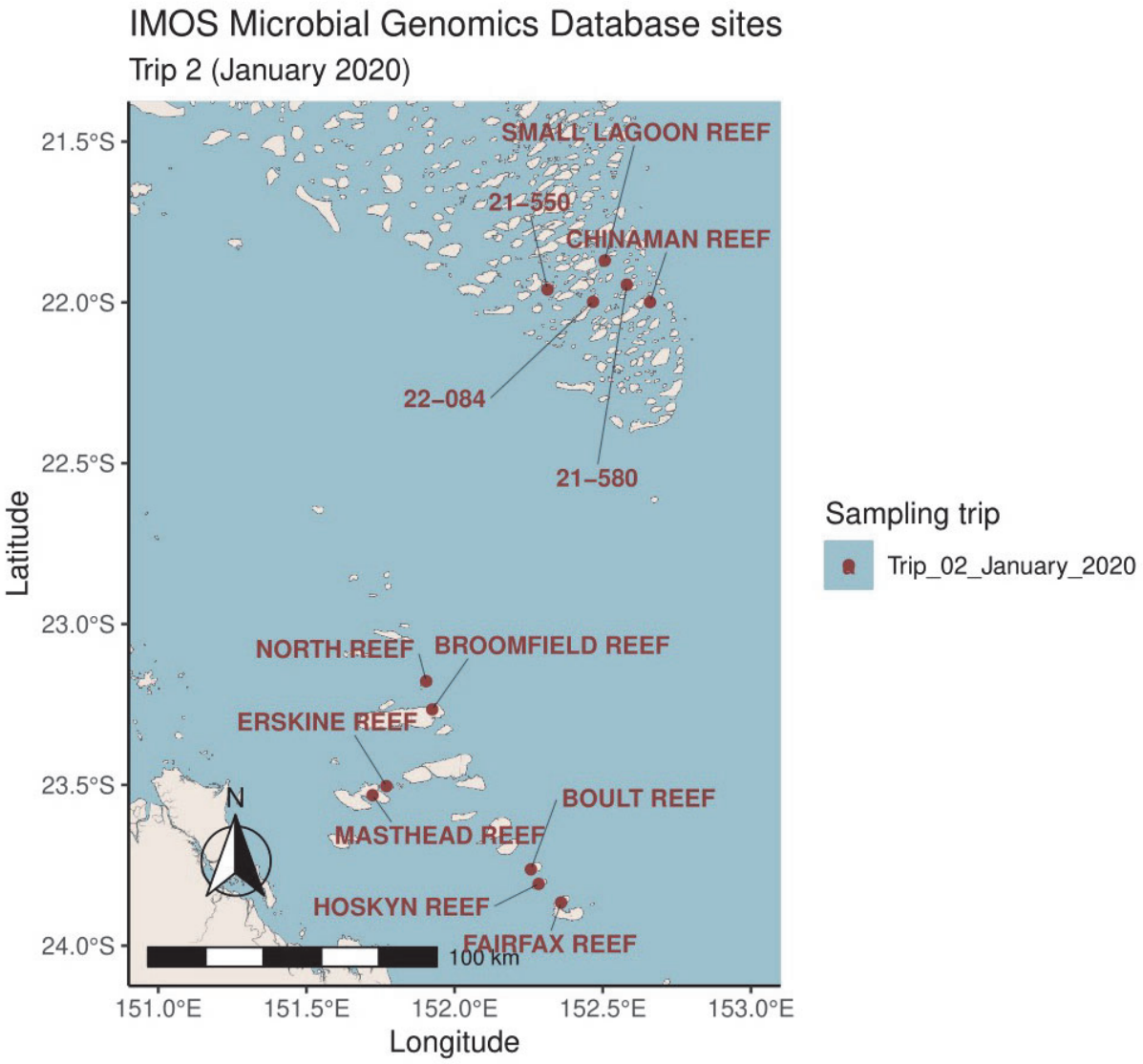

Figure S2: Reef sites pertinent to the Trip 2 sampling event, conducted in January 2020, in the Swains and Capricorn Bunker sectors of the southern GBR.

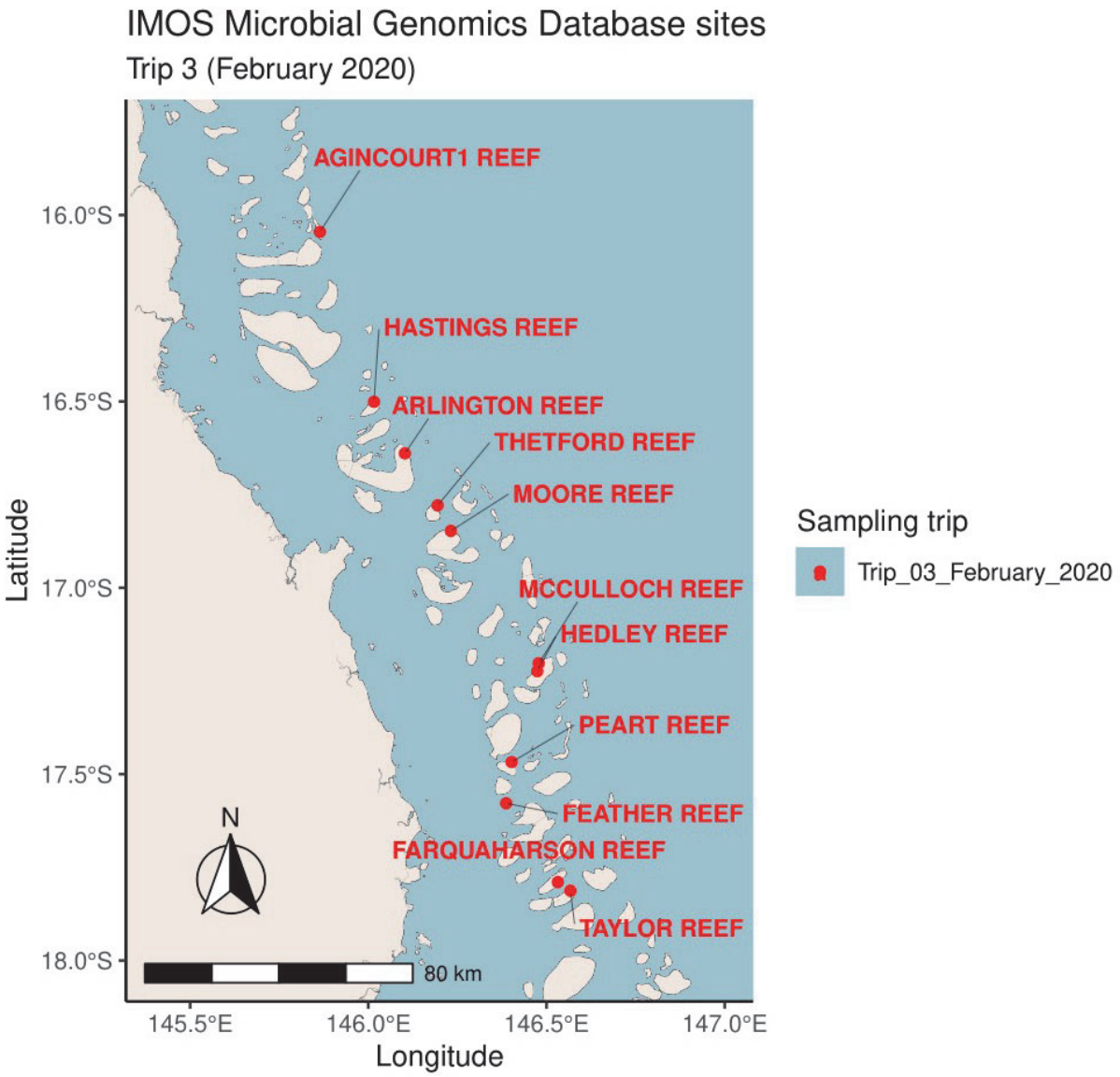

Figure S3: Reef sites pertinent to the Trip 3 sampling event, conducted in February 2020, in the Cairns and Inninsfail sectors of the central GBR.

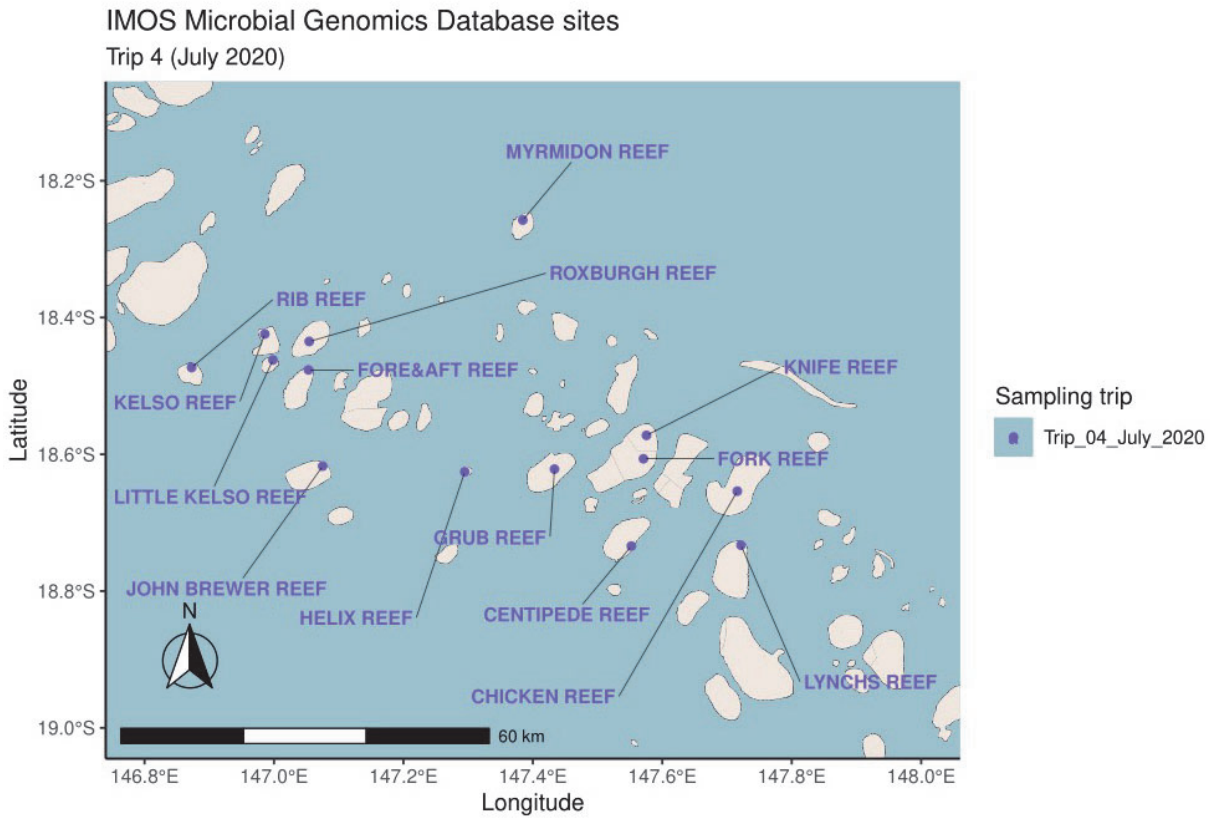

Figure S4: Reef sites pertinent to the Trip 4 sampling event, conducted in July 2020, in the Townsville sector of the central GBR.

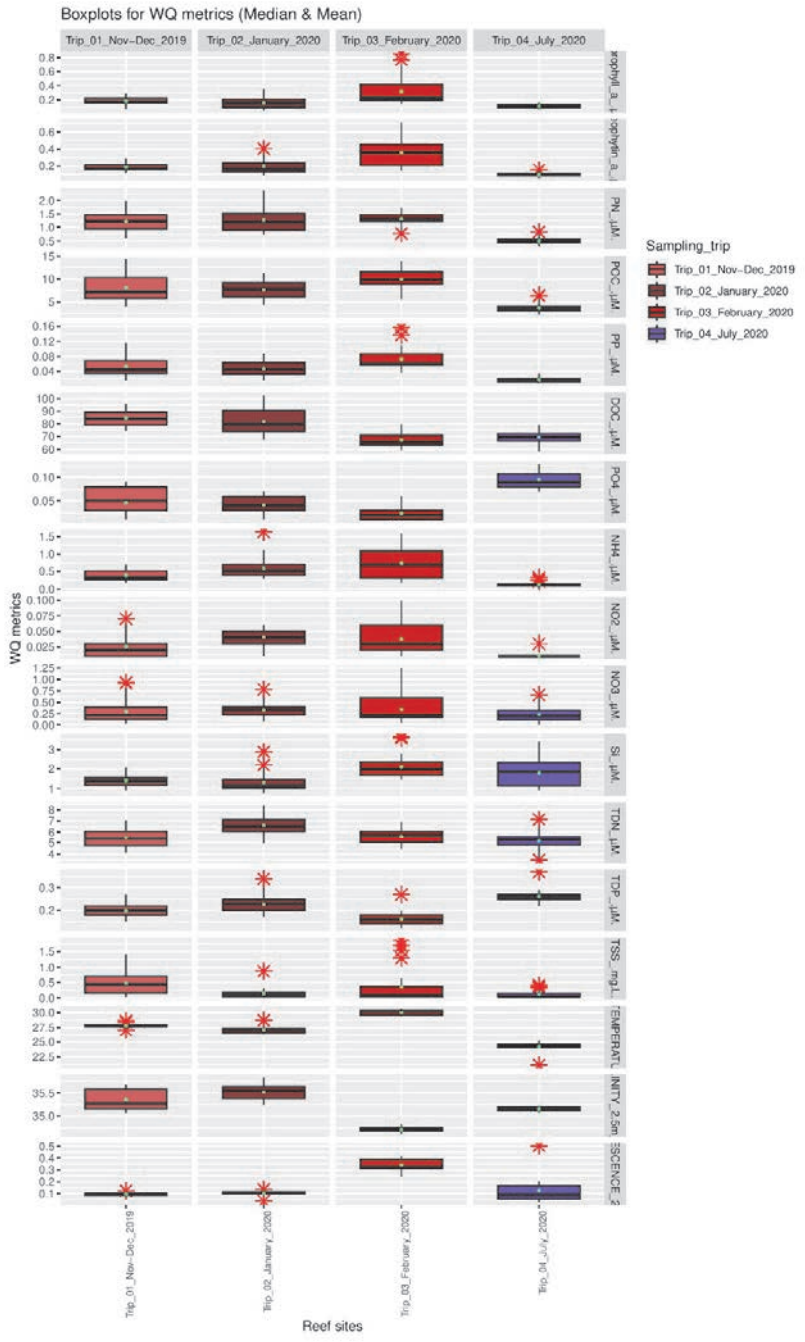

Figure S5: Physico-chemical data. Median ± SD values of 17 physico-chemical variables collected. Values are summarised across the four sampling trips, with the colour code corresponding to Fig. 2.1 in the main text.

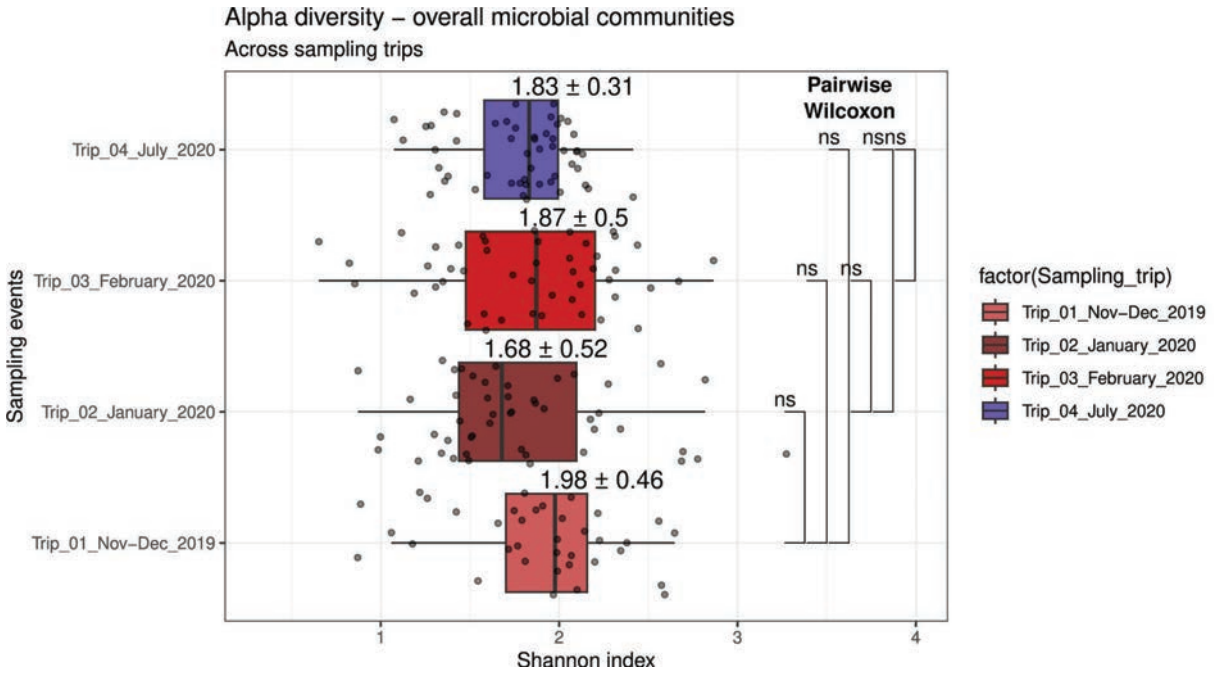

Figure S6. Boxplots illustrate microbial diversity (Shannon Index) for genera *for overall microbial communities*, across sampling trips. The symbols *, **, ***, and **** denote levels of statistical significance in pairwise Wilcoxon rank sum tests when testing variation of Shannon diversity scores for overall microbial communities across the four sampling trips: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$, indicating increasing levels of significance. 'ns' indicates non-significant results, where $p \geq 0.05$.

Table S3. Median and standard deviation for Shannon Index values, computed within trips.

Sampling_trip	**Median**	**SD**
Trip_01_Nov-Dec_2019	1.976648	0.4646255
Trip_02_January_2020	1.678609	0.5192536
Trip_03_February_2020	1.871546	0.4974857
Trip_04_July_2020	1.831793	0.3093137

Table S4. Pairwise Wilcoxon rank sum tests to compare median Shannon Diversity between sampling trips, computed for overall communities.

group1	group2	n1	n2	**statistic p**	**p.adj p.adj.signif**
Trip_01_Nov-Dec_2019	Trip_02_January_2020	36	48	1030 0.135	0.81 ns
Trip_01_Nov-Dec_2019	Trip_03_February_2020	36	47	887 0.711	1.00 ns
Trip_01_Nov-Dec_2019	Trip_04_July_2020	36	52	1106 0.150	0.81 ns
Trip_02_January_2020	Trip_03_February_2020	48	47	1010 0.383	1.00 ns
Trip_02_January_2020	Trip_04_July_2020	48	52	1139 0.454	1.00 ns
Trip_03_February_2020	Trip_04_July_2020	47	52	1361 0.332	1.00 ns

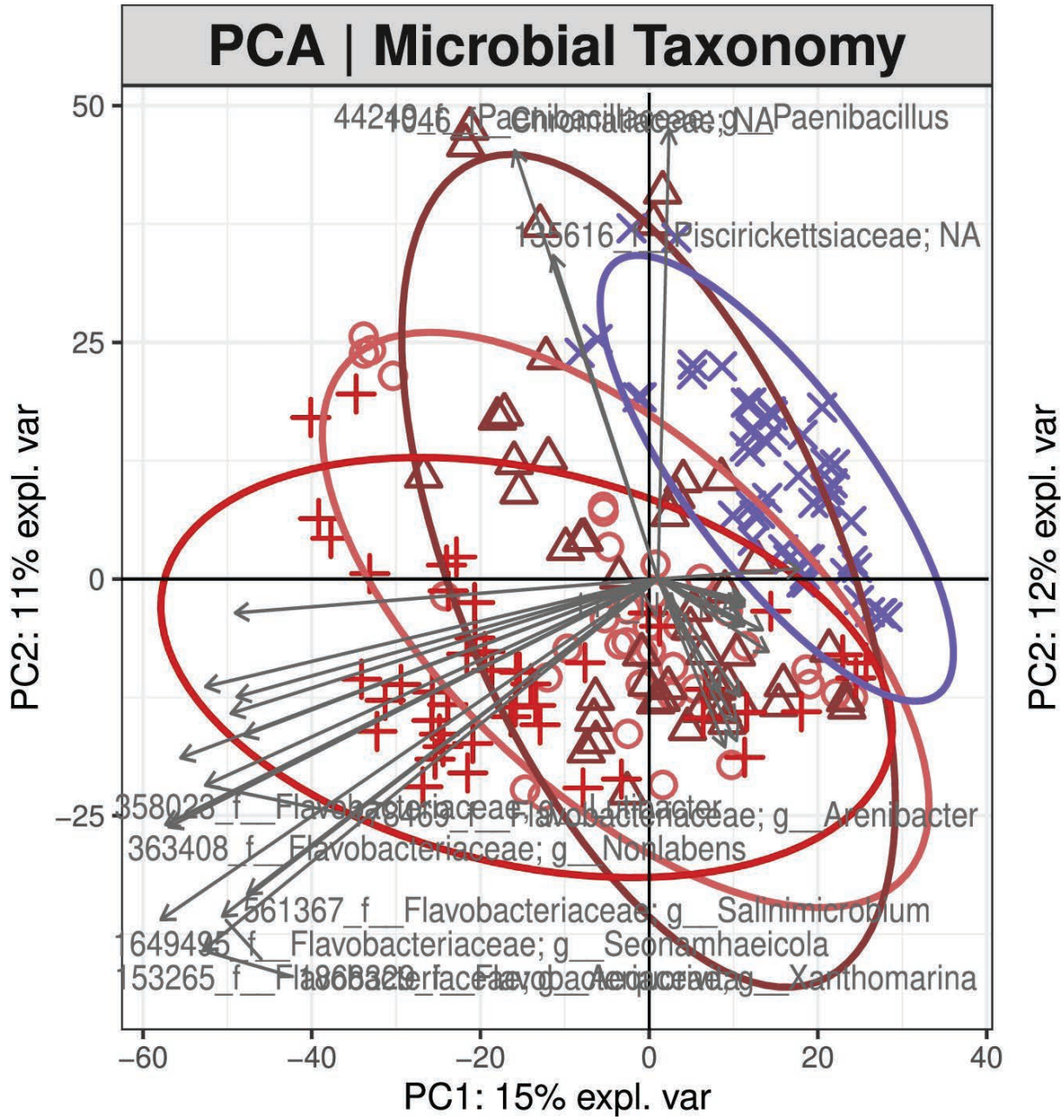

Figure S7: Biplots from the Principal Components Analysis (PCA) show the main clustering patterns of reef sites based on microbial taxonomic community composition. The plots highlight which microbial taxa are enriched in specific reef sites, coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively.

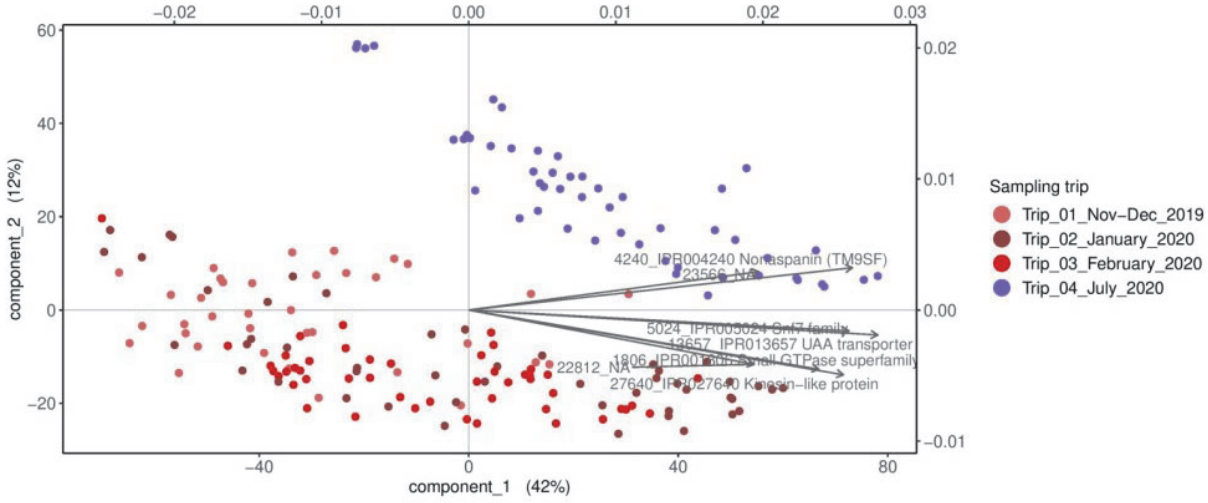

Figure S8: Biplots from the Principal Components Analysis (PCA) show the main clustering patterns of reef sites based on microbial functional community composition (i.e. GO terms). The plots highlight which microbial genes are enriched in specific reef sites, coloured in red or blue tones to denote trips that occurred during the austral summer (wet season) or austral winter (dry season), respectively.

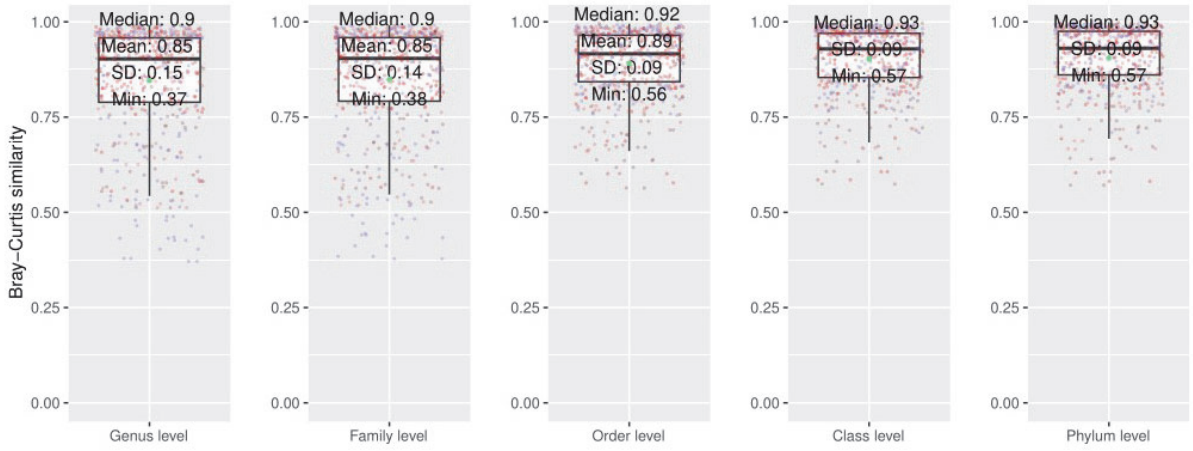

Figure S9: Bray-Curtis Similarity Index shows within-reef community similarity (0 = dissimilar; 1 = identical) for microbial taxonomy at genus, family, order, class, and phylum-level classifications. Data points are coloured coded to correspond the colouring scheme in Fig. 1 in the main text, and the green dot represents the mean Bray-Curtis similarity.

9 Appendix B – Supplementary Material for Chapter 3

PCA - Principal Components Analysis | What are the main clustering patterns across our samples?

Principal Components Analysis (PCA) was applied in an R package mixOmics[292] as an unsupervised approach to visualise the main clustering patterns between reef sites based on microbial community profiles. The number of optimal PCA components was determined using the tune.pca() function in mixOmics.

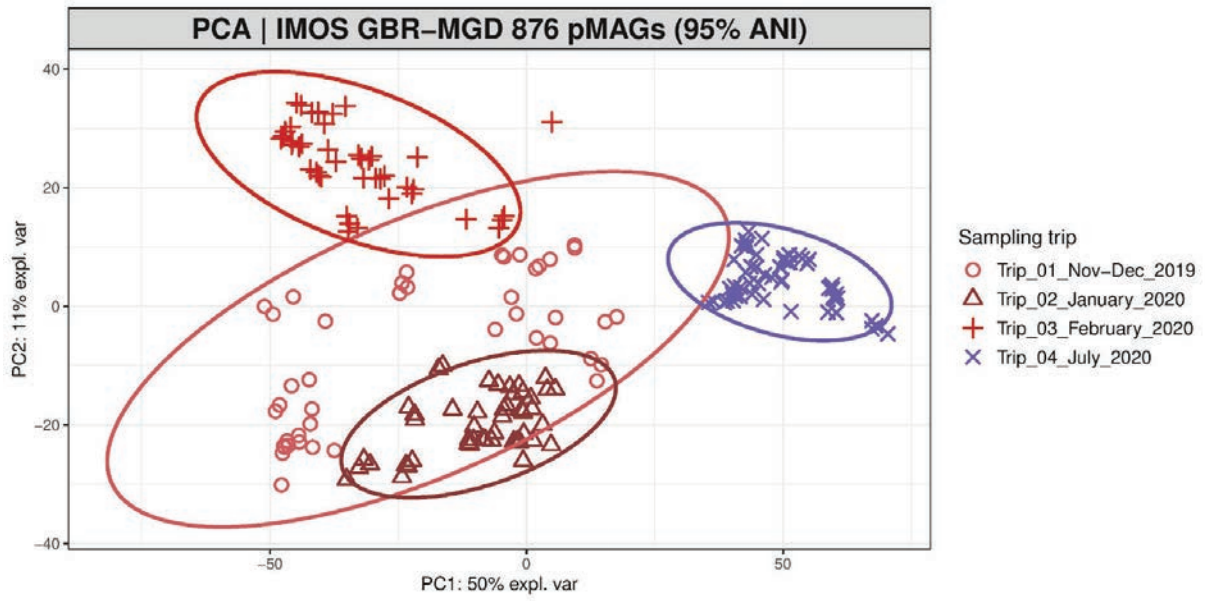

Figure S1. Main clustering patterns of seawater microbial communities, colored per sampling transect. The PCA ordination plots show clear differences between microbial communities sampled during the summer/wet season (red) and winter/dry season (blue), with 50% of variance being attributable to dimension 1. Samples collected in the peak of summer (Trip 3) additionally separate from early summer sampling (Trips 1 and 2) on PCA dimension 2.

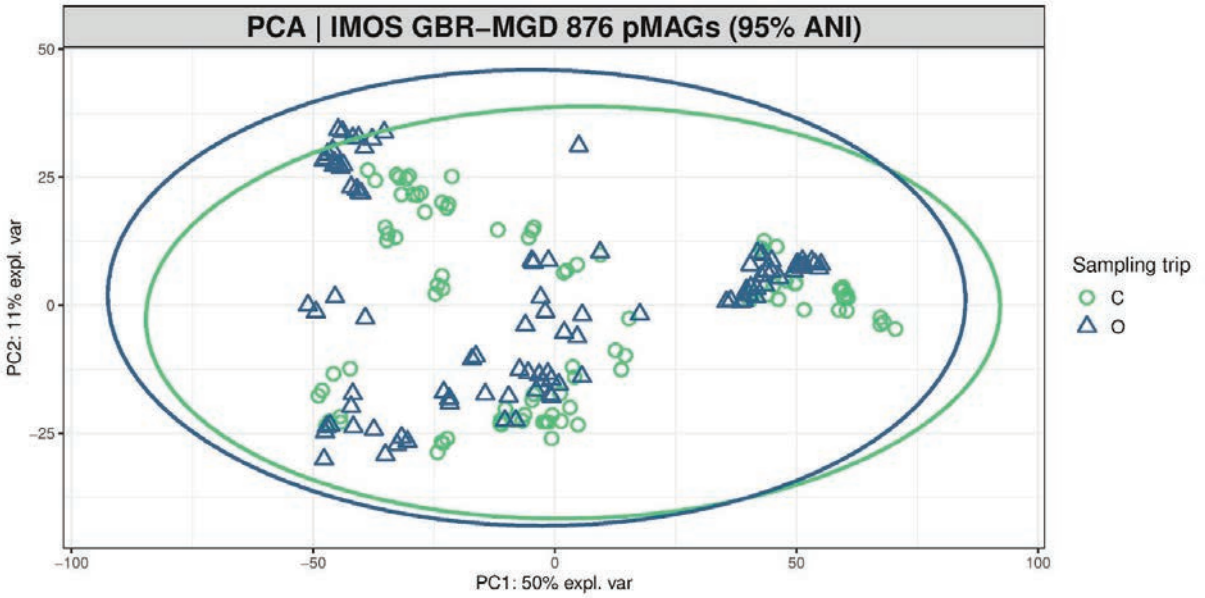

Figure S2. Clustering patterns of seawater microbial communities based on reef zoning. PCA ordination does not show clear clustering between No-Take Marine Reserves (C - closed to fishing, green) and fished reefs (O - open to fishing, blue).

(s)PLS-DA - (**S**parse) **P**artial **L**east **S**quares **D**iscriminant **A**nalysis | Can we discriminate between No-Take Marine Reserves (NTMRs) and fished reefs using a supervised approach?

PLS-DA

As PCA ordination shows that our sites cluster based on geographic proximity (i.e. sector) and time (i.e. sampling trip) (**Fig. S1**) and not based on reef zoning (**Fig. S2**) as our categorical outcome of interest, we then explored if sPLS-DA[291], as a supervised approach, will identify microbial indicators of No-Take Marine Reserves (NTMRs) vs fished reefs.

Tuning the number of components in PLS-DA

The perf() function evaluates the performance of PLS-DA - i.e., its ability to rightly classify 'new' samples into their category (NTMRs and fished zones) using repeated cross-validation. We initially choose a large number of components (here ncomp = 10) and assess the model as we gradually increase the number of components. Here, we used a 4-fold CV repeated 50 times.

The plot (**Fig. S3**) shows that the error rate keeps dropping as we increase the number of components, which may suggest strong batch effects in the data as this many dimensions would typically not be needed to discriminate only two categorical outcomes (NTMRs and fished reefs).

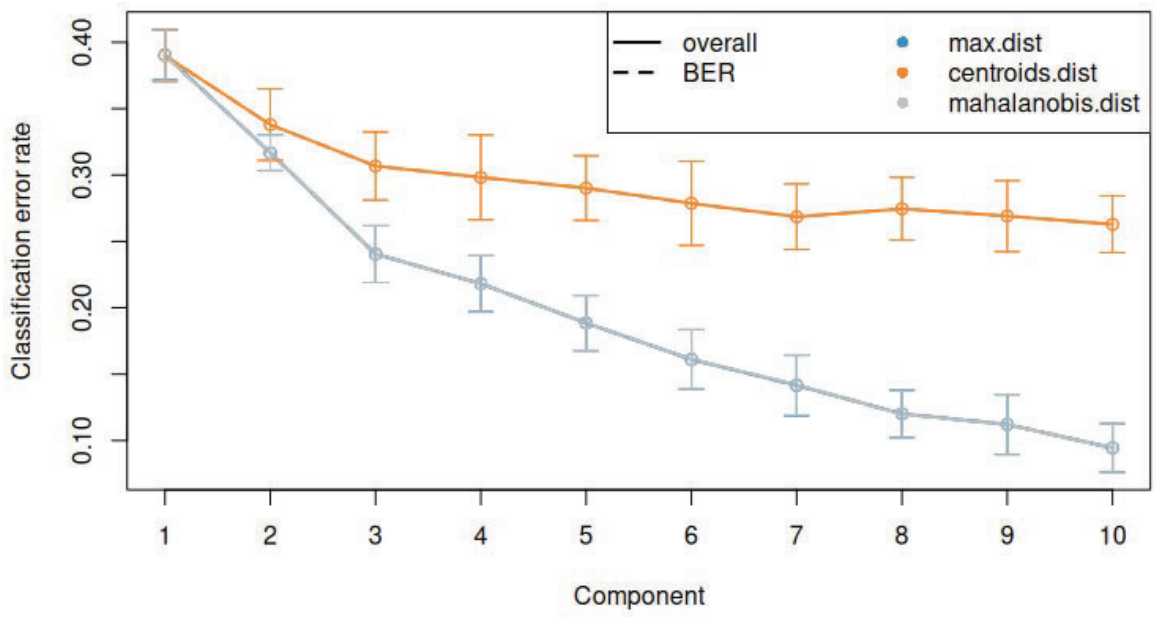

Figure S3. Tuning the number of components in PLS-DA on the IMOS GBR-MGD microbial data (876 pMAGs, de-replicated at 95% ANI). For each component, repeated cross-validation (50 ×4–fold CV) is used to evaluate the PLS-DA classification performance (overall and balanced error rate BER, and for each type of prediction distance: max.dist, centroids.dist and mahalanobis.dist) to discriminate between No-Take Marine Reserves (NTMRs) and fished reefs based on the seawater microbiomes. Bars show the standard deviation across the repeated folds.

In PLS-DA sample plots (**Fig. S4**), we can observe improved clustering according to reef protection status (**Fig. S4**; top), compared with PCA (**Fig. S2**). This is to be expected since PLS-DA is a supervised approach and includes the class information of each sample, and aims to discriminate between them. From the *plotIndiv()* function, we observe some discrimination between NTMRs and fished reefs mostly on component 1 (x-axis), however we can still see the trip effect (**Fig. S4**; bottom). The axis labels indicate the amount of variation explained per component, however, the interpretation of this amount is not as important as in PCA, as PLS-DA aims to maximise the covariance between components associated to X (predictor dataset, i.e. the 876 IMOS-MGD MAGs) and Y (categorical "response", i.e. no-take and take zones), rather than the variance of X (shown in PCA plots).

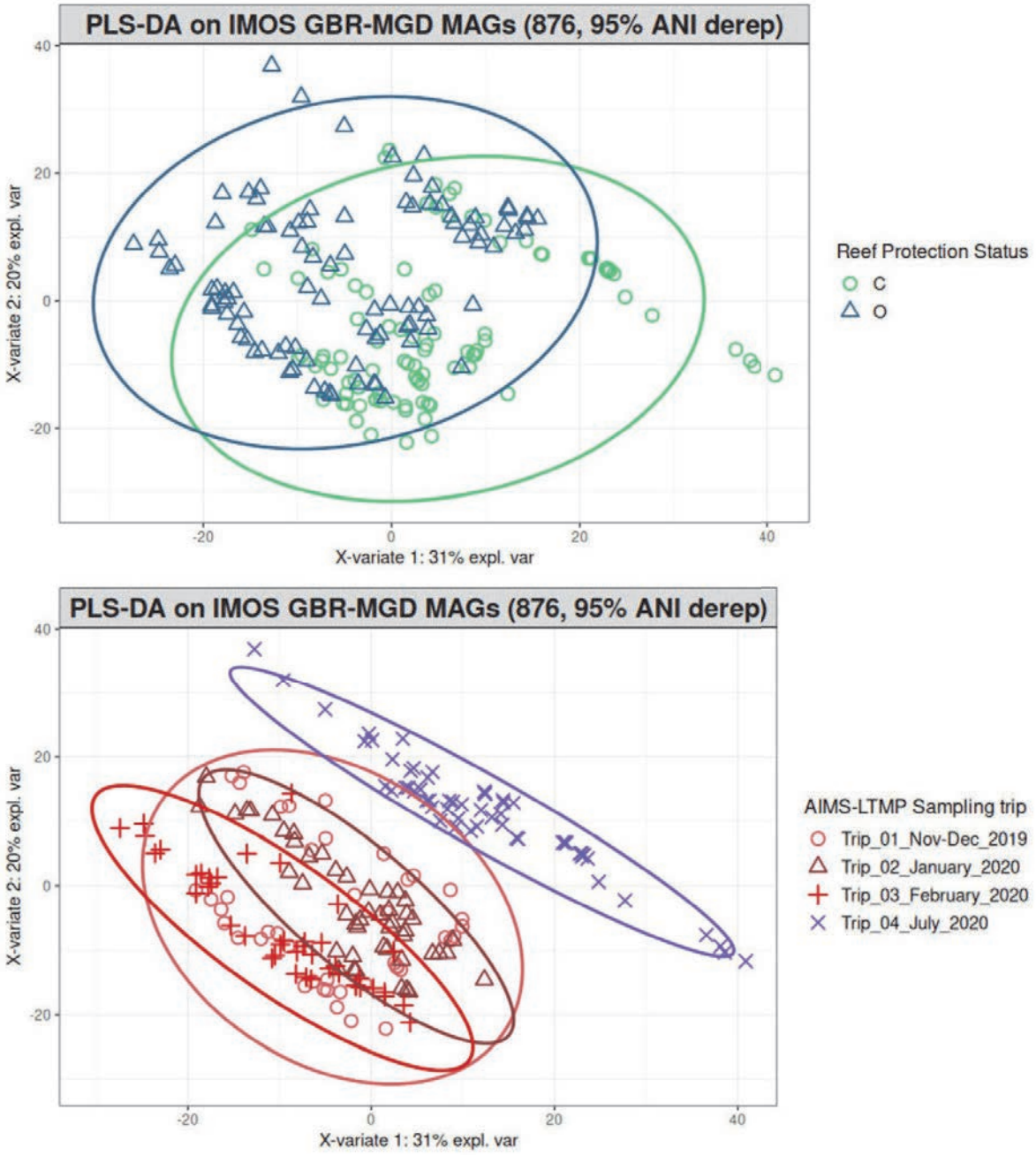

Figure S4. Sample plots from PLS-DA performed on the IMOS GBR-MGD microbial data (876 pMAGs, dereplicated at 95% ANI) as X, to discriminate reef zoning status as Y. Samples are projected into the space spanned by the first two components, coloured by their reef zoning status (above) or sampling trip (below). While we do observe separation of samples based on reef zoning (above), samples also cluster based on time of sampling and geographic proximity (below), suggesting batch effects due to confounding effects of space and time.

sPLS-DA - can we refine these clusters by selecting the most influential pMAGs to discriminate reef zoning?

As many of the pMAGs in X may be noisy or uninformative to discriminate between NTMRs and fished reefs, an sPLS-DA analysis (sparse variant) may help refine the sample clusters and select a small subset of variables relevant to discriminate each class.

Tuning the number of variables to select

We estimate the classification error rate with respect to the number of selected variables in the model with the function *tune.splsda()*. The tuning is being performed one component at a time inside the function and the optimal number of variables to select is automatically retrieved after each component run.

Previously, we determined the optimal number of components to be ncomp = 10 with PLS-DA. Here we set ncomp = 15 to further assess if this would be the case for a sparse model, and use 4-fold cross validation repeated 50 times. We first define a grid of keepX values, and we tested 34 keepX values in total: a fine grid (1-10) followed by a coarser sequence (20-250 in increments of 10).

In (**Fig. S5**), we display the mean classification error rate on each component, bearing in mind that each component is conditional on the previous components calculated with the optimal number of selected variables. The diamond in the figure below indicates the best keepX value to achieve the lowest error rate per component. This type of graph helps not only to choose the 'optimal' number of variables to select, but also to confirm the number of components ncomp. From the following code (tune.splsda.open.closed_IMOS_P$choice.ncomp$ncomp), we can assess that the optimal number of components was 10 according to a one-sided T–test. The numerical output from tune.splsda() (**Table S2**) globally shows that the classification error rate continues to decrease after the second component in sparse PLS-DA, yet since we are only discriminating between two categorical outcomes, retaining a small number (i.e. one or two) components is recommended to avoid overfitting. Figures S6 and S7 further confirm the spatiotemporal batch effects in the data.

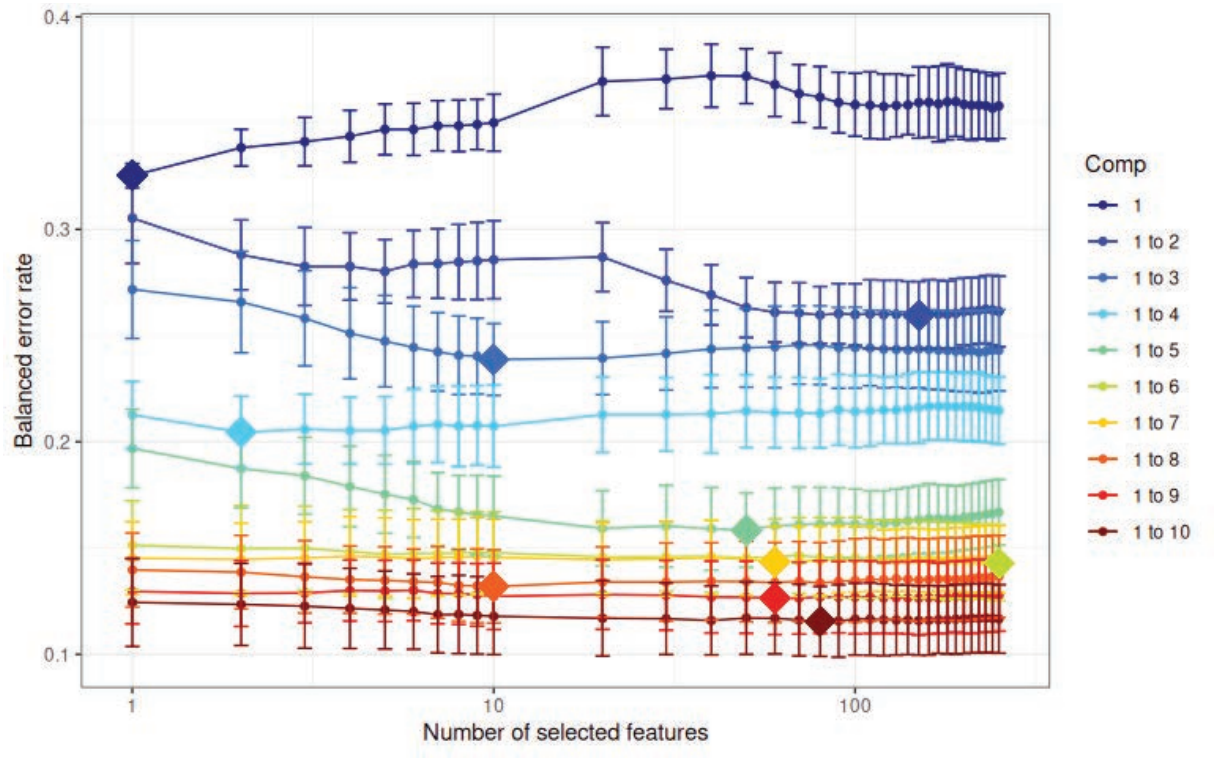

Figure S5. Tuning keepX for the sPLS-DA performed on the IMOS GBR-MGD pMAGs (876 genomes, drep at 95% ANI). Each coloured line represents the balanced error rate (y-axis) per component across all tested keepX values (x-axis) with the standard deviation based on the repeated cross-validation folds (4-fold x 50 repeats). The diamond indicates the optimal keepX value on a particular component which achieves the lowest classification error rate as determined with a one-sided t–test. As sPLS-DA is an iterative algorithm, values represented for a given component (e.g. comp 1 to 2) include the optimal keepX value chosen for the previous component (comp 1).

Table S2. Numerical output associated with Fig. S5, showing sPLS-DA tune() results: mean error rate for each component and each tested keepX value given the past (tuned) components.

comp1	comp2	comp3	comp4	comp5	comp6	comp7	comp8	comp9	comp10
0.33	0.31	0.27	0.21	0.20	0.15	0.15	0.14	0.13	0.12
0.34	0.29	0.27	0.20	0.19	0.15	0.14	0.14	0.13	0.12
0.34	0.28	0.26	0.21	0.18	0.15	0.15	0.14	0.13	0.12
0.34	0.28	0.25	0.21	0.18	0.15	0.15	0.14	0.13	0.12
0.35	0.28	0.25	0.21	0.18	0.15	0.15	0.13	0.13	0.12
0.35	0.28	0.24	0.21	0.17	0.15	0.14	0.13	0.13	0.12

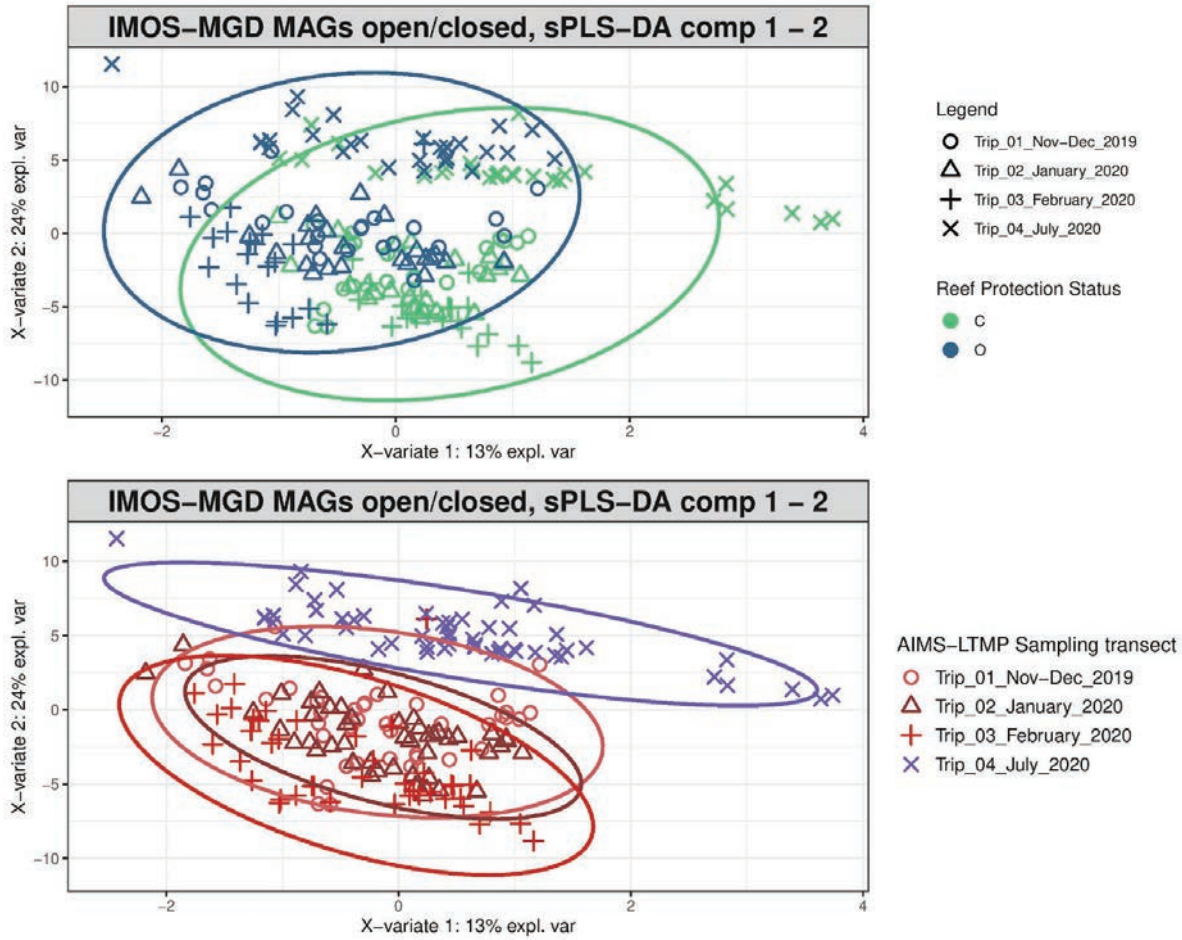

Figure S6. Sample plots from sPLS-DA performed on the IMOS GBR-MGD microbial data (876 pMAGs, dereplicated at 95% ANI) as X, to discriminate reef zoning status as Y. Samples are projected into the space spanned by the first two components. The plots represent 95% ellipse confidence intervals around each sample class: (above) NTMRs (in green) vs fished reefs (in blue); and (below) sampling trip. While we do see separation between NTMRs and fished reefs as our outcome of interest (above), we can also observe spatio-temporal batch effects (below).

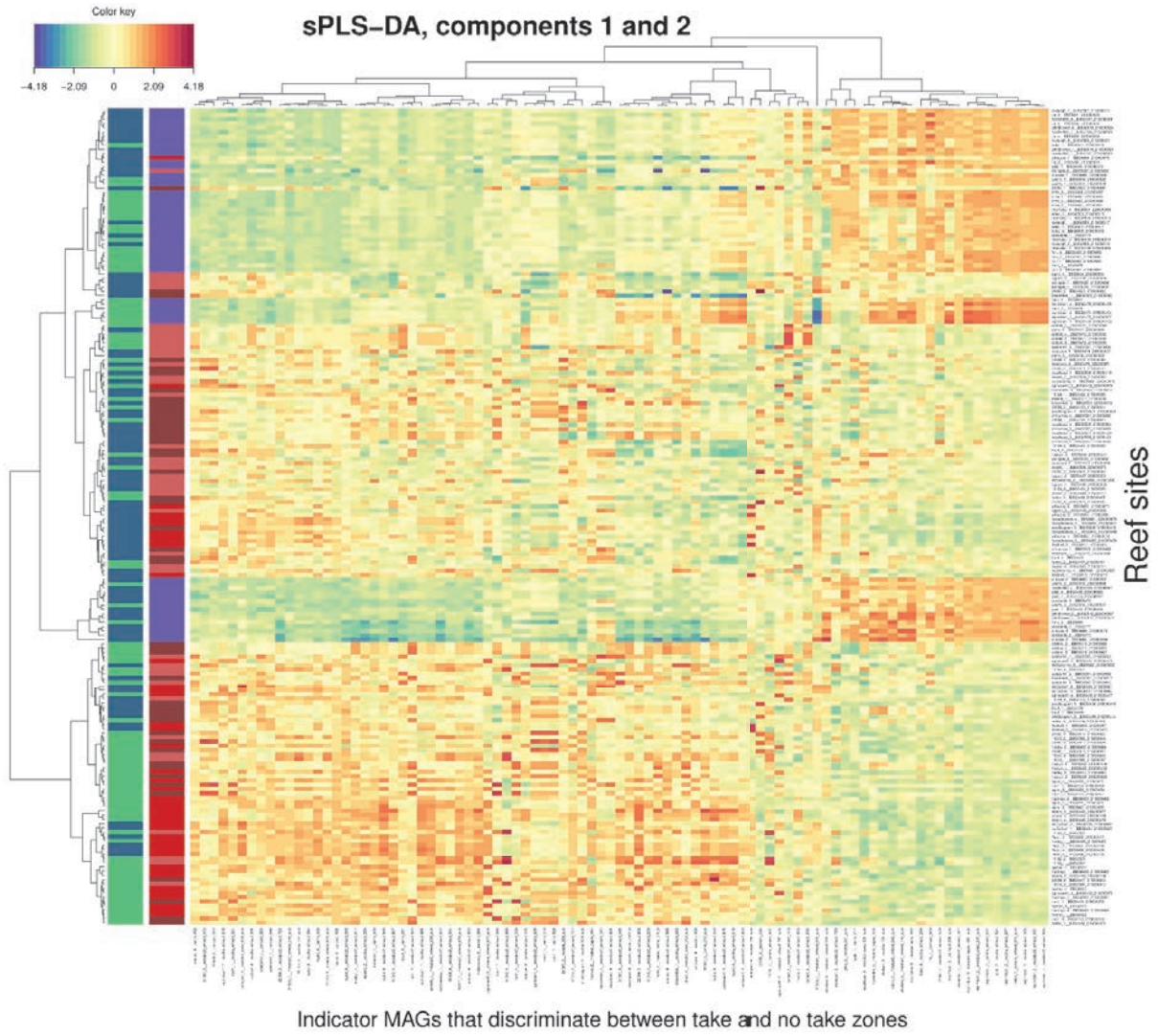

Figure S7. Clustered Image Map showing abundance patterns of reef zoning microbial indicators, as inferred from sPLS-DA performed on the IMOS GBR-MGD microbial data (876 pMAGs, de-replicated at 95% ANI) as X, to discriminate No-Take Marine Reserves (NTMRs) and fished reefs as categorical Y. A hierarchical clustering based on the pMAG enrichment values for the selected indicator microbes, with samples (48 reef sites x 4 replicates) in rows coloured according to their reef protection status (NTMRs in green, fished reefs in blue) and sampling trip (Austral summer/wet season samples in red tones, winter samples in blue). The heatmap is clustered using Euclidean distance with Complete agglomeration method. As previously observed, spatio-temporal patterns are stronger drivers than zoning and thus represent batch effects.

Multivariate **INT**egration **(MINT) sPLS-DA** | Discriminating between reefs that are open or closed to fishing (**sPLS-DA**), while accounting for sector-specific effects (**MINT**)

Tuning the number of dimensions

The perf() function is used to estimate the performance of the MINT-sPLS-DA model using Leave One Group Out Cross Validation (LOGOCV, i.e. by training MINT sPLS-DA on six out of seven sectors and validating the model performance on the left-out subset, hence seven times until each of the seven GBR sectors is left out once), and to choose the optimal number of components for our final model.

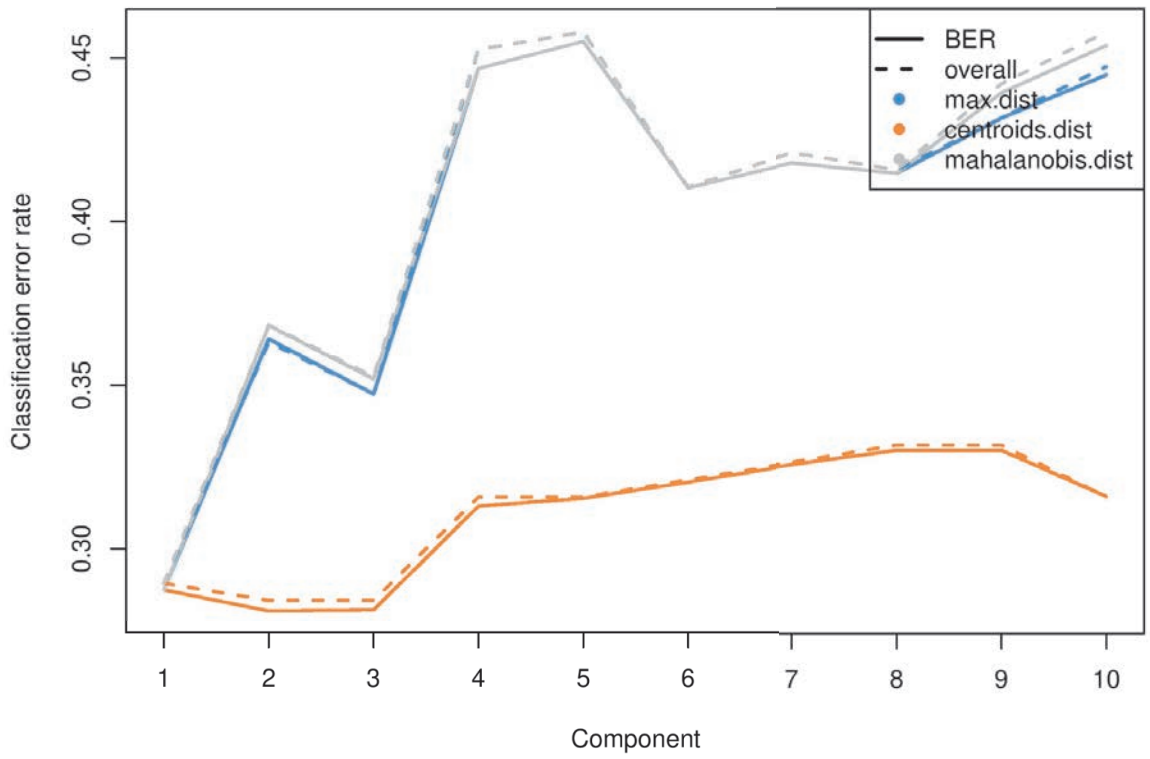

Fig S8. Choosing the number of components in mint.splsda using perf() with LOGOCV to discriminate between reefs that are open or closed to fishing, using a dataset of 876 IMOS GBR-MGD microbial genomes (pMAGs95%ANI). Classification error rates (overall and balanced - BER) are represented on the y-axis with respect to the number of components on the x-axis for each prediction distance. Overall and balanced error

rates show largely the same trend as the design is balanced (i.e. the same number of NTMR and fished reefs in each GBR sector). The plot shows that the error rate reaches a minimum (~29%) with two or three dimensions with the centroids prediction distance. We therefore retained 2 PCs in downstream analysis.

Table S3. Numerical output associated with Fig. S1. Here, we show overall MINT sPLS-DA error rates when discriminating between reefs that are open or closed to fishing (Y), using 876 IMOS GBR-MGD pMAGs95%ANI as X, using 3 prediction distances (max.dist, centroids.dist, mahalanobis.dist), and across the seven GBR sectors. Sector-specific model accuracies were expressed as 1 – error (with centroids dist), and for each of the 10 tested MINT sPLS-DA components, we also show MINT sPLS-DA classification accuracy averaged across 7 GBR sectors.

MINT sPLS-DA comp	Study (GBR sector)	Max dist	Centroids dist	Mahalanobis dist	Accuracy (1 – error with centroids dist)	Average accuracy
comp1	01_Cape_Grenville	0.42	0.42	0.42	0.58	
comp1	02_Princess_Charlotte_bay	0.20	0.20	0.20	0.80	
comp1	03_Cairns	0.35	0.35	0.35	0.65	
comp1	04_Innisfail	0.07	0.07	0.07	0.93	0.71
comp1	05_Townsville	0.30	0.30	0.30	0.70	
comp1	06_Swains	0.35	0.35	0.35	0.65	
comp1	07_Capricorn_Bunker	0.32	0.32	0.32	0.68	
comp2	01_Cape_Grenville	0.58	0.42	0.54	0.58	
comp2	02_Princess_Charlotte_bay	0.67	0.20	0.73	0.80	
comp2	03_Cairns	0.10	0.25	0.10	0.75	
comp2	04_Innisfail	0.22	0.15	0.22	0.85	0.71
comp2	05_Townsville	0.32	0.27	0.32	0.73	
comp2	06_Swains	0.45	0.40	0.50	0.60	
comp2	07_Capricorn_Bunker	0.36	0.32	0.36	0.68	
comp3	01_Cape_Grenville	0.63	0.38	0.63	0.63	
comp3	02_Princess_Charlotte_bay	0.60	0.27	0.60	0.73	
comp3	03_Cairns	0.25	0.25	0.25	0.75	
comp3	04_Innisfail	0.30	0.15	0.30	0.85	0.71
comp3	05_Townsville	0.30	0.29	0.30	0.71	
comp3	06_Swains	0.35	0.40	0.35	0.60	
comp3	07_Capricorn_Bunker	0.18	0.29	0.21	0.71	
comp4	01_Cape_Grenville	0.63	0.42	0.63	0.58	
comp4	02_Princess_Charlotte_bay	0.93	0.27	0.93	0.73	
comp4	03_Cairns	0.25	0.30	0.25	0.70	
comp4	04_Innisfail	0.48	0.22	0.48	0.78	0.68
comp4	05_Townsville	0.39	0.29	0.39	0.71	
comp4	06_Swains	0.45	0.40	0.45	0.60	
comp4	07_Capricorn_Bunker	0.29	0.36	0.29	0.64	
comp5	01_Cape_Grenville	0.71	0.42	0.71	0.58	
comp5	02_Princess_Charlotte_bay	1.00	0.33	1.00	0.67	
comp5	03_Cairns	0.30	0.30	0.30	0.70	0.67
comp5	04_Innisfail	0.48	0.19	0.48	0.81	
comp5	05_Townsville	0.38	0.29	0.38	0.71	
comp5	06_Swains	0.45	0.40	0.45	0.60	

comp5	07_Capricorn_Bunker	0.21	0.36	0.21	0.64	
comp6	01_Cape_Grenville	0.63	0.42	0.63	0.58	0.67
comp6	02_Princess_Charlotte_bay	0.80	0.40	0.80	0.60	
comp6	03_Cairns	0.30	0.30	0.30	0.70	
comp6	04_Innisfail	0.41	0.19	0.41	0.81	
comp6	05_Townsville	0.34	0.29	0.34	0.71	
comp6	06_Swains	0.40	0.40	0.40	0.60	
comp6	07_Capricorn_Bunker	0.25	0.36	0.25	0.64	
comp7	01_Cape_Grenville	0.63	0.46	0.63	0.54	0.66
comp7	02_Princess_Charlotte_bay	0.87	0.40	0.87	0.60	
comp7	03_Cairns	0.20	0.30	0.20	0.70	
comp7	04_Innisfail	0.41	0.19	0.41	0.81	
comp7	05_Townsville	0.34	0.29	0.34	0.71	
comp7	06_Swains	0.45	0.40	0.45	0.60	
comp7	07_Capricorn_Bunker	0.32	0.36	0.32	0.64	
comp8	01_Cape_Grenville	0.63	0.46	0.63	0.54	0.65
comp8	02_Princess_Charlotte_bay	0.93	0.40	0.93	0.60	
comp8	03_Cairns	0.20	0.30	0.20	0.70	
comp8	04_Innisfail	0.44	0.19	0.44	0.81	
comp8	05_Townsville	0.32	0.29	0.32	0.71	
comp8	06_Swains	0.35	0.45	0.35	0.55	
comp8	07_Capricorn_Bunker	0.32	0.36	0.32	0.64	
comp9	01_Cape_Grenville	0.67	0.46	0.67	0.54	0.65
comp9	02_Princess_Charlotte_bay	0.93	0.40	0.93	0.60	
comp9	03_Cairns	0.20	0.30	0.20	0.70	
comp9	04_Innisfail	0.44	0.19	0.52	0.81	
comp9	05_Townsville	0.30	0.29	0.30	0.71	
comp9	06_Swains	0.35	0.45	0.40	0.55	
comp9	07_Capricorn_Bunker	0.43	0.36	0.39	0.64	
comp10	01_Cape_Grenville	0.75	0.46	0.75	0.54	0.67
comp10	02_Princess_Charlotte_bay	0.80	0.33	0.80	0.67	
comp10	03_Cairns	0.20	0.30	0.20	0.70	
comp10	04_Innisfail	0.56	0.19	0.56	0.81	
comp10	05_Townsville	0.29	0.29	0.29	0.71	
comp10	06_Swains	0.50	0.45	0.55	0.55	
comp10	07_Capricorn_Bunker	0.36	0.29	0.39	0.71	

Table S4. MINT sPLS-DA - error rate (centroids distance) across GBR sectors, and separately for C (reefs closed to fishing) and O (open to fishing).

	Study	comp1	comp2	comp3	comp4	comp5	comp6	comp7	comp8	comp9	comp10
C	01_Cape_Grenville	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
O	01_Cape_Grenville	0.25	0.25	0.17	0.25	0.25	0.25	0.33	0.33	0.33	0.33
C	02_Princess_Charlotte_bay	0.00	0.00	0.13	0.13	0.25	0.38	0.38	0.38	0.38	0.25

O	02_Princess_Charlotte_bay	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
C	03_Cairns	0.25	0.00	0.00	0.13	0.13	0.13	0.13	0.13	0.13	0.13
O	03_Cairns	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
C	04_Innisfail	0.07	0.13	0.13	0.20	0.13	0.13	0.13	0.13	0.13	0.13
O	04_Innisfail	0.08	0.17	0.17	0.25	0.25	0.25	0.25	0.25	0.25	0.25
C	05_Townsville	0.39	0.32	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
O	05_Townsville	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
C	06_Swains	0.25	0.38	0.38	0.38	0.50	0.50	0.50	0.50	0.50	0.50
O	06_Swains	0.42	0.42	0.42	0.42	0.33	0.33	0.33	0.42	0.42	0.42
C	07_Capricorn_Bunker	0.31	0.31	0.25	0.38	0.38	0.38	0.38	0.38	0.38	0.25
O	07_Capricorn_Bunker	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Average error [C]		0.27	0.25	0.26	0.31	0.33	0.35	0.35	0.35	0.35	0.31
Average error [O]		0.31	0.32	0.31	0.33	0.32	0.32	0.33	0.34	0.34	0.34
Average accuracy [C]		0.73	0.75	0.74	0.69	0.67	0.65	0.65	0.65	0.65	0.69
Average accuracy [O]		0.69	0.68	0.69	0.67	0.68	0.68	0.67	0.66	0.66	0.66

Tuning the number of features per dimension

We can choose the keepX parameter using the tune() function for a MINT object. The function performs LOGOCV for different values of test.keepX (we specified test.keepX = seq(10, 300, 10)) provided on each component (we tested 5 components), and no repeat argument is needed. Based on the mean classification error rate (overall error rate or BER) and a centroids distance, we output the optimal number of variables keepX to be included in the final model.

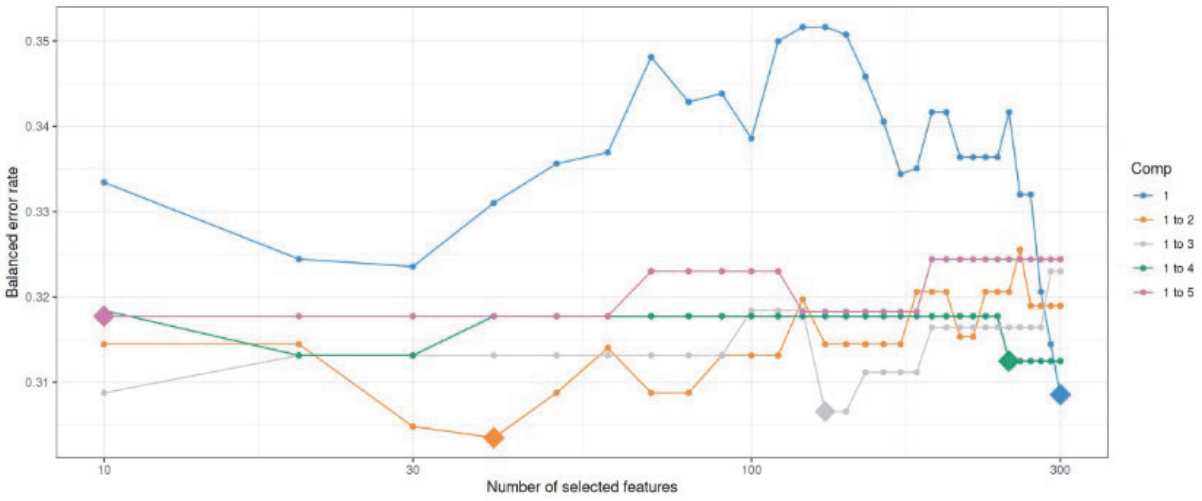

Figure S9. Tuning plot of the MINT sPLS-DA models with up to 5 components, testing a grid value of 10 to 300 indicators (with sequential increases of 10). Diamonds represent the optimal number of features on a given component. Balanced error rate found on the vertical axis and is the metric to be minimised.

Table S5. Numerical output associated with Fig S9, also showing sector-specific MINT sPLS-DA classification errors and average error/accuracy across sectors.

GBR_sector	comp1	comp2	comp3	comp4	comp5
CA	0.33	0.27	0.21	0.27	0.27
CB	0.32	0.32	0.32	0.35	0.35
CG	0.42	0.42	0.46	0.42	0.46
IN	0.18	0.23	0.23	0.23	0.23
PC	0.21	0.21	0.21	0.21	0.21
SW	0.44	0.4	0.44	0.44	0.44
TO	0.29	0.29	0.29	0.29	0.29
Average Error	0.31	0.31	0.31	0.32	0.32
Accuracy (1 – average error)	0.69	0.69	0.69	0.68	0.68

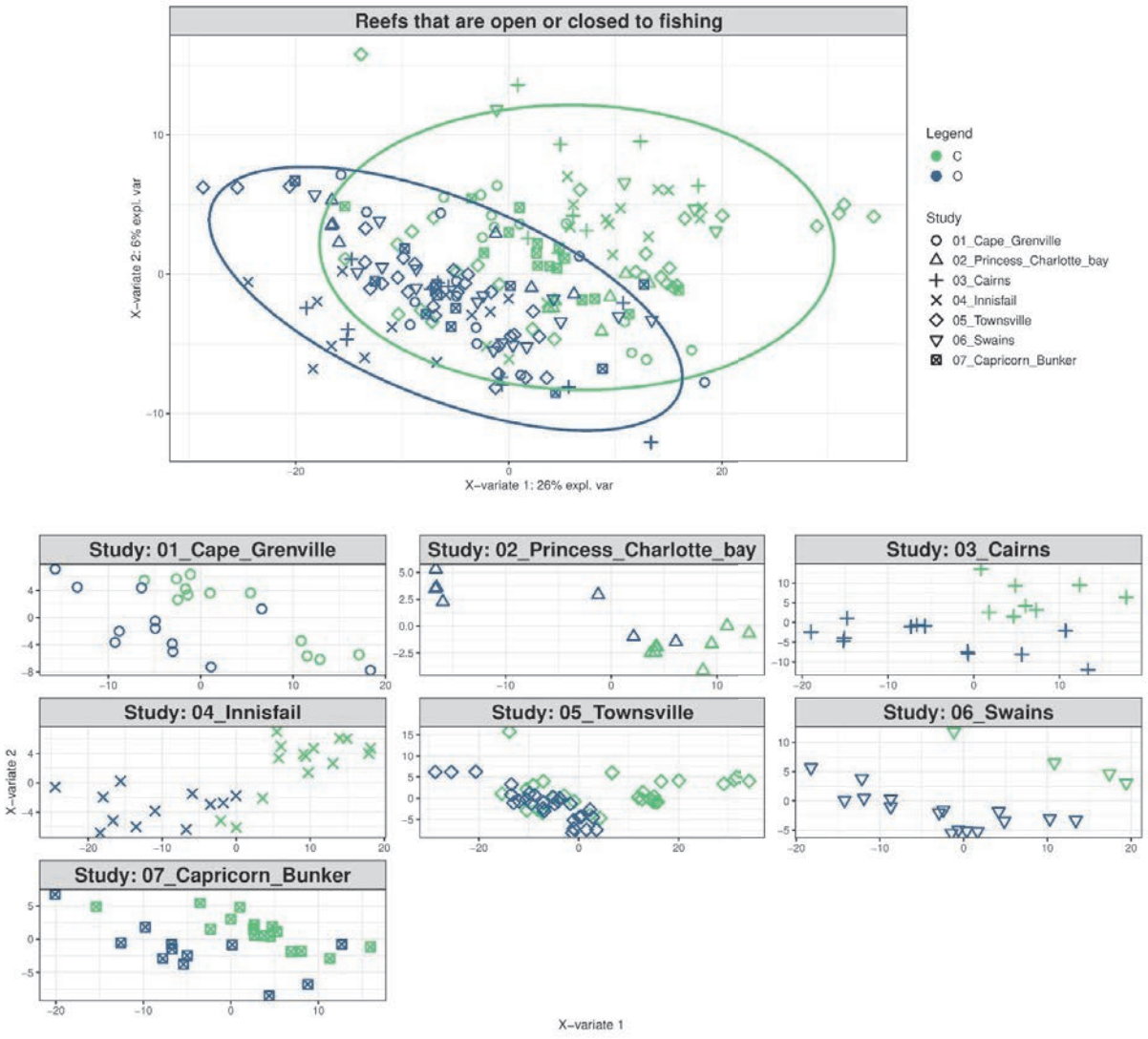

Figure S10. Sample plots from the MINT sPLS-DA performed on the 876 IMOS GBR-MGD seawater pMAGs$_{95\%ANI}$, aiming to find discriminatory microbes between reefs that are open or closed to fishing. Samples (48 reef sites x 4 replicates) are projected into the space spanned by the first two components. Reef sites are coloured by their protection level (open or closed to fishing) and symbols indicate the membership of reef sites to their corresponding LTMP trip/transect. (**top**) Global components from the model with 95% ellipse confidence intervals around each sample class. (**bottom**) Partial components per study show a good agreement across GBR sectors. Component 1 discriminates between reefs that are open or closed to fishing.

Performance of the final MINT sPLS-DA model

Use of the auroc() function will yield a visualisation of classification performance when undergoing the LOGOCV procedure from above. The interpretation of this output may not be particularly insightful in relation to the performance evaluation of mixOmics methods, but can complement the

statistical analysis. For example, the MINT sPLS-DA classification of fished vs. NTMR sites had ~73 % accuracy in classifying samples in their corresponding zoning category.

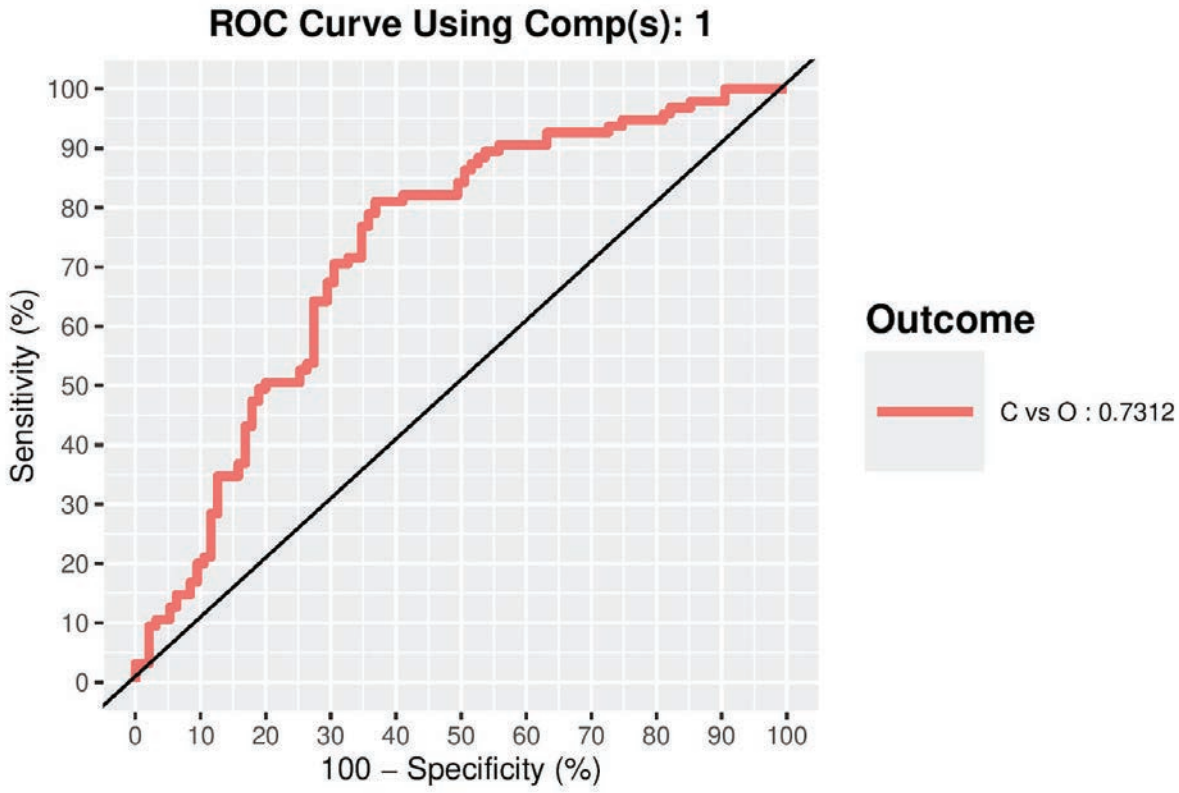

Fig. S11. ROC curve and AUC from the MINT sPLS-DA performed on the IMOS MGD MAGs (876 pMAGs95%ANI) for global component 1 for the fished vs. NTMRs reefs comparison. Numerical outputs include the AUC (0.7312) and a Wilcoxon test p–value (p = 3.69 × 10−8) for fished vs. NTMRs reefs class comparison that are performed per component.

Multivariate **INT**egration **(MINT) sPLS** | Correlating microbial and environmental data (**sPLS**), while accounting for sector-specific effects (**MINT**)

Table S6. Mean ± standard deviation (SD) of MINT sPLS partial correlation scores between microbial indicators of zoning and environmental variables. This table complements the heatmap in the main text (**Fig 3.3B**), with positive values (red) indicating positive associations, and negative values (blue) indicate negative associations. NTMRs = No-Take Marine Reserves.

Reef Variable	Fished Reef Indicators (n = 114)	NTMR Indicators (n = 236)
POC_µM	0.08 ± 0.13	-0.12 ± 0.17
SEAWATER_TEMPERATURE_2.5m_RV	0.06 ± 0.13	-0.1 ± 0.16
Foliose_non_Acropora	0.08 ± 0.1	-0.12 ± 0.13
Carnivore	0.01 ± 0.11	-0.04 ± 0.14
TDN_µM	0.01 ± 0.11	-0.03 ± 0.13

Sand	0.01 ± 0.13	-0.03 ± 0.15
Turf_algae	0.17 ± 0.12	-0.24 ± 0.15
PP_µM	0.14 ± 0.11	-0.2 ± 0.14
SALINITY_2.5m_RV	0.2 ± 0.13	-0.28 ± 0.17
Phaeophytin_A_µg_L	0.16 ± 0.08	-0.22 ± 0.1
Massive_non_Acropora	-0.11 ± 0.09	0.12 ± 0.1
Piscivore	-0.11 ± 0.07	0.13 ± 0.07
FLUORESCENCE_2.5m_RV	-0.06 ± 0.07	0.06 ± 0.08
NO2_µM	0.14 ± 0.06	-0.18 ± 0.05
Lobate_Soft_Coral	0.02 ± 0.07	-0.01 ± 0.08
PO4_µM	-0.01 ± 0.11	0.04 ± 0.13
Coralline_algae	-0.14 ± 0.17	0.22 ± 0.22
Tabulate_Acropora	-0.13 ± 0.14	0.19 ± 0.18
Digitate_Acropora	-0.12 ± 0.16	0.18 ± 0.2
Herbivore	-0.15 ± 0.08	0.21 ± 0.09
Detritivore	-0.15 ± 0.12	0.21 ± 0.15
TDP_µM	-0.17 ± 0.14	0.24 ± 0.17
Submassive_non_Acropora	-0.24 ± 0.13	0.33 ± 0.16
Encrusting_Acropora	-0.21 ± 0.1	0.28 ± 0.12
Millepora	-0.2 ± 0.11	0.27 ± 0.14

Genera**l**ised **M**ixed **M**odel**s** (**GLMMs**) - do NTMRs and fished reefs differ in benthic cover (hard coral, algae), fish biomass, and densities of herbivorous fish?

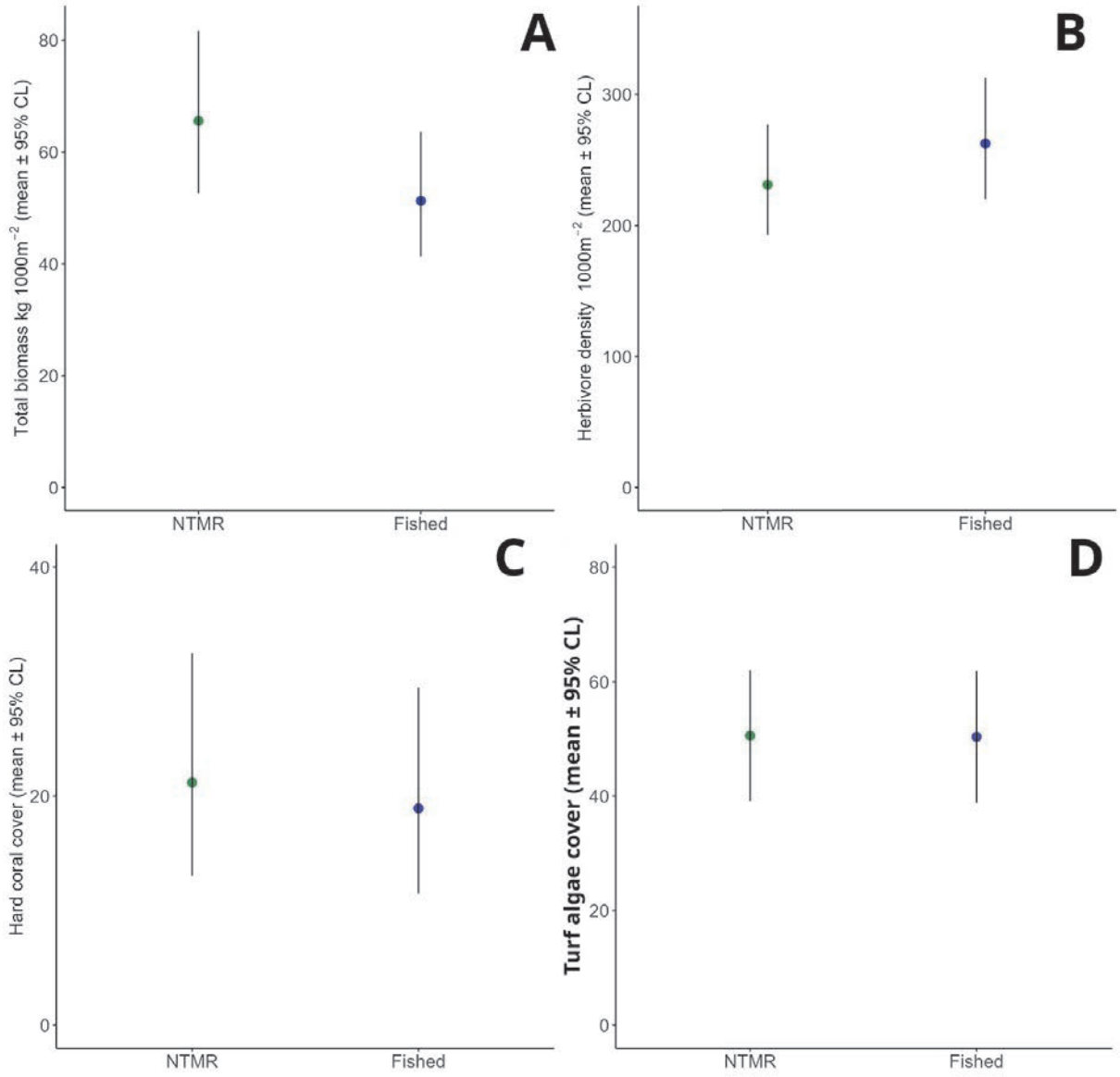

Figure S12. Comparison of (A) total fish biomass, (B) herbivore density, (C) hard coral, and (D) turf algae cover between No-Take Marine Reserves (NTMRs) and fished reef zones, using Generalised Mixed Models (GLMMs). Bars show estimated marginal means (± SE) from a Gamma GLMM, with closed zones (NTMRs) supporting 24.6% higher biomass than open zones (O) ($p = 0.014$). Differences in other metrics (B-D) were insignificant. Random effects accounted for spatial hierarchy (sector, shelf, reef/site/transect); error distributions used Gamma-log link.

Total fish biomass

Raw data

Table S7. Summary statistics (descriptive) for overall fish biomass.

Status	Mean (g)	Median (g)	SD	n	SE
Closed	100,915	61,616	129,789	345	6,988
Open	71,652	51,186	116,017	360	6,115

GLMM Results (Gamma Distribution)

Running the model

biom.tmb <- glmmTMB(

biomass ~ OPENORCLOSED_AFTER2004+(1|A_SECTOR)+(1|SHELF)+(1|REEF_NAME/SITE_NO/TRANSECT_NO),

data = tot.biom,

family = Gamma(link = "log")

)

summary(biom.tmb)

Model Specification

Family: Gamma (log link)

Formula:

biomass ~ OPENORCLOSED_AFTER2004 + (1 | A_SECTOR) + (1 | SHELF) + (1 | REEF_NAME/SITE_NO/TRANSECT_NO)

Dataset: total fish biomass

Observations: 705

Model Fit Statistics

(Note: AIC/BIC not available)

Table S8. Random Effects Variance.

Group	Variance	Std. Deviation
A_SECTOR	0.0497	0.223
SHELF	2.07×10^{-7}	~0
TRANSECT_NO:SITE_NO:REEF_NAME	0.324	0.569
SITE_NO:REEF_NAME	0.154	0.392
REEF_NAME	0.0412	0.203

Groups:

- A_SECTOR: 7 levels
- SHELF: 2 levels
- REEF_NAME: 47 levels

Dispersion parameter (σ^2): 3.64×10^{-8}

Table S9. Fixed Effects

Term	Estimate	Std. Error	z-value	p-value
(Intercept)	11.091	0.112	98.87	< 0.001 ***
OPENORCLOSED_AFTER2004O	-0.246	0.100	-2.47	0.014 *

Key Interpretation

Significant effect of protection status:

- Open (fished) areas had **24.6% lower biomass** (95% CI: 4.4%-44.8%) than NTMRs (*z* = -2.47, *p* = 0.014).
- Most variance explained by:
  - Transect-level effects (56.9% SD).
  - Site-level effects (39.2% SD).

Herbivorous Fish Density

Raw data

Table S10. Summary statistics (descriptive) for herbivorous fish densities.

Protection Status	Mean density	Median density	SD	n	SE
Closed (NTMR)	259	244	141	345	7.60
Open (Fished)	294	266	157	360	8.28

GLMM Results (Negative Binomial)

Run a glmm using library(glmm.tmb)

herb.tmb <- glmmTMB(

Density ~ OPENORCLOSED_AFTER2004+(1|A_SECTOR)+(1|SHELF)+(1|REEF_NAME/SITE_NO/TRANSECT_NO),

data = herbs,

family = poisson(link = "log")

)

Looking at the results now:

summary(herb.tmb)

Model Specification

Family: Negative Binomial (log link)

Formula:

Density ~ OPENORCLOSED_AFTER2004 + (1 | A_SECTOR) + (1 | SHELF) + (1 | REEF_NAME/SITE_NO/TRANSECT_NO)

Dataset: herbivorous fish (density)

Observations: 705

Table S11. Model Fit Statistics

Statistic	Value
AIC	8659.4
BIC	8695.9
Log-Likelihood	-4321.7
Deviance	8643.4

Table S12. Random Effects Variance

Group	Variance	Std. Deviation
A_SECTOR	2.125×10^{-2}	0.146
SHELF	3.510×10^{-14}	~0
TRANSECT_NO:SITE_NO:REEF_NAME	3.192×10^{-9}	~0
SITE_NO:REEF_NAME	7.097×10^{-2}	0.266
REEF_NAME	7.796×10^{-2}	0.279

Groups:

- A_SECTOR: 7 levels
- SHELF: 2 levels
- REEF_NAME: 47 levels

Table S13. Fixed Effects

Term	Estimate	Std. Error	z-value	p-value
(Intercept)	5.443	0.092	58.95	< 0.001 ***
OPENORCLOSED_AFTER2004	0.127	0.098	1.29	0.196

Dispersion parameter: 6.9

Key Interpretation

- No significant effect of protection status (OPENORCLOSED_AFTER2004O) on herbivore density (*z* = 1.29, *p* = 0.196).
- Most variance explained by reef-level random effects (REEF_NAME and SITE_NO:REEF_NAME).

Hard Coral Cover

Raw data

Table S14. Summary statistics (descriptive) for hard coral cover.

Protection Status	Mean Cover (%)	Median (%)	SD	n	SE
Closed (NTMR)	23.7	22.0	14.6	345	0.79
Open (Fished)	21.2	17.2	15.4	360	0.81

GLMM Results (Binomial Distribution)

---

hc.tmb <- glmmTMB(cbind(n.points,total.points-.points)~OPENORCLOSED_AFTER2004 + (1|A_SECTOR)+(1|SHELF)+(1|REEF_NAME/SITE_NO/TRANSECT_NO),

family='binomial',

data=tot.hc)

summary(hc.tmb)

Model Specification:

Family: Binomial (logit link)

Formula:

cbind(n.points, total.points - n.points) ~ OPENORCLOSED_AFTER2004 + (1 | A_SECTOR) + (1 | SHELF) + (1 | REEF_NAME/SITE_NO/TRANSECT_NO)

Dataset: Hard coral cover

Observations: 705

Table S15. Model Fit Statistics

Statistic	Value
AIC	6467.1
BIC	6499.0
Log-Likelihood	-3226.5
Deviance	6453.1

Df.resid	698

Table S16. Random Effects Variance

Groups	Variance	Std. Dev.
SHELF	0.0869	0.295
REEF_NAME	0.1488	0.386
SITE_NO:REEF_NAME	0.0739	0.272
TRANSECT_NO:SITE_NO:REEF_NAME	0.1584	0.398
A_SECTOR	0.0313	0.177

Groups:

- A_SECTOR, 7 levels;
- SHELF, 2 levels;
- TRANSECT_NO:SITE_NO:REEF_NAME, 705 levels;
- SITE_NO:REEF_NAME, 141 levels;
- REEF_NAME, 47 levels;

Table S17. Fixed Effects

Term	Estimate	Std. Error	z-value	p-value
(Intercept)	0.0222	0.238	0.093	0.926
OPENORCLOSED_AFTER2004O	-0.0085	0.129	-0.066	0.947

Estimated Marginal Means (Probability Scale)

Status	Probability	SE	95% CI
Closed	0.506	0.060	0.391 - 0.620
Open	0.503	0.060	0.389 - 0.618

Contrasts

Comparison	Estimate	SE	z-value	p-value
Closed - Open	0.00213	0.0323	0.066	0.947

Key Findings

- **No significant difference** in hard coral cover between NTMR and fished areas (z = -0.066, p = 0.947).
- Model explains moderate variation through reef- and site-level random effects.
- Estimated probabilities nearly identical (50.6% NTMR vs 50.3% fished).

Turf algae

Raw data

Table S18. Summary statistics (descriptive) for turf algae cover.

Protection Status	Mean Cover (%)	Median (%)	SD	n	SE
Closed (NTMR)	50.8	51.8	17.7	345	0.95
Open (Fished)	52.8	53.0	14.2	360	0.75

GLMM Results (Binomial Distribution)

```
hc.tmb <- glmmTMB(cbind(n.points,total.points-.points)~OPENORCLOSED_AFTER2004 + (1|A_SECTOR)+(1|SHELF)+(1|REEF_NAME/SITE_NO/TRANSECT_NO),
family='binomial',
data=ta)
summary(hc.tmb)
```

Model specification

- **Family**: Binomial (logit link)
- **Formula**: cbind(n.points, total.points - n.points) ~ OPENORCLOSED_AFTER2004 + (1 | A_SECTOR) + (1 | SHELF) + (1 | REEF_NAME/SITE_NO/TRANSECT_NO)
- **Dataset**: turf algae

- **Observations**: 705

Table S19. Model Fit Statistics

Statistic	Value
AIC:	6467.1
BIC:	6499.0
Log-Likelihood:	-3226.5

Table S20. Random Effect Variance

Groups	Variance	Std. dev.
SHELF	0.0869	0.295
REEF_NAME	0.1488	0.386
SITE_NO:REEF_NAME	0.0739	0.272
TRANSECT_NO:SITE_NO:REEF_NAME	0.1584	0.398
A_SECTOR	0.0313	0.177

Table S21. Fixed Effects

Term	Estimate	Std. Error	z-value	p-value
(Intercept)	0.0222	0.238	0.093	0.926
OPENORCLOSED_AFTER2004O	-0.0085	0.129	-0.066	0.947

Estimated Marginal Means (Probability Scale)

Status	Probability	SE	95% CI
Closed	0.506	0.060	0.391 - 0.620
Open	0.503	0.060	0.389 - 0.618

Contrasts

Comparison	Estimate	SE	z-value	p-value
Closed - Open	0.00213	0.0323	0.066	0.947

Key Findings

No significant difference in turf algae cover between protection zones (z = -0.066, p = 0.947).

Redfield ratio | Is the origin of nutrients different between NTMRs and fished reefs?

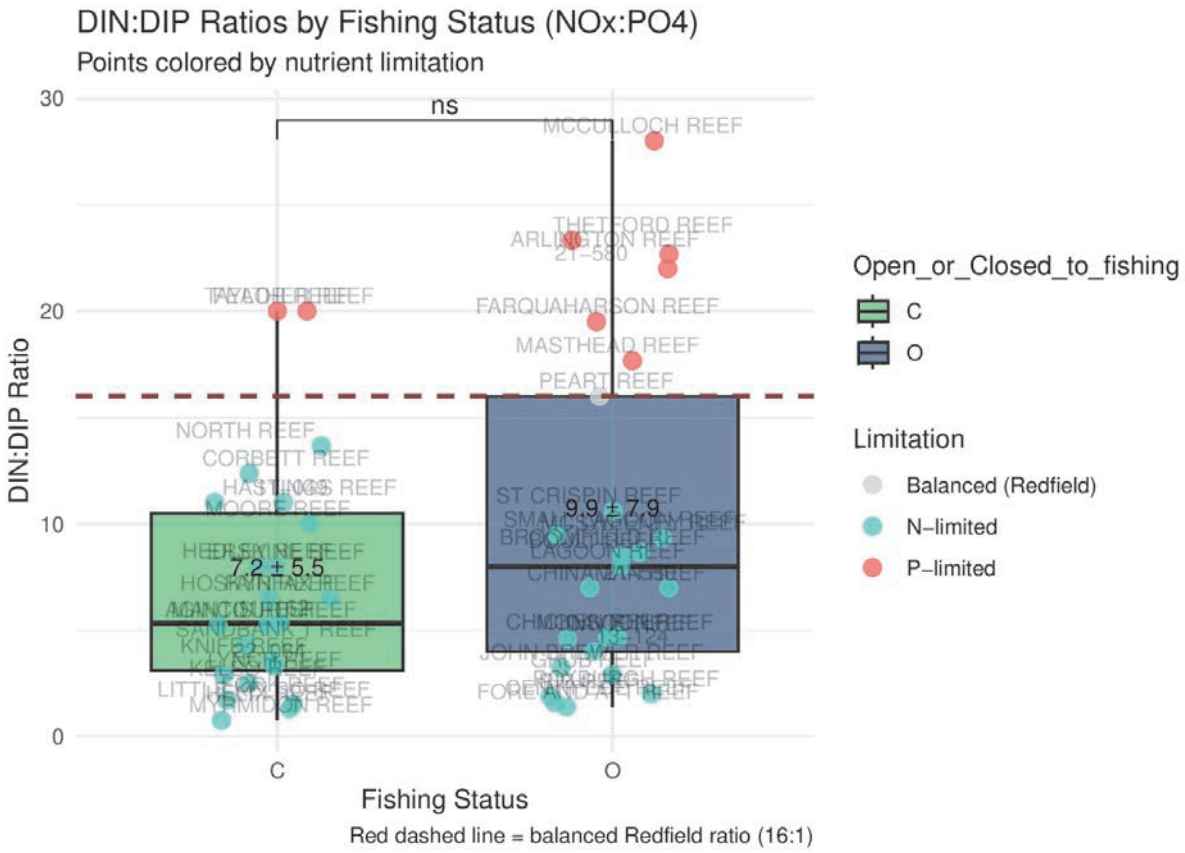

Fig. S13. The DIN:DIP ratio between NTMRs and fished reefs showing which sites were N- or P-limited compared to the balanced redfield ratio of 16:1 (marked with a red dashed line). Group-level comparisons were tested with a Wilcoxon rank sum test, and significance levels are indicated as: *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; "ns" = not significant.

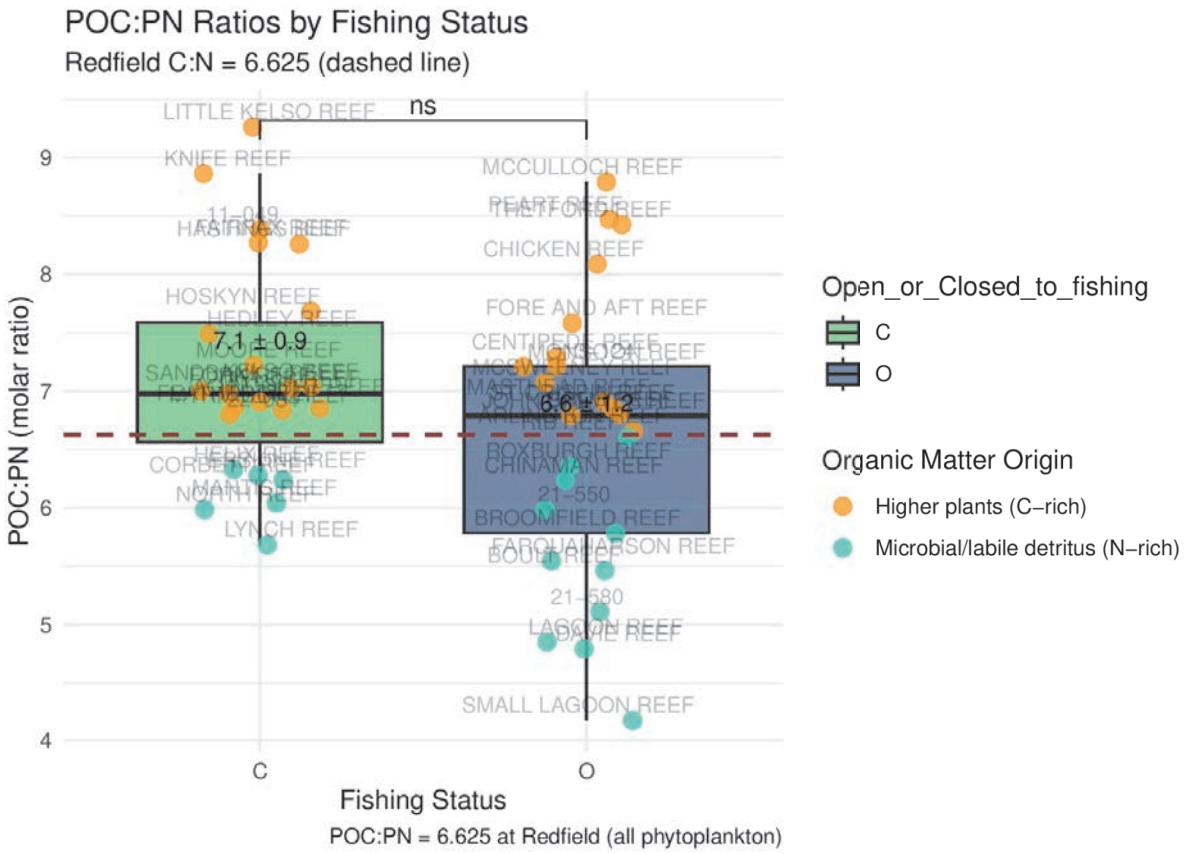

Fig. S14. The POC:PN ratio between NTMRs and fished reefs showing which sites were C- or N-enriched. POM with more carbon indicates origins primarily of plant material (macroalgae, seagrass) whereas N-enriched POM is more likely of bacterial origin, and containing more labile detritus. The POC:PN ratio of 6.625 at Redfield would be totally phytoplanktonic in origin, and is marked with a red dashed line. Group-level comparisons were tested with a Wilcoxon rank sum test, and significance levels are indicated as: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; "ns" = not significant.

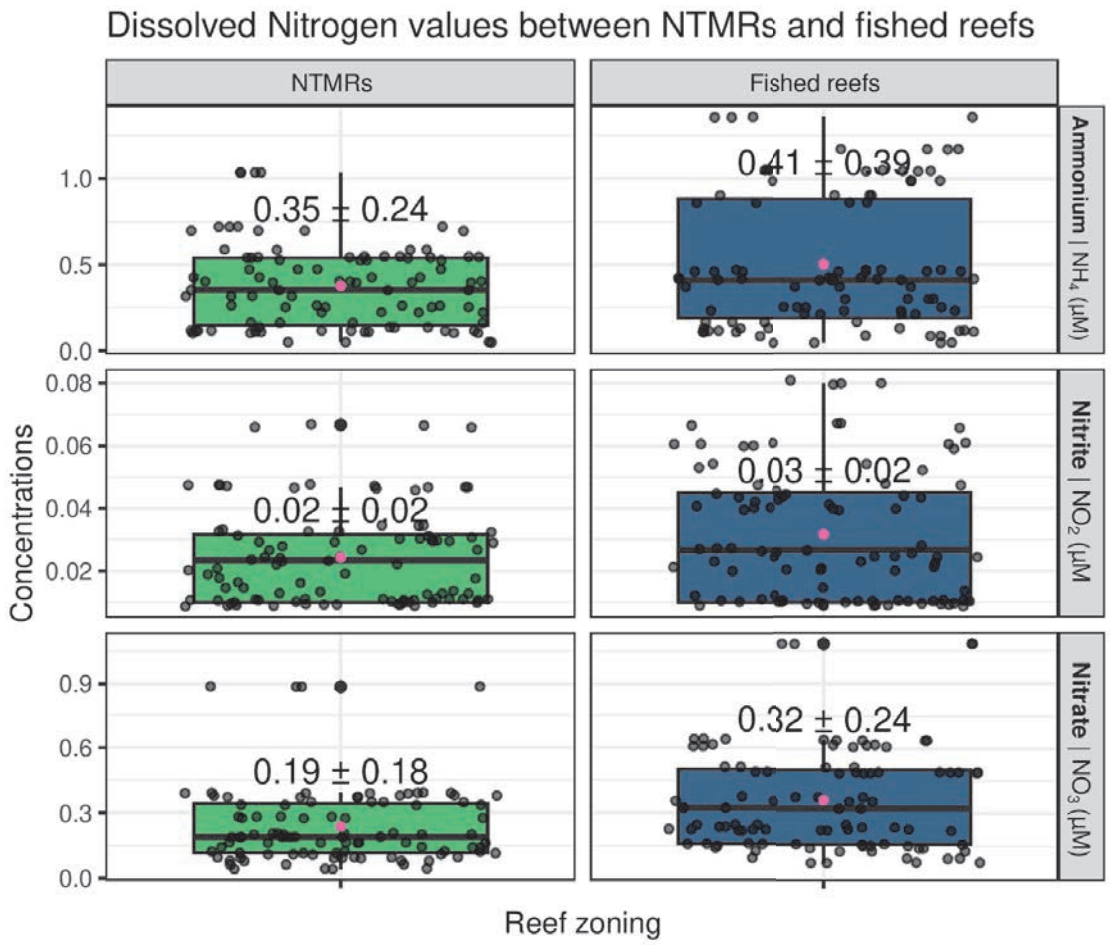

Fig. S15. Dissolved Nitrogen data. Median ± SD values for ammonium, nitrite, and nitrate, summarised across the reef zones (NTMRs vs fished reefs). Group-level comparisons were tested with a Wilcoxon rank sum test, presented in Table S22 (below).

Table S22. Wilcoxon rank sum tests comparing dissolved Nitrogen (ammonium, nitrite, and nitrate) values between NTMRs and fished reefs. Significance levels are indicated as: *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; "ns" = not significant.

variable group1 group2 n1 n2 statistic p.adj.bonferroni p.adj.signif

NH4_µM C O 95 95 4041 0.214 ns

NO2_µM C O 95 95 3824 0.065 ns

NO3_µM C O 95 95 2980 0.0000527 ****

Microbial co-occurrence networks

The positive:negative cohesion ratio computed for each reef sample (48 reefs x 4 replicates), and visualised separately for each of the 7 GBR sectors (**Fig. S16**) and across the 4 sampled trips (**Fig. S18**). The boxplots show inner quartiles and median positive:negative cohesion ratio (shown as an absolute

value) on the y axis, and a higher value indicates a prevalence of positive (i.e. symbiosis, metabolic co-dependency) compared to negative (i.e. competition) interactions in the microbial community. Specifically for reef zoning, this positive:negative cohesion ratio was consistently higher in NTMRs (median ± SD of positive:negative cohesion in: Cape Grenville - CG: 1.17 ± 0.10; Princess Charlotte bay - PC: 1.12 ± 0.08; Cairns - CA: 1.52 ± 0.05; Innisfail - IN: 1.20 ± 0.08; Townsville - TO: 1.31 ± 0.22; Swains - SW: 1.22 ± 0.10) compared to fished reefs (median ± SD of positive:negative cohesion ratios equalling to CG: 1.00 ± 0.16; PC: 0.97 ± 0.07; CA: 1.36 ± 0.13; IN: 1.03 ± 0.04; TO: 0.99 ± 0.19; SW: 1.13 ± 0.17) across 6 GBR sectors, apart from the Capricorn Bunker (CB) sector (with median ± SD of positive:negative cohesion being slightly higher in fished reefs: 1.08 ± 0.11, compared to NTMRs: 1.05 ± 0.09) (**Fig. S16**). Based on the results of the Wilcoxon Rank Sum tests, the positive:negative cohesion ratios were significantly higher in NTMR compared to fished zones in the following sectors: CG (W = 111, p-adj = 0.02), PC (W = 47, p-adj = 0.03), CA (W = 86, p-adj = 0.002), IN (W = 162, p-adj = 0.0002), and TO (W = 569, p-adj = 0.003). However, no significant difference was found in the southern GBR sectors SW (W = 59, p-adj = 0.427) and CB (W = 93, p-adj = 0.91).

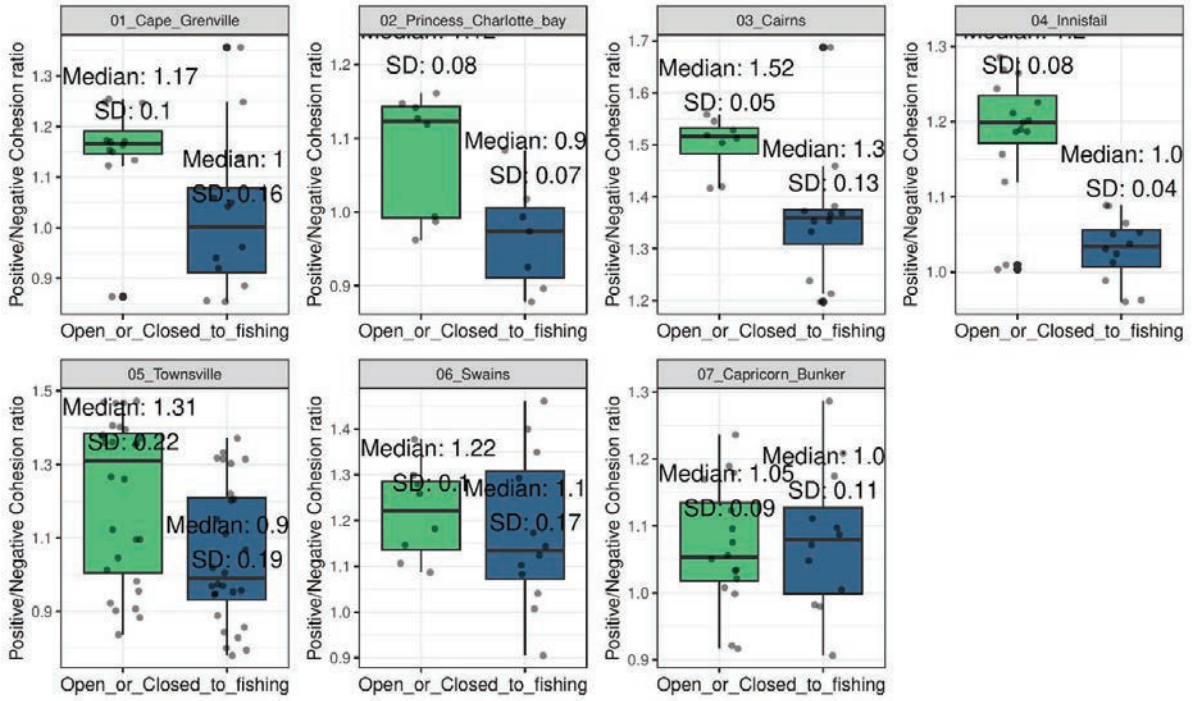

Figure S16. Sample-level positive:negative cohesion ratios indicate the prevalence of positive (increase in positive to negative cohesion ratio) in or negative (increase in positive to negative cohesion ratio) interactions within reef bacterioplankton between fished reefs (blue) and NTMRs (green), in each of the 7 GBR sectors we sampled.

Further, the higher positive:negative cohesion was also observed for the sites sampled in the winter trip, suggesting a prevalence of positive interactions (mutualism, co-occurrence due to metabolic exchange) in the winter (**Fig. S17**) when nutrients are depleted (**Fig. S18**). In contrast, we see a potential increase of negative/mutually exclusive interactions (predator/prey, pathogen/host, parasite/host, and etc.)

in the summer trips (**Fig. S17**) when nutrients are elevated (**Fig. S18**), potentially indicating that opportunistic microbes are competing for available nutrients that are elevated in the summer transects.

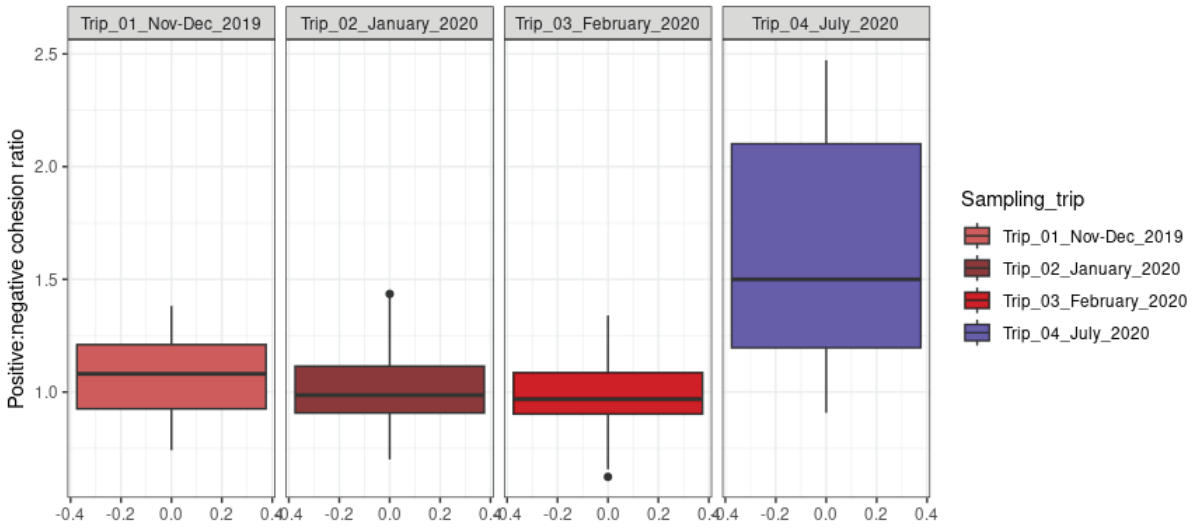

Figure S17. Sample-level positive:negative cohesion ratios indicate the prevalence of positive (increase in positive to negative cohesion ratio) in or negative (increase in positive to negative cohesion ratio) interactions within reef bacterioplankton between summer sampling transects (Trips 1-3, red) and the winter trip (Trip 4; blue).

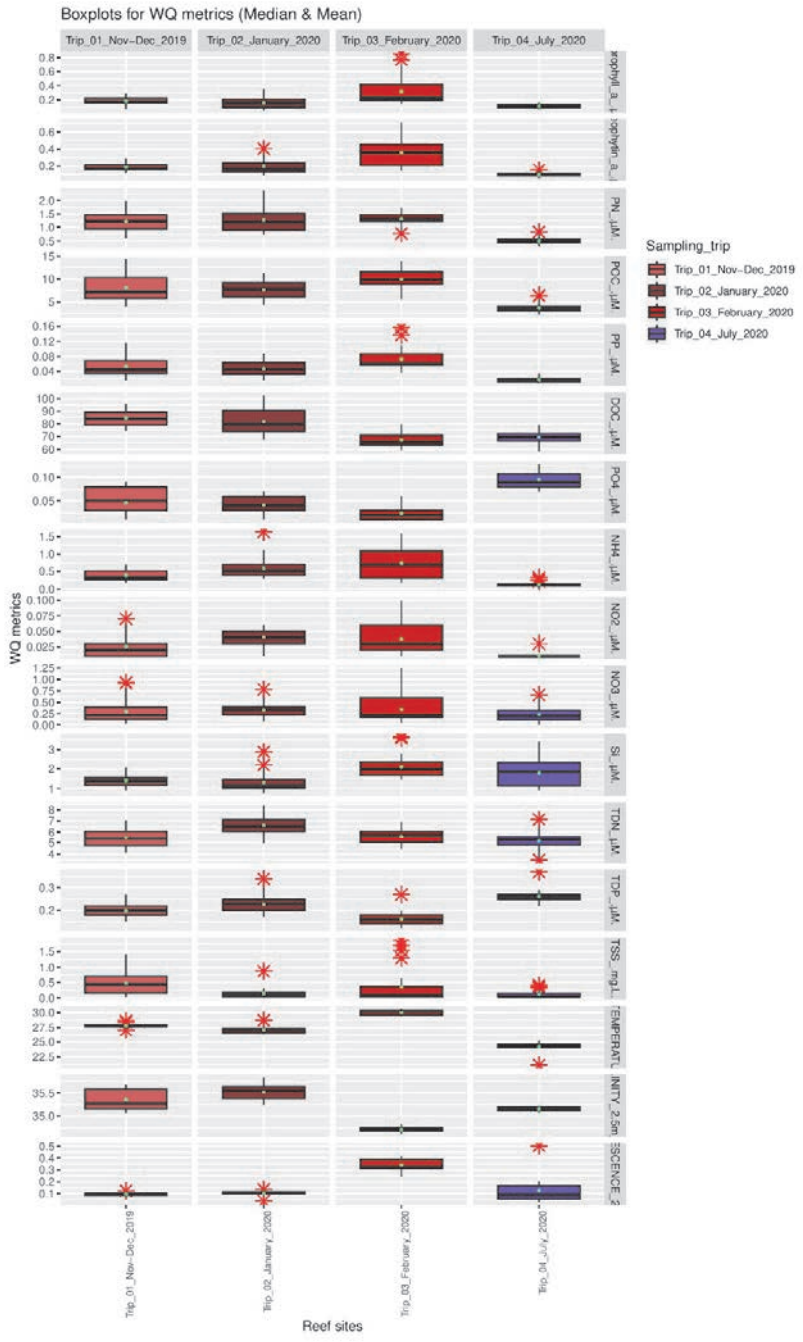

Figure S18: Physico-chemical data. Median ± SD values of 17 physico-chemical variables collected. Values are summarised across the four sampling trips, with the colour code corresponding to Fig. 3.1 in the main text. Acronyms explained: ammonia (NH_4^+), nitrite (NO_2^-), nitrate (NO_3^-), total dissolved nitrogen (TDN), phosphate (PO_4^{3-}), total dissolved phosphorus (TDP), dissolved organic carbon (DOC), silicate (Si), total suspended solids (TSS), chlorophyll a (Chl-*a*), phaeophytin a (Phaeo), particulate organic carbon (POC), particulate nitrogen (PN), and particulate phosphorus (PP).

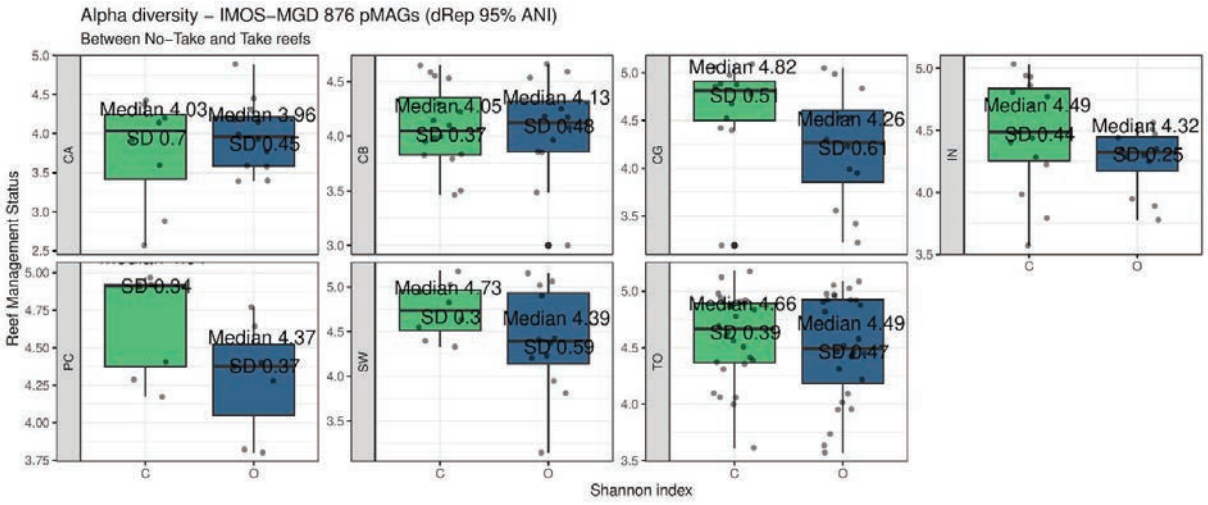

Figure S19. Microbial diversity between the zones. Alpha diversity (Shannon index) of seawater microbiomes between NTMRs (green) and fished reef sites (blue) across the 7 sampled GBR sectors.

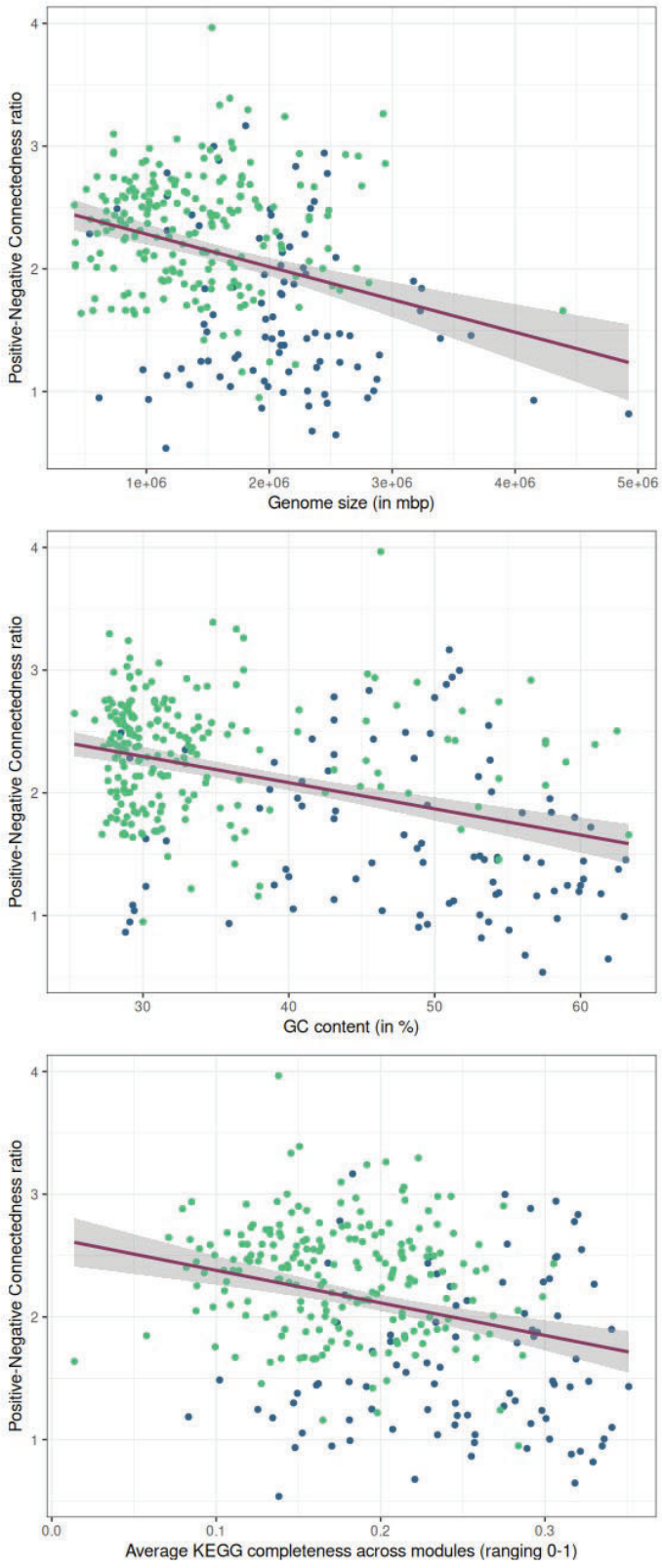

Figure S20. Linear regression models linking microbial genome features including genome size (**A**), GC content (**B**), and metabolic pathway completeness (**C**) to network connectedness (positive:negative edge ratio) as the response

variable, for metagenome-assembled genomes (MAGs) indicative of No-Take Marine Reserves (NTMRs; green) and fished reefs (blue).

KEGG Pathway analysis | Which metabolic traits are enriched in microbial indicators of NTMRs vs. fished reefs?

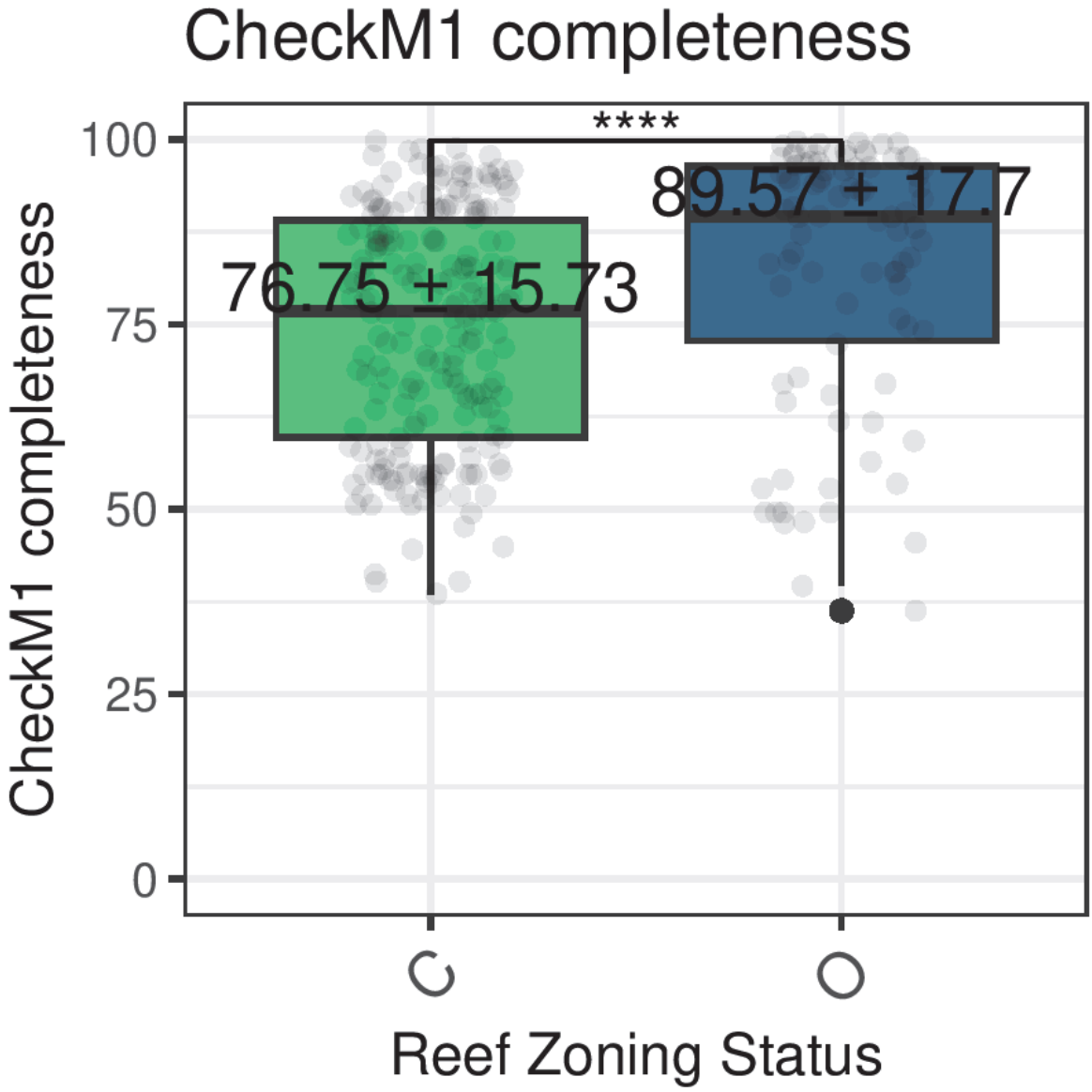

Fig. S21. Boxplots comparing CheckM1 genome completeness score distributions between microbial indicators of NTMRs vs fished reefs. Significance levels from Wilcoxon rank sum tests are indicated as: *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; "ns" = not significant.

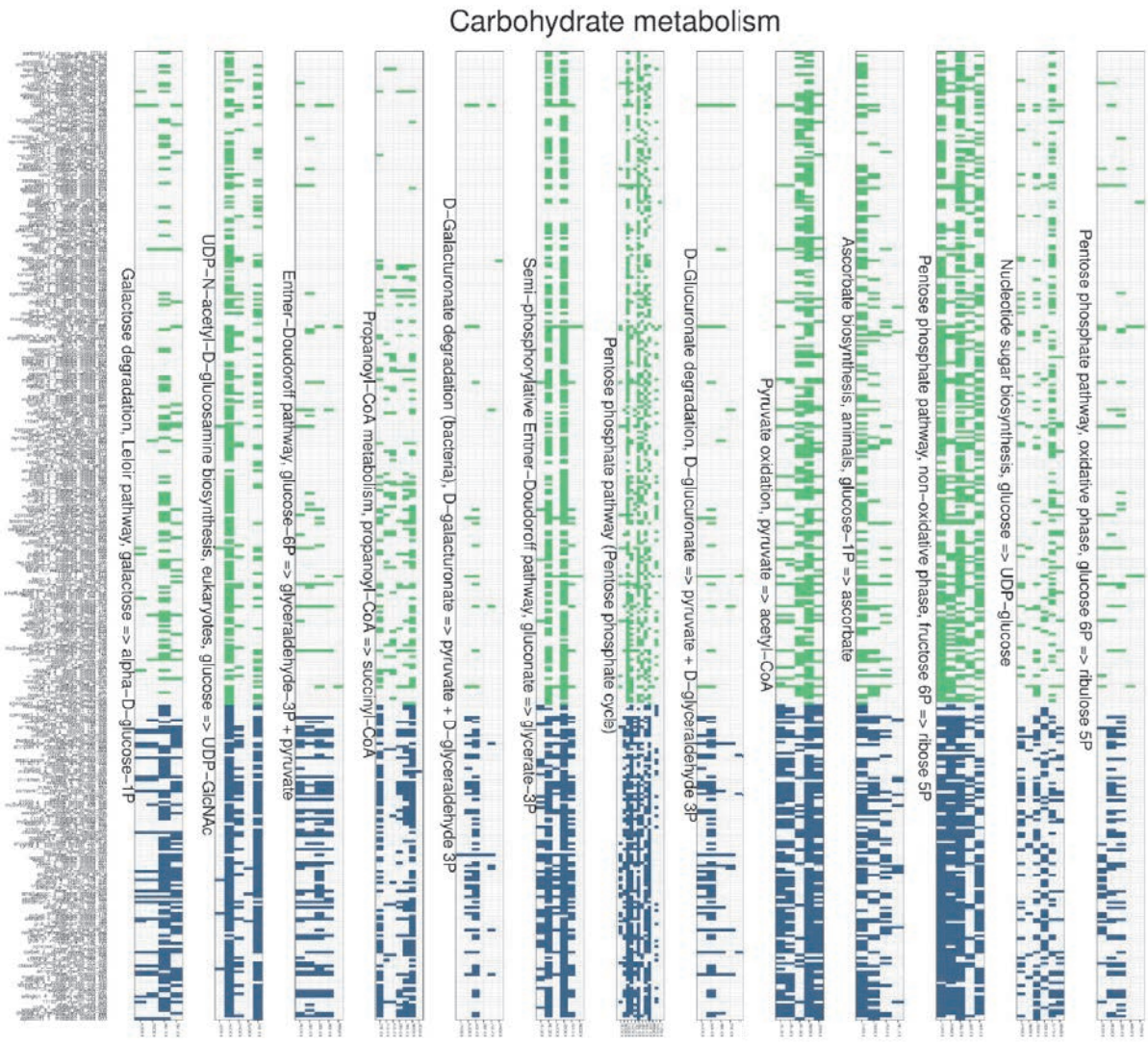

Fig. S22. Gene level module completeness (Carbohydrate metabolism).

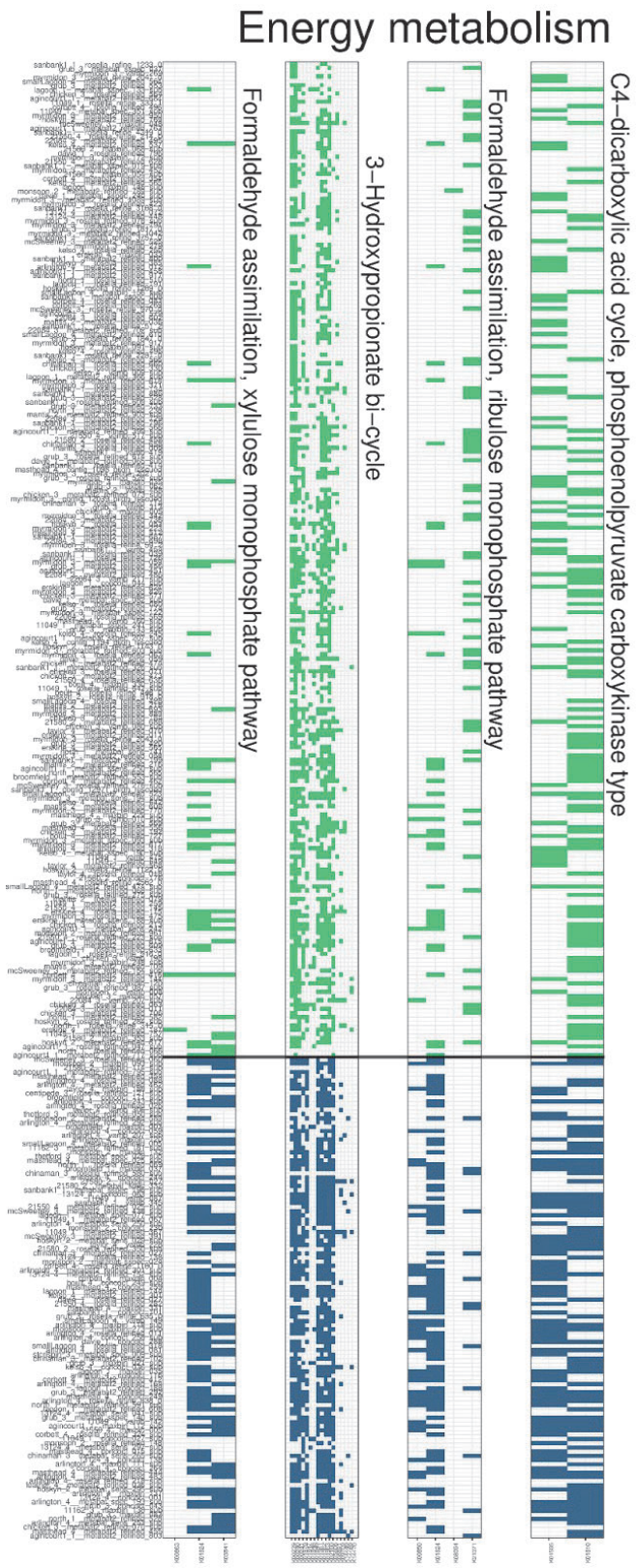

Fig. S23. Gene level module completeness (Energy metabolism).

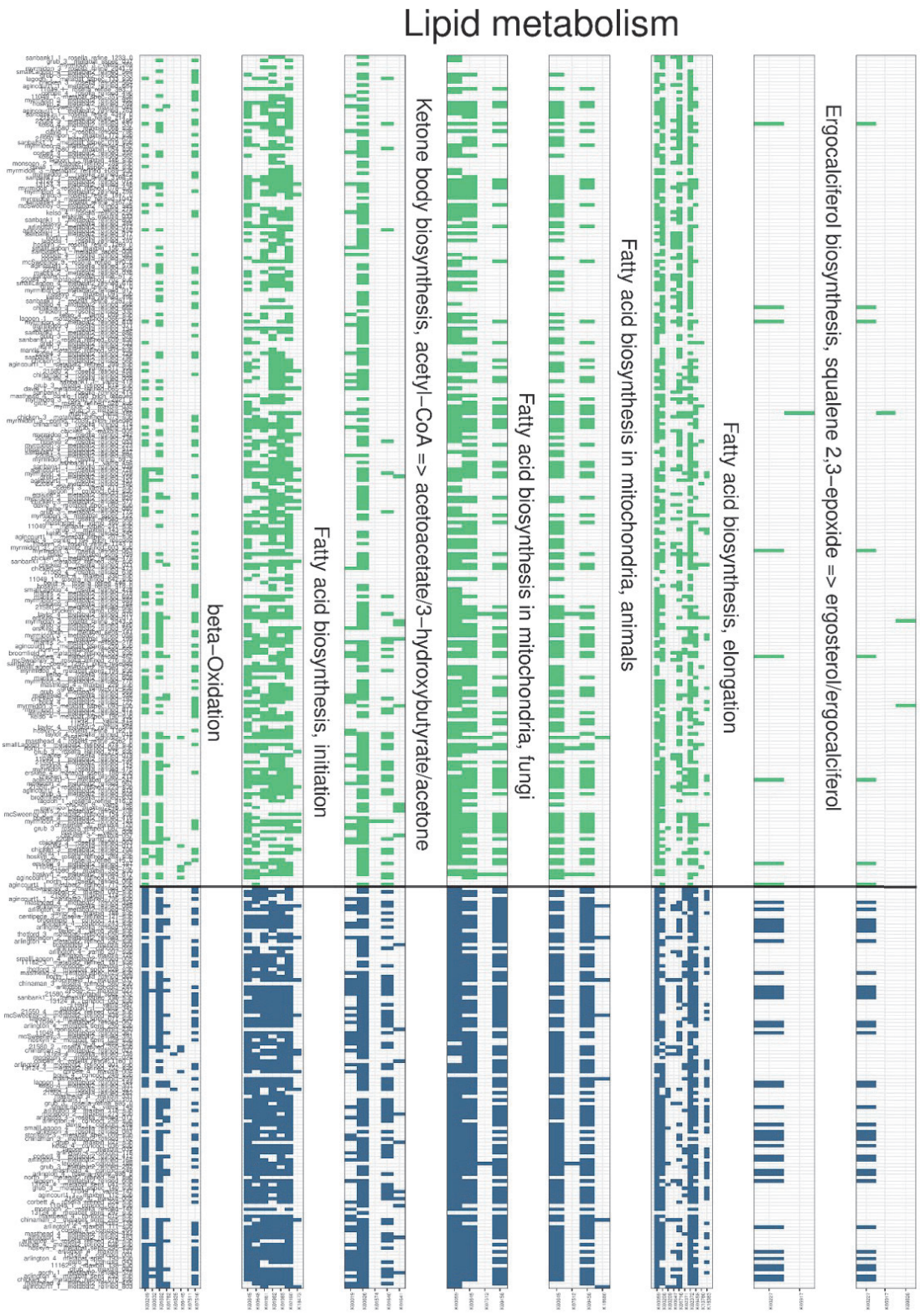

Fig. S24. Gene level module completeness (Lipid metabolism).

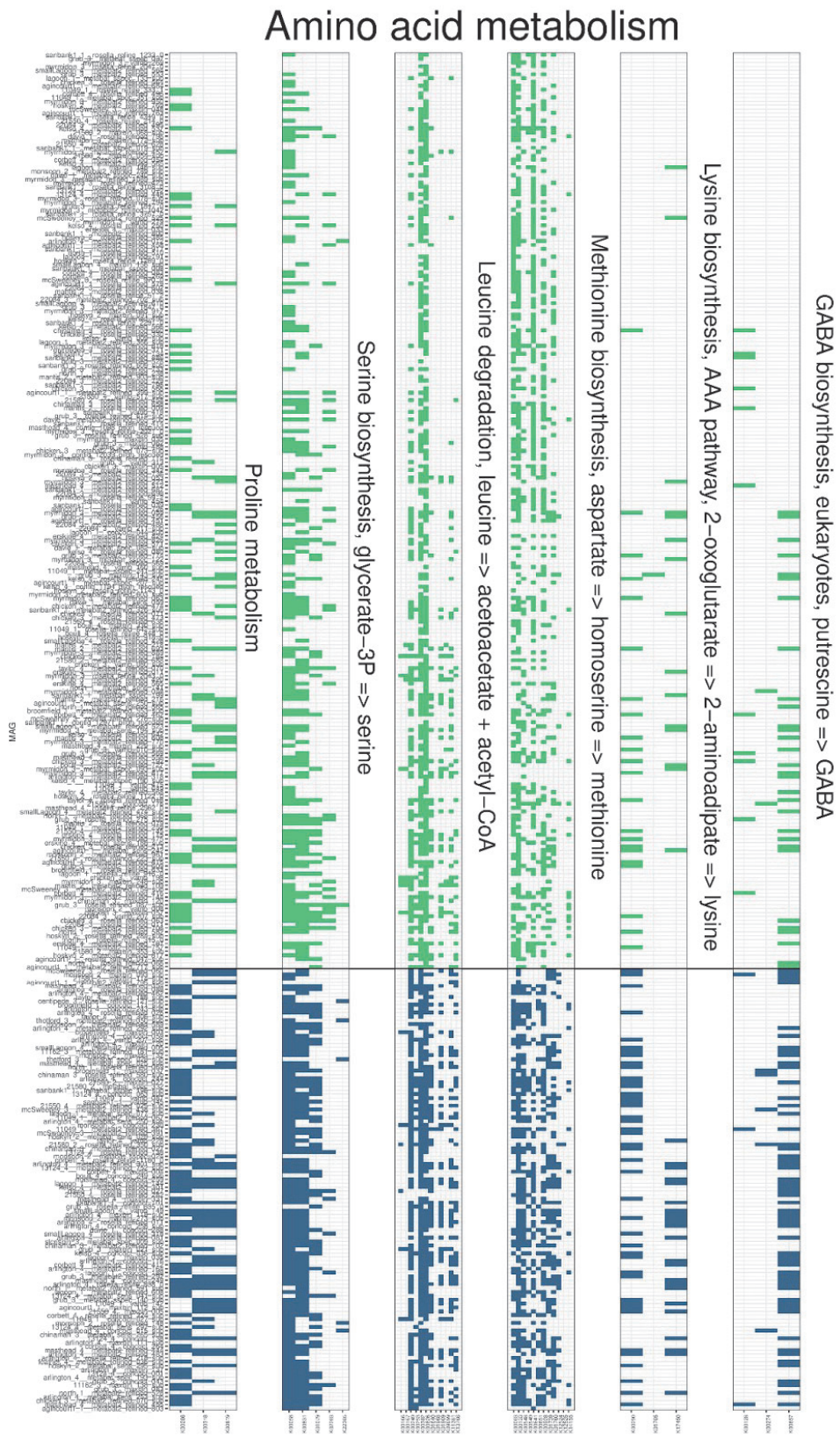

Fig. S25. Gene level module completeness (Amino acid metabolism).

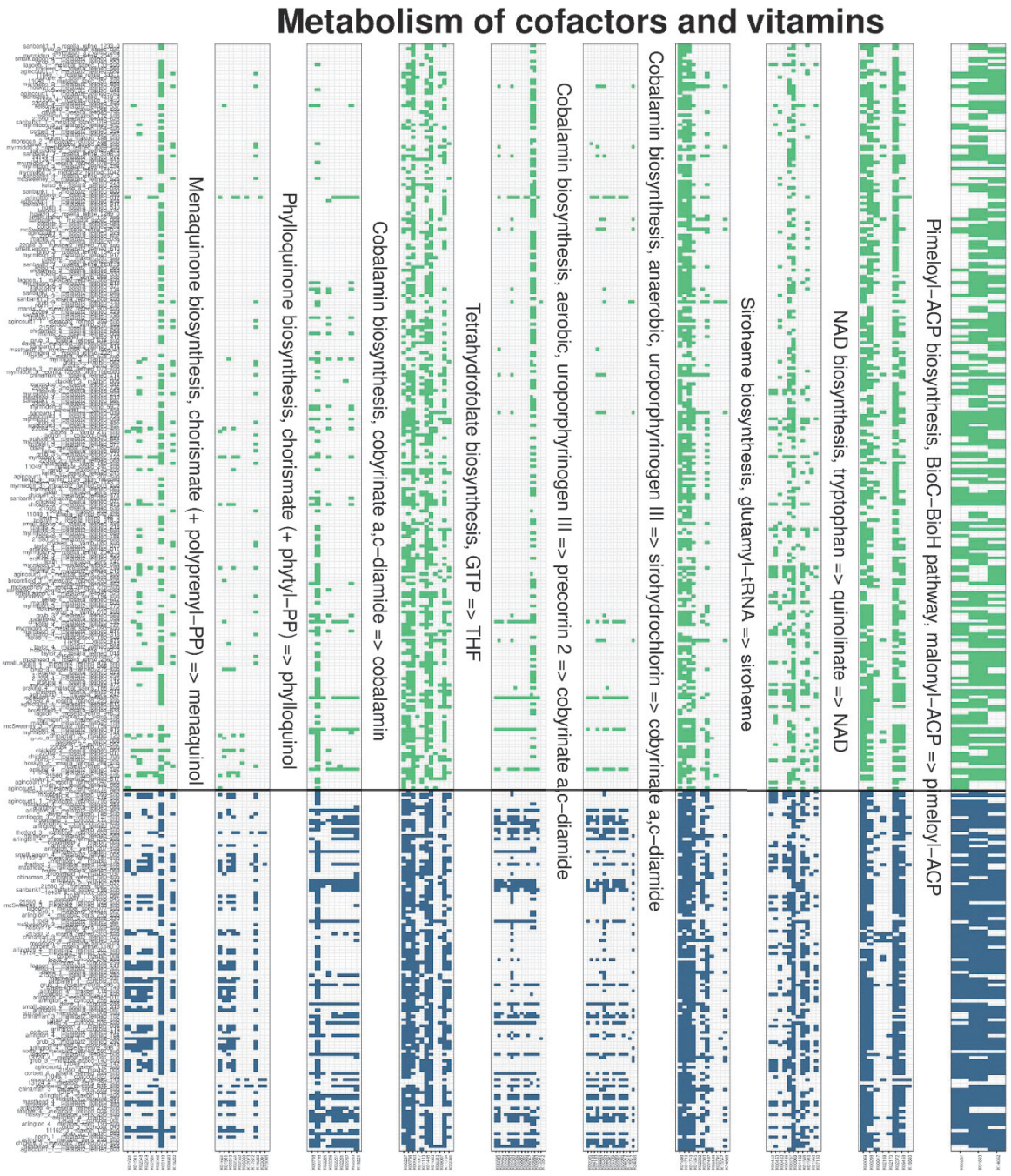

Fig. S26. Gene level module completeness (Metabolism of cofactors and vitamins).

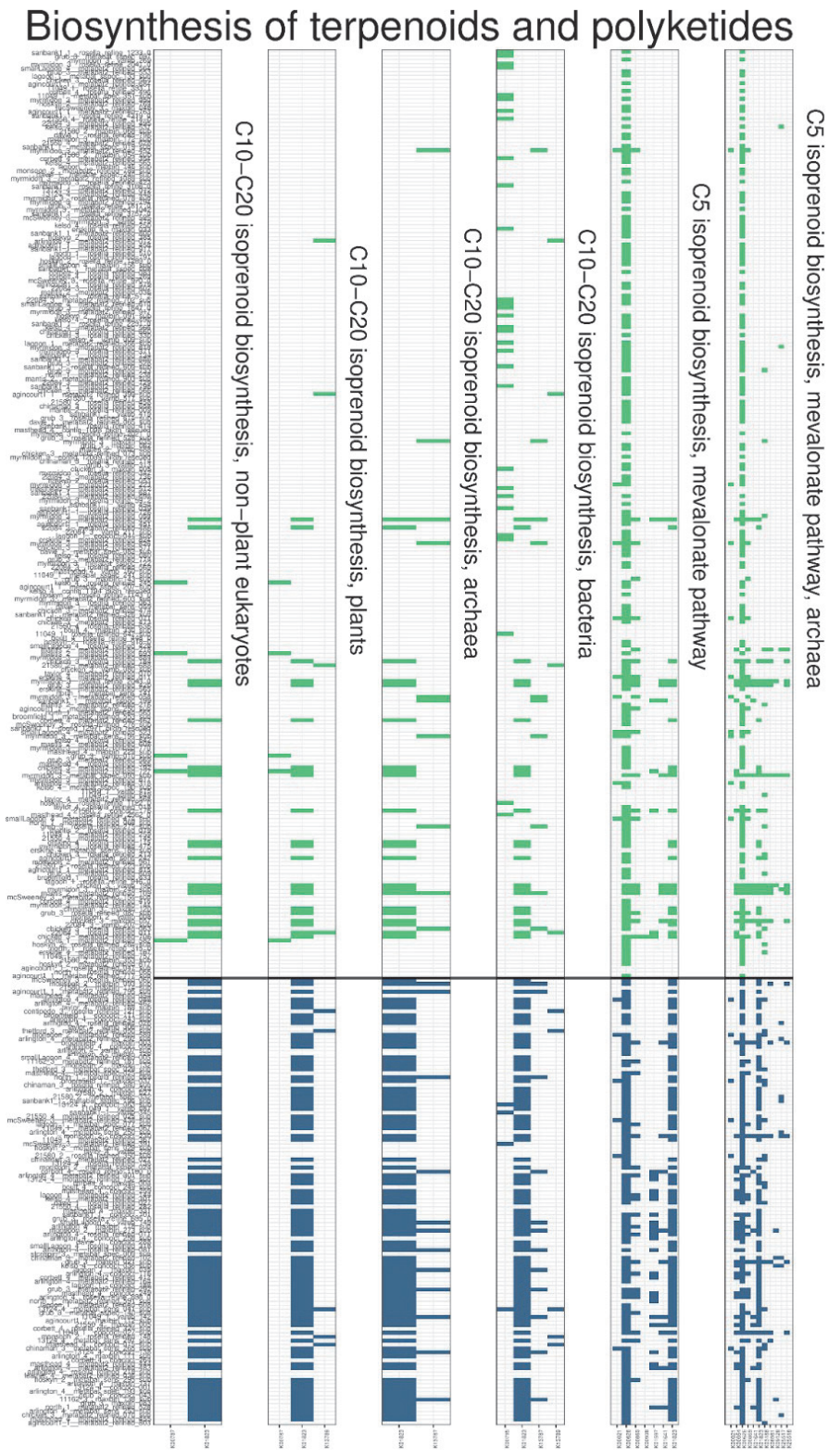

Fig. S27. Gene level module completeness (Biosynthesis of terpenoids and polyketides).

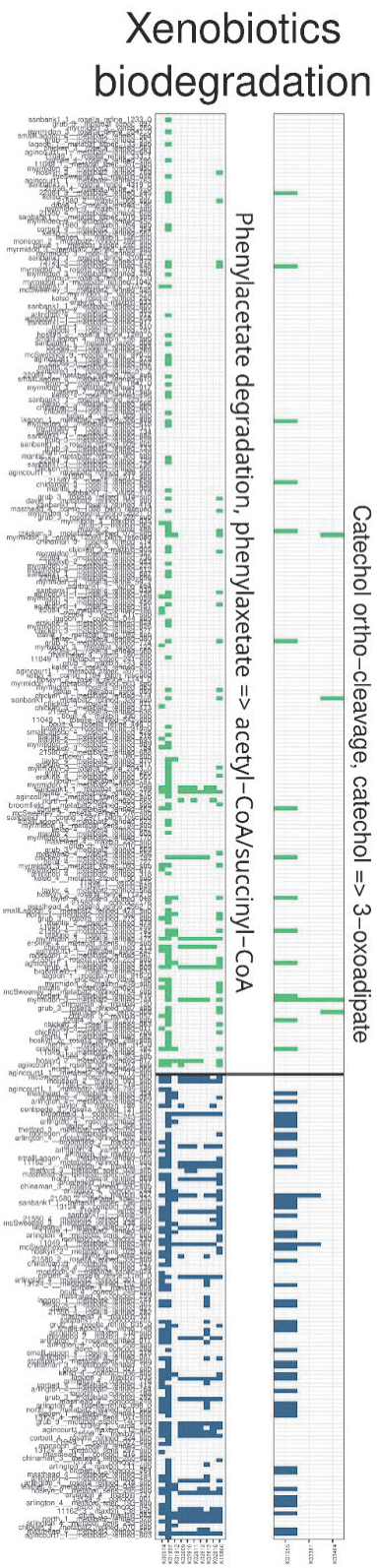

Fig. S28. Gene level module completeness (Xenobiotics biodegradation).

Fig. S29. Functional potential for nitrogen acquisition in zoning-indicator MAGs. Heatmap of nitrate, nitrite, ammonium, and urea transporter and metabolic genes, which were colored if present (NTMR-enriched microbes in green; fished-reef enriched microbes in blue).

Robust **O**ptimum (**RO**) method | Microbial niche modelling

For each microbial predictor, we computed the microbial niche tolerance ranges using the robust optimum method[307], as explained in the main text. This allowed us to define the lower and upper niche limits—the environmental conditions below and above which the $pMAGs_{95\%ANI}$ cannot survive due to unfavorable conditions—as well as the ecological niche optimum, which corresponds to the environmental value at which that microbe is found at its highest relative abundances, i.e. optimal conditions for the existence, development, growth, and proliferation of that microbe (Ter Braak & van Dam, 1989). A comparative niche analysis focused on the 350 indicator $pMAG_{s95\%ANI}$ that distinguish between NTMRs and fished reefs, particularly in relation to dissolved nitrogen variables (NH_4, NO_2, and NO_3) which were the explanatory drivers on fished reefs. Pairwise comparisons of niche preferences for these dissolved nitrogen variables between NTMRs and fished reefs were visualised with boxplots in ggplot2[308] (v3.5.1).

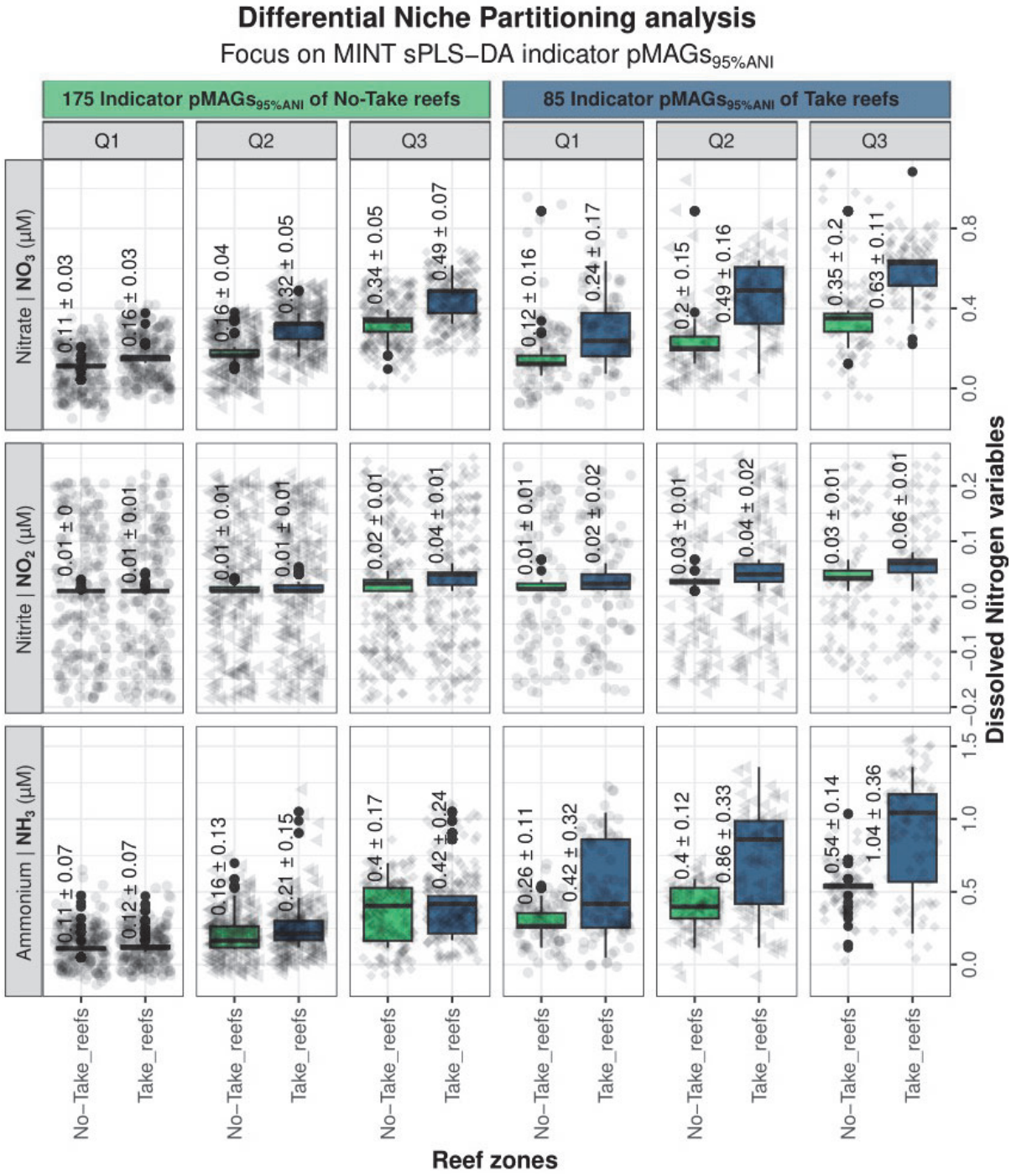

Figure S30. Comparative niche analysis for dissolved nitrogen variables between microbial indicators of fished reefs vs NTMRs. Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the 236 microbial indicators of NTMRs (left) and 114 indicators enriched in fished reefs (right)—for ammonium (bottom), nitrite (middle), and nitrate (top). Niche bound values are visualised using distinct point shapes. This plot shows that microbial indicators of fished reefs have a preference towards higher values of all dissolved Nitrogen variables in fished reefs.

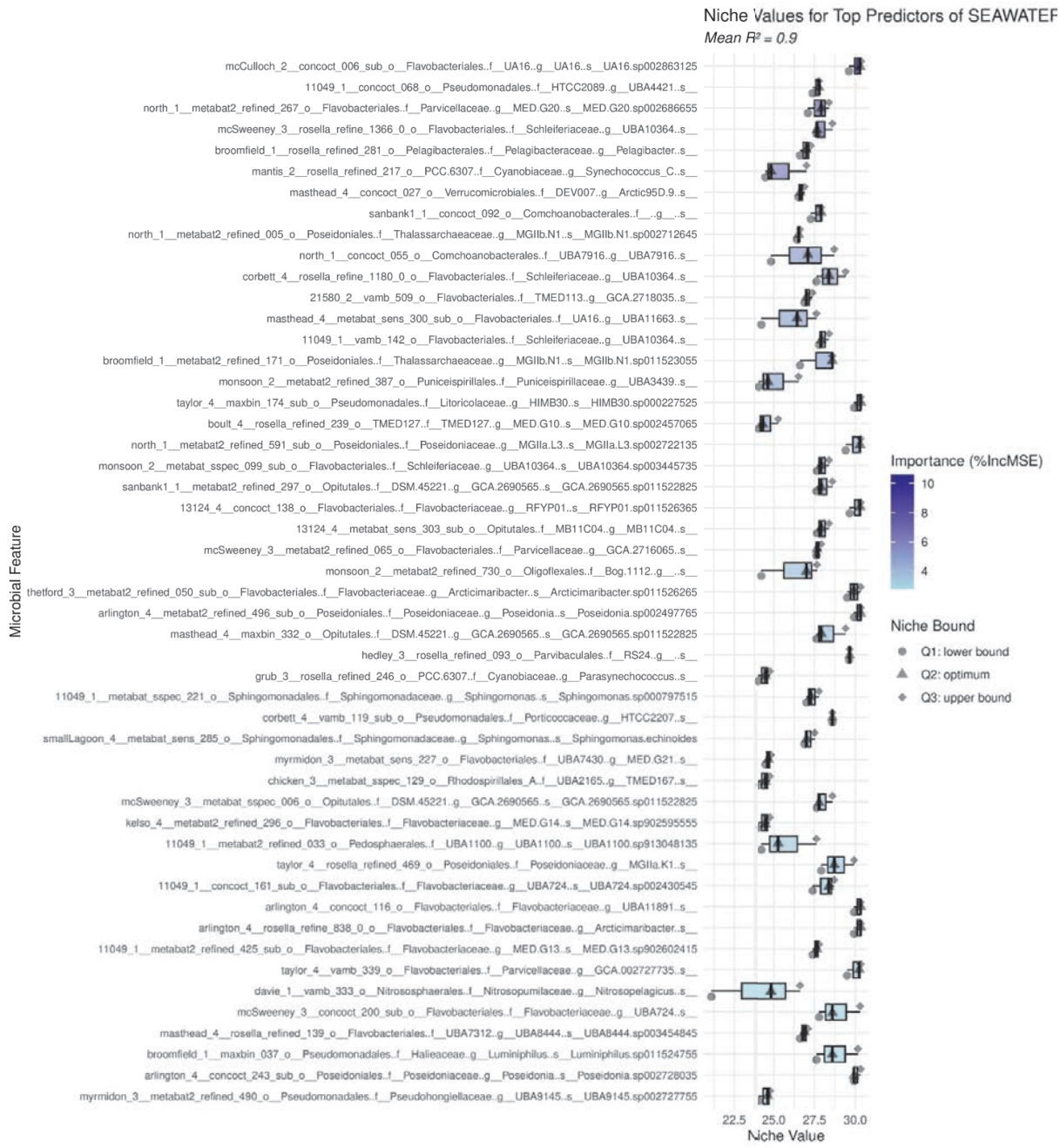

Figure S31. Microbial predictors of surface seawater temperature (SST). Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the top 50 microbial predictors per pMAG for SST. Niche bound values are visualised using distinct point shapes: circles (Q1), triangles (Q2), and squares (Q3). In addition, microbial predictors are additionally colored by random forest importance (%IncMSE), using a light-to-dark blue gradient (light blue = low importance, dark blue = high importance).

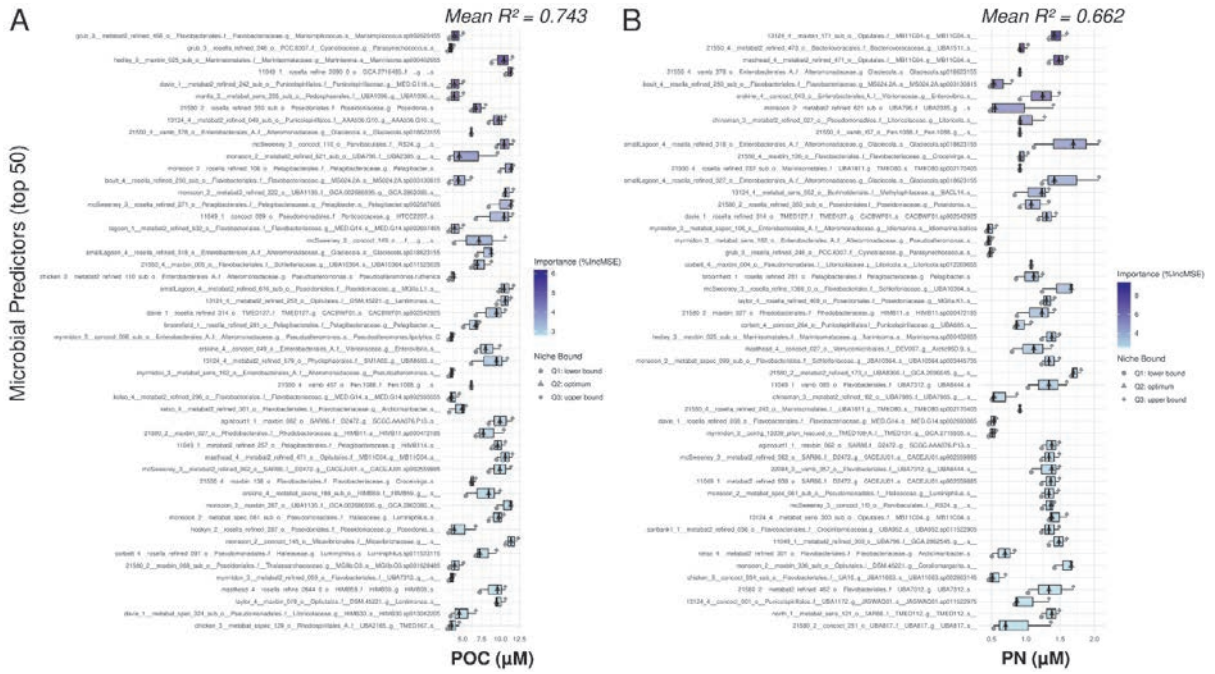

Figure S32. Microbial predictors of particulate nutrients. Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the top 50 microbial predictors per pMAG for (**A**) Particulate organic carbon (POC) and (**B**) particulate nitrogen (PN). Niche bound values are visualised using distinct point shapes: circles (Q1), triangles (Q2), and squares (Q3). In addition, microbial predictors are additionally colored by random forest importance (%IncMSE), using a light-to-dark blue gradient (light blue = low importance, dark blue = high importance).

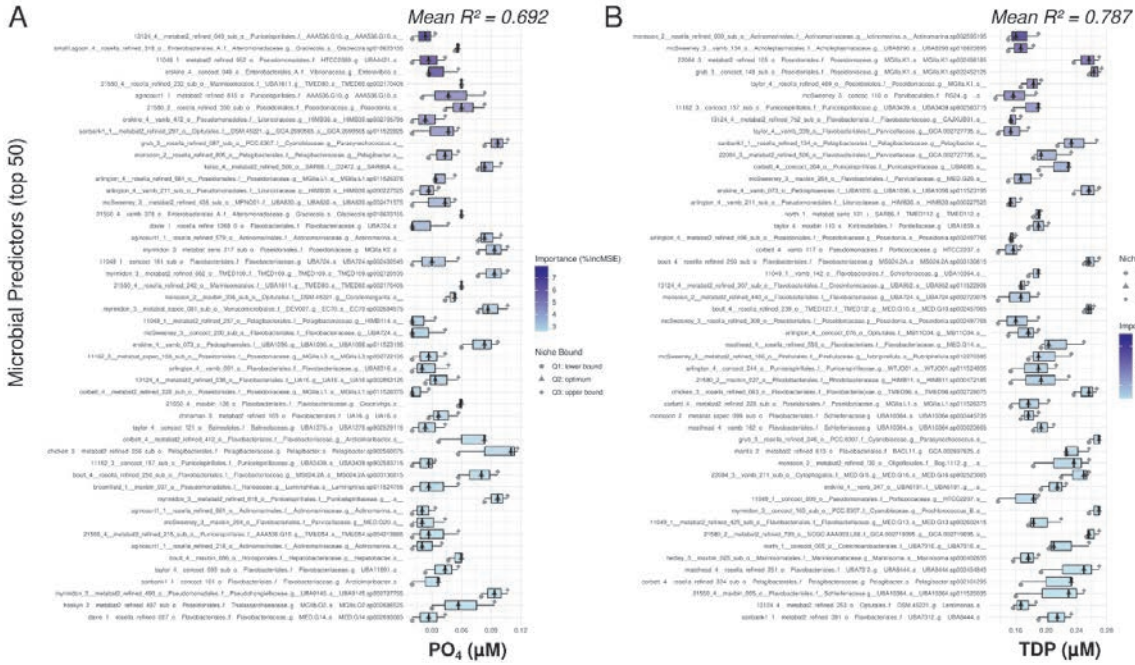

Figure S33. Microbial predictors of dissolved phosphorus. Boxplots show niche tolerance ranges (Q1: lower bound, Q2: optimum, Q3: upper bound) for the top 50 microbial predictors per pMAG for (A) phosphate (PO_4^{3-}) and (B) total dissolved phosphorus (TDP). Niche bound values are visualised using distinct point shapes: circles (Q1), triangles (Q2), and squares (Q3). In addition, microbial predictors are additionally colored by random forest importance (%IncMSE), using a light-to-dark blue gradient (light blue = low importance, dark blue = high importance).

10 Appendix C – Supplementary Material for Chapter 4

PCA - **P**rincipal **C**omponents **A**nalysis | What are the main clustering patterns across our samples?

Principal Components Analysis (PCA) was applied in an R package mixOmics[292] (v6.26.0) as an unsupervised approach to visualise the main clustering patterns between reef sites based on microbial and viral community profiles. The number of optimal PCA components was determined using the tune.pca() function in mixOmics.

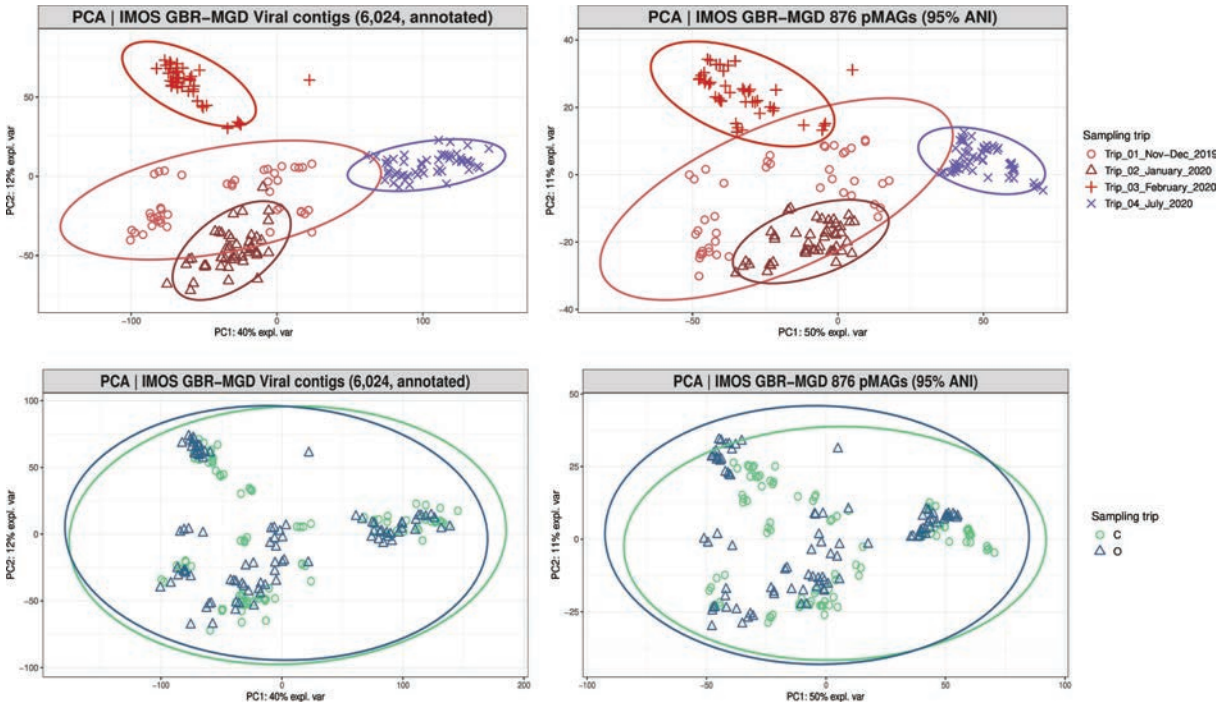

Figure S1. The PCA ordination plots shows the main clustering patterns of seawater microbial communities (viruses - left; pMAGs - right), colored per sampling transect (above) or reef zoning (below). We can observe clear differences both between microbial and viral communities which cluster between summer/wet season (red) and winter/dry season (blue) on PCA dimension 1, with samples collected in the peak of summer (Trip 3) additionally separating from early summer sampling (Trips 1 and 2) on PCA dimension 2, for both pMAGs and Virus datasets. Conversely, PCA ordination does not show clear clustering between No-Take Marine Reserves (C - closed to fishing, green) and fished reefs (O - open to fishing, blue).

PCA plots show that both microbial and viral communities cluster based on time of sampling with winter samples clearly forming a well separated cluster (**Fig. S1**). When looking at prokaryotic community composition to identify which bacterial and archaeal taxa are driving this structure, we observe that Trips 1-3 (November 2019 - February 2020) were characterised by high relative abundances of a cyanobacterium *Synechococcus* (e.g., *Parasynechococcus* and *Synechococcus_C*), genus UBA8309 (classified within Puniceispirillaceae) and *Sphingomonas*. A noticeable shift in the dominant prokaryotic

taxa was observed during the winter Trip 4 (July 2020), when *Prochlorococcus* (g__Prochlorococcus_A) becomes more abundant (at the expense of *Synechococcus*), in addition to the alphaproteobacterial groups SAR11 (*Pelagibacter*) (**Fig. S2**; above).

The viral community was overwhelmingly dominated by the class Caudoviricetes (tailed bacteriophages) across all four sampling time points. The class Megaviricetes (giant viruses) was consistently identified as the second most abundant viral group. Other viral classes, including Faserviricetes, Malgrandaviricetes, Herviviricetes, Tokiviricetes, and Pokkesviricetes, were present in comparatively low abundances (**Fig. S2**; below).

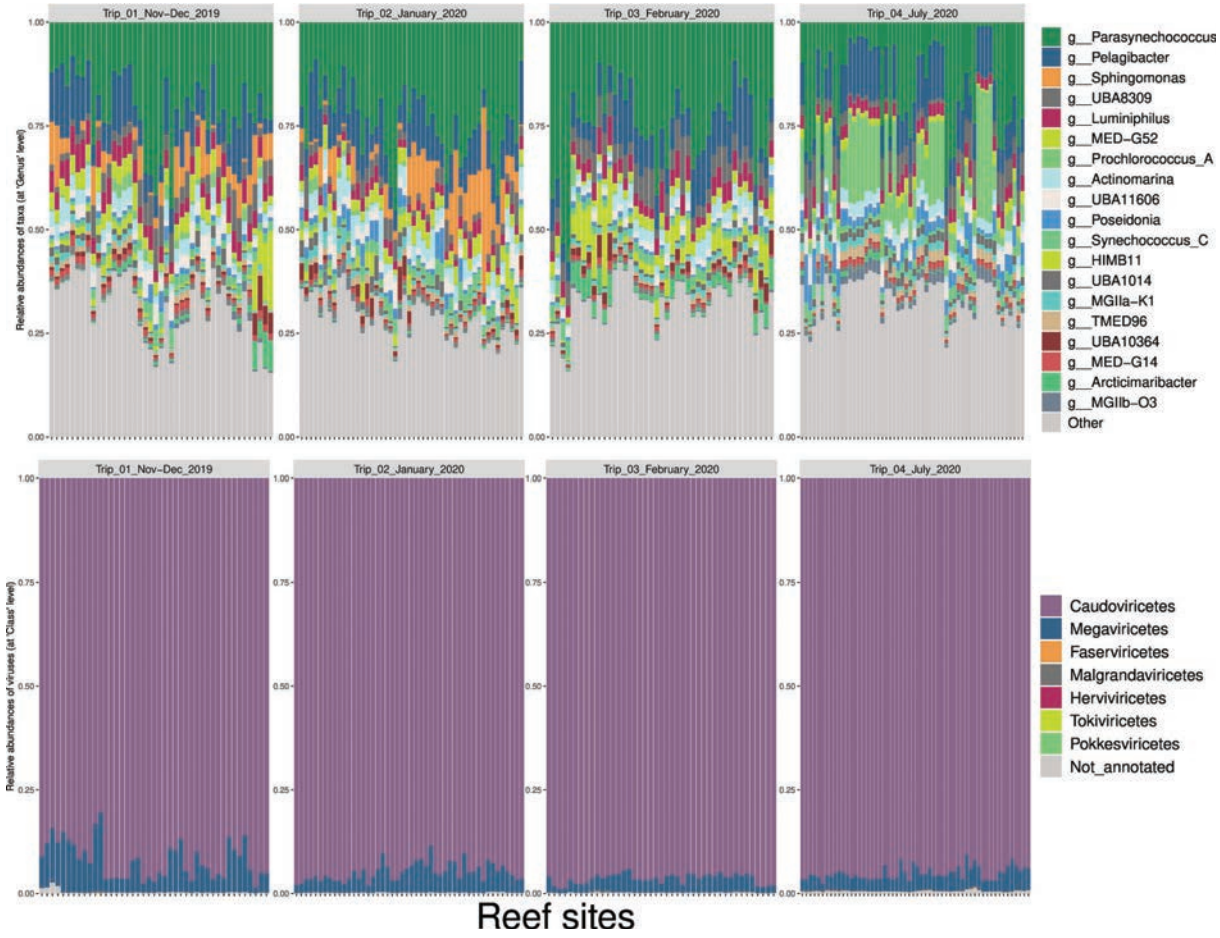

Figure S2. Microbial and viral community composition. Stacked barplots illustrate microbial (above) and viral (below) relative abundances (y-axis) for each sample (x-axis), with reef sites grouped by their corresponding sampling trip. These barplots represent the following the top 20 most abundant microbial genera (with all other classified within "Other"), all of the identified viral classes, including viral contigs not annotated at Class level.

Despite most viral contigs being classified predominantly to the class level (e.g., Caudoviricetes) (**Fig S2**; below), our analysis reveals clear seasonal patterns. Different viral contigs show strongly differential abundance across sampling trips, a pattern primarily driven by season (summer/wet vs. winter/dry) (**Fig. S3**). Furthermore, the iPHoP host prediction results indicate that these seasonally abundant viruses are predicted to infect the microbial hosts that are also most prevalent in that same season (**Fig. S4**). For instance, viral contigs most abundant in winter are predicted to target winter-dominant cyanobacterium *Prochlorococcus*, in addition to streamlined microbial hosts such as *Pelagibacter* (SAR11) and SAR86 (**Fig. S4; Trip 4**).

Figure S3. Relative abundance of viral contigs that drive seasonal differences in community composition. Bubble plot shows the relative abundance of viral contigs identified as major contributors (loadings > 0.91) to PCA separation of viral community samples. Viral contigs are grouped vertically by sampling trip in which they were most abundant (their 'specialism'), and samples are grouped horizontally by their collection sampling trip. Bubble size corresponds to relative abundance within each sample. This visualization highlights viruses that are strongly associated with specific seasonal conditions (summer in red vs. winter in blue).

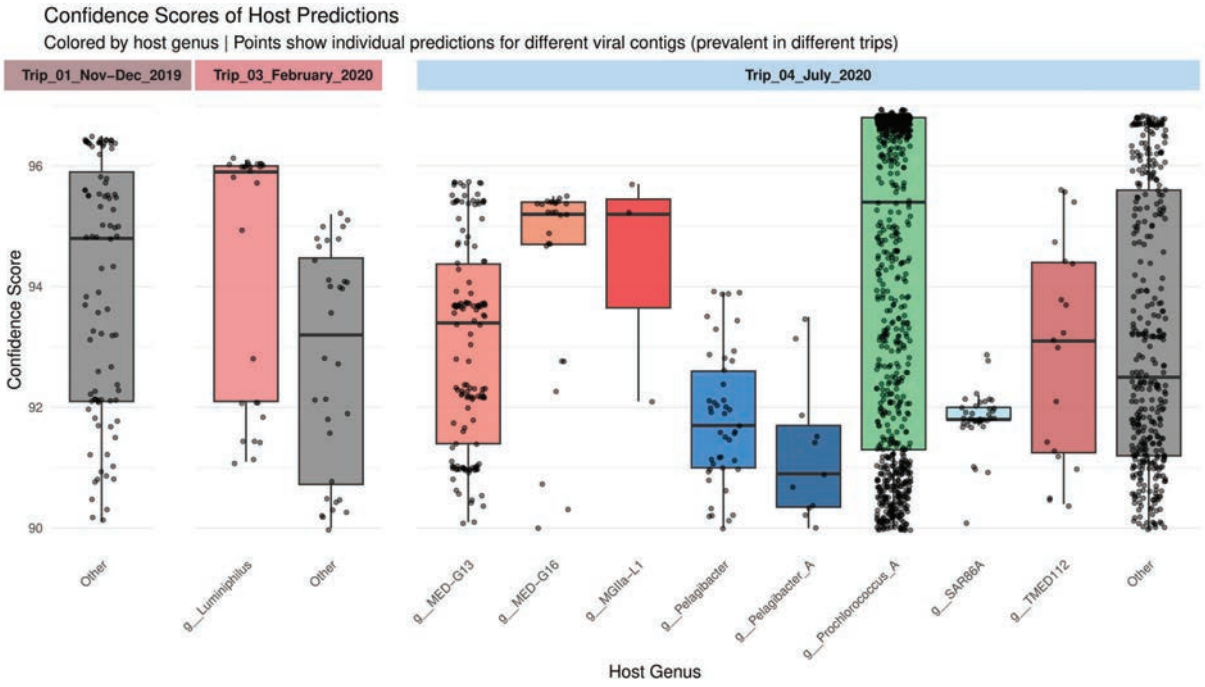

Figure S4. Predicted microbial hosts of seasonally prevalent viral contigs. Boxplots show the distribution of iPHoP confidence scores for host predictions at the genus level, stratified by the season (summer or winter) in which the viral contig was most abundant (as inferred from **Fig. S3**). The main goal was to identify the putative hosts of summer- and winter-specific viral contigs.

In summary, our results demonstrate a strong coupling between viral and prokaryotic community dynamics driven by seasonal succession (**Fig. S1**; above). Both microbial and viral communities cluster primarily by season (**Fig. S1**; above), with distinct summer- and winter-dominant viral contigs (**Fig. S3**) infecting the most abundant seasonal prokaryotic hosts, such as *Pelagibacter*, SAR86, and *Prochlorococcus* in winter, and *Luminiphilus* in summer (**Figs. S2, S4**). This tight ecological linkage suggests viral activity is a key factor shaping and responding to the seasonal turnover of microbial hosts. Lastly, variation explained by spatiotemporal proximity was greater than that explained by reef zoning status (**Fig. S1**).

DIABLO | **D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent c**O**mponents

Integrating microbial and viral data to identify multi-omics signatures discriminating between reefs that are open or closed to fishing.

Parameter choice

Design matrix

The choice of the design can be motivated by different aspects, including:

- Biological *a priori* knowledge: Should we expect seawater microbes (pMAGs) and viruses to be highly correlated?

- Analytical aims: As further developed in Singh et al. (2019)[386], a compromise needs to be achieved between a classification and prediction task, and extracting the correlation structure of the data sets. A full design with weights = 1 will favour the latter, but at the expense of classification accuracy, whereas a design with small weights (closer to 0) will lead to a highly predictive signature.

As we are more interested in predicting the zoning status and potentially increasing the current accuracy of 72% (achieved with MINT sPLS-DA for pMAGs and viruses separately), here we choose a design with smaller weights (0.1 weighted model).

However, we will still unravel a correlated signature even with this design, as we require both data sets (pMAGs, viruses) to explain the same outcome Y (zoning status), as well as maximising pairs of covariances between data sets.

Tuning the number of DIABLO components

As in the PLS-DA framework, we first fit a block.plsda model without variable selection to assess the global performance of the model and choose the number of components to retain. We run ***perf()*** with 5-fold cross validation repeated 50 times for up to 15 components and with our specified design matrix. Similar to PLS-DA, we obtain the performance of the model with respect to the different prediction distances. Results suggest a large number of components needs to be retained, and the model does not stabilise (**Fig. S5; Table S2**).

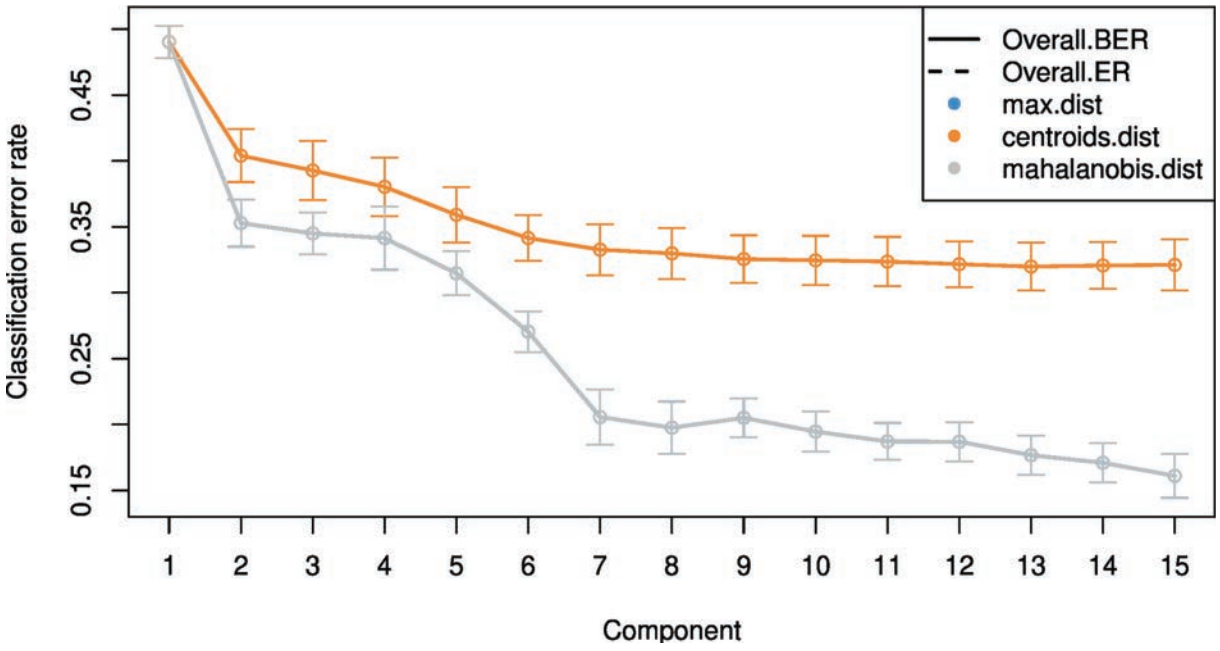

Figure S5. DIABLO: tuning the number of components. Choosing the number of components in block.plsda using perf() with 50× 5-fold CV function in the IMOS GBR-MGD dataset, with two omics blocks: seawater pMAGs and viruses. Classification error rates (overall and balanced) are represented on the y-axis with respect to the number of components on the x-axis for each prediction distance (as in PLS-DA). Bars show the standard deviation across the 50 repeated folds. The plot shows that the error rate reaches a minimum at 14-15 dimensions. The performance plot indicates that two components should be sufficient in the final model, and that the mahalanobis and max distances might lead to better prediction. As our design is balanced (i.e. the same number of No-Take and Take reefs), both overall and balanced error rates (BER) can be considered for further analysis. We will choose BER.

Table S1. DIABLO: optimal number of components. Table output is according to the prediction distance and type of error rate (overall or balanced), as well as a prediction weighting scheme.

	max.dist	centroids.dist	mahalanobis.dist
Overall error	15	11	15
Overall balanced error	15	11	15

Tuning the number of variables to select

We then choose the optimal number of variables to select in each data set using the tune.block.splsda function. The function tune() is run with a 5-fold cross validation, but repeated only 10 times (nrepeat = 10) for illustrative and computational reasons here. For a thorough tuning process, we advise increasing the nrepeat argument to 10–50, or more. We choose a keepX grid that is relatively fine at the start, then coarse. If the data sets are easy to classify, the tuning step may indicate the smallest number of variables to separate the sample groups. Hence, we start our grid at the value 5 to avoid a too small signature that may preclude biological interpretation.

The number of features to select on each component is returned and stored for the final DIABLO model: list.keepX.DIABLO <- tune.diablo$choice.keepX

Table S2. DIABLO: Optimal number of features as per tune(). Features are selected in block.splsda in each of the blocks, and for each component.

	comp1	comp2	comp3	comp4	comp5	comp6	comp7	comp8	comp9	comp10
pMAGs	10	7	65	9	5	40	25	7	8	6
Virus	5	5	56	16	26	9	8	90	10	98

Final model

In summary, the final DIABLO model was run with 10 components, using a design matrix weight of 0.1. The number of features selected per component for each data block was determined by the tuning

step, resulting in the selection of 10, 7, 65, 9, 5, 40, 25, 7, 8, and 6 features for the pMAGs block; and 5, 5, 56, 16, 26, 9, 8, 90, 10, and 98 features for the virus block, across the respective components.

DIABLO graphs

Sample plots

plotDiablo() is a diagnostic plot to check whether the correlations between components from each data set were maximised as specified in the design matrix. We specify the dimension to be assessed with the ncomp argument.

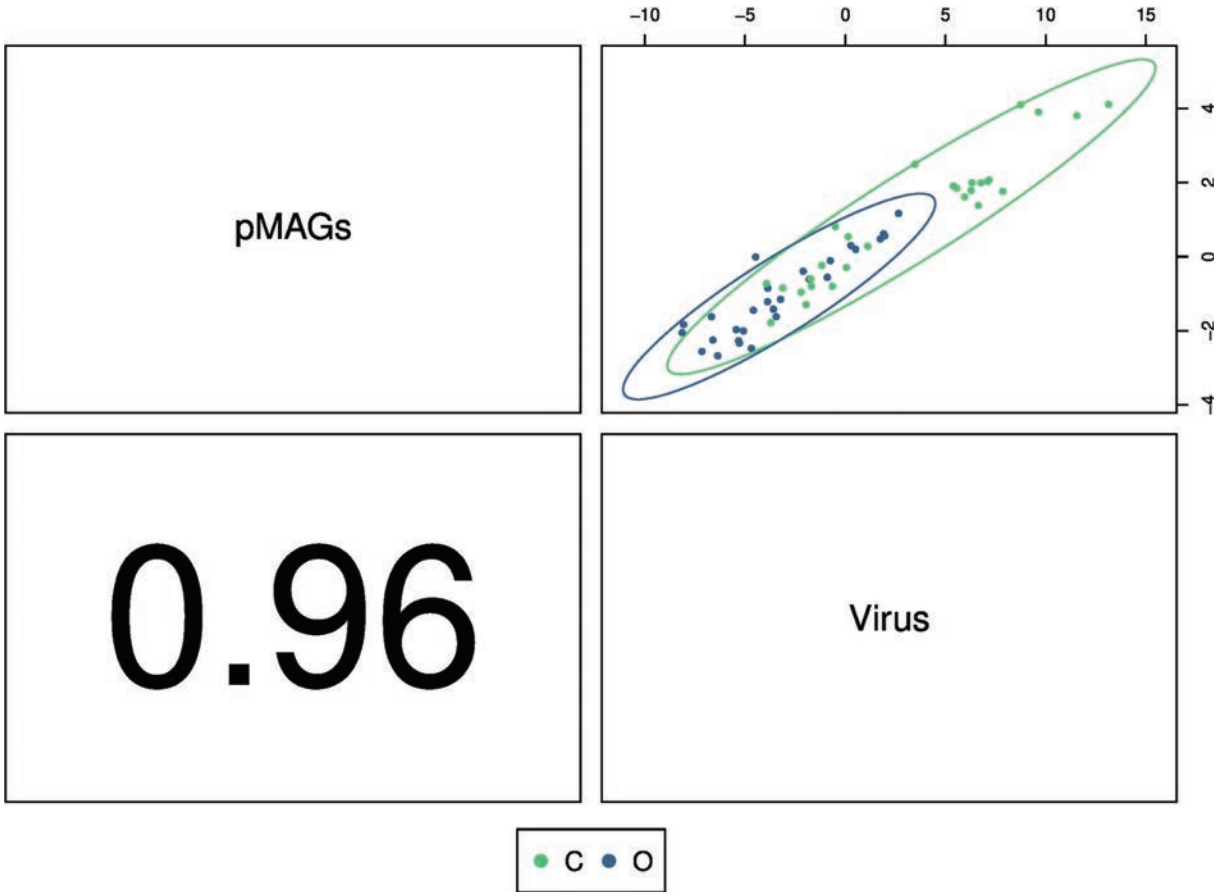

Figure S6. Diagnostic plot from multiblock sPLS-DA. Samples are represented based on the specified component (here ncomp = 1) for each data set (pMAGs and Virus). Samples are coloured by reef zoning status (green for NTMR reefs - closed to fishing, and blue for fished reefs) and 95% confidence ellipse plots are represented. The bottom left numbers indicate the correlation coefficients between the first components from both data sets. In this example, abundances of pMAGs and viruses are highly correlated in the first dimension.

plotIndiv()

These sample plots project each sample into the space spanned by the components from each block, resulting in a series of graphs corresponding to each data set. This type of graphic allows us to

better understand the information extracted from each data set and its discriminative ability. Here we see that, while NTMRs and fished reefs do cluster separately, reefs also cluster based on time of sampling and geographic proximity (i.e., Sampling trip), suggesting spatiotemporal confounding effects in the DIABLO model.

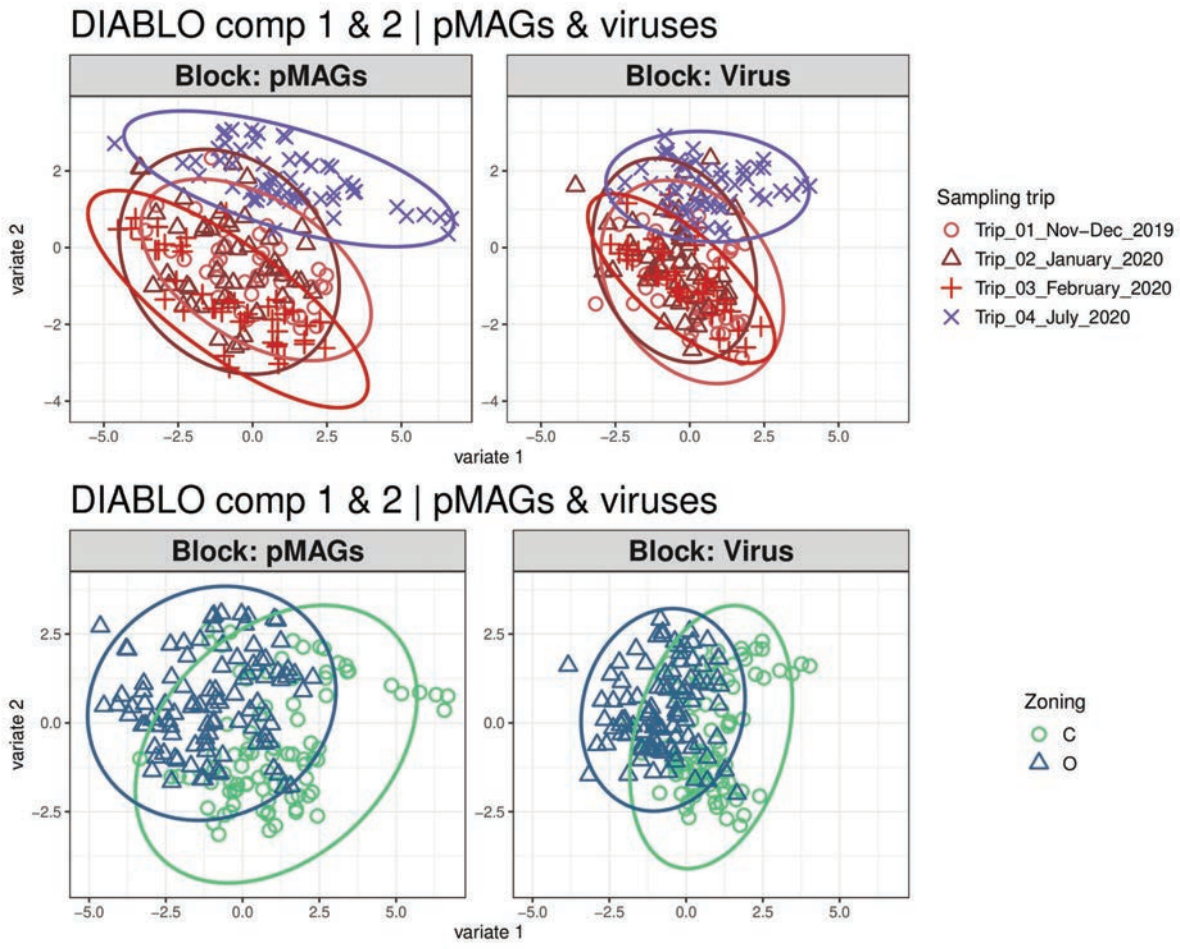

Figure S7. Sample plots from DIABLO. Block.splsda was performed with seawater pMAGs and viruses as two omics blocks (X), used to discriminate reef zoning status as categorical Y. The samples are visualised according to their scores on the first two components for both data sets. Samples are coloured by sampling trip (above) and reef zoning status (below). The plot shows the degree of agreement between the different omics data sets and the discriminative ability of each data set.

plotArrow()

In the arrow plot, the start of the arrow indicates the centroid between both data sets for a given sample and the tip of the arrow the location of that same sample but in each omics block. Such graphics highlight the agreement between all data sets at the sample level when modelled with multiblock sPLS-DA. Similar to the sample plots above (**Fig. S7**), we also observe that while the model does discriminate

between NTMRs and fished reefs (**Fig. S8**; left), spatiotemporal batch effects inherent to sampling trips are also observed (**Fig. S8**; right).

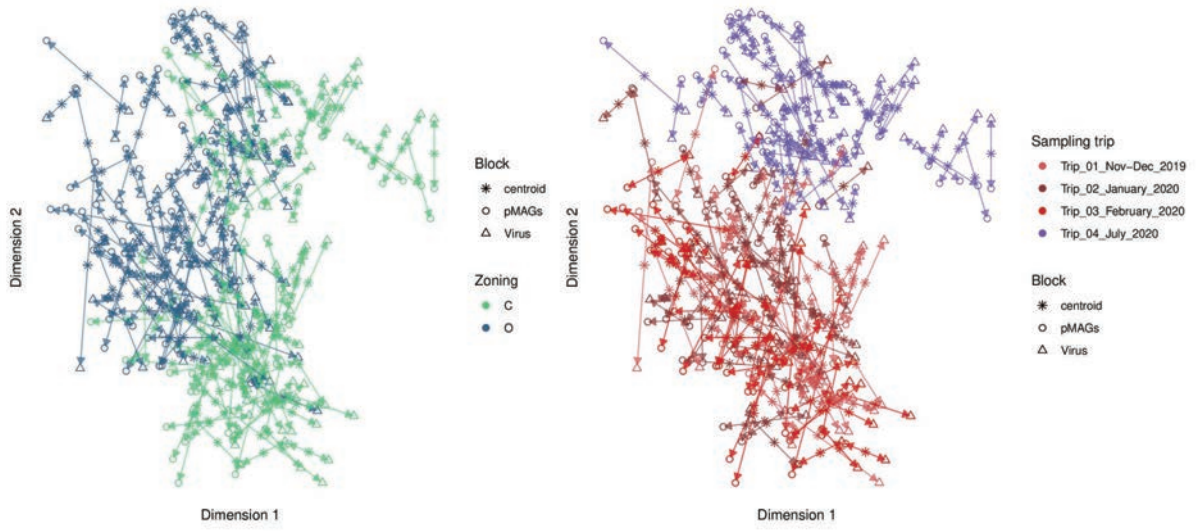

Figure S8. DIABLO Arrow plot. Multiblock sPLS-DA was performed on the IMOS GBR-MGD data: with seawater pMAGs and viruses as two omics blocks (X), used to discriminate reef zoning status as categorical Y. The samples are projected into the space spanned by the first two DIABLO components for each data set then overlaid across data sets. The start of the arrow indicates the centroid between all data sets for a given sample and the tip of the arrow the location of the same sample in each block. Arrows further from their centroid indicate some disagreement between the data sets. Samples are coloured by reef zoning status (left) of sampling trip (right).

Variable plots

The visualisation of the selected variables is crucial to mine their associations in multiblock sPLS-DA. All the plots presented provide complementary information for interpreting the results.

plotVar()

The correlation circle plot highlights the contribution of each selected variable to each component. Important variables should be close to the large circle. Here, only the variables selected on components 1 and 2 are depicted (across all blocks). Clusters of points indicate a strong correlation between variables. For better visibility we chose to hide the variable names.

The correlation circle plot (**Fig. S9**) shows some positive correlations (between selected pMAGs and viruses) and negative correlations between pMAGs and viruses on component 1. The correlation structure is less obvious on component 2, but we observe some key selected features (pMAGs and viruses) that seem to highly contribute to component 2.

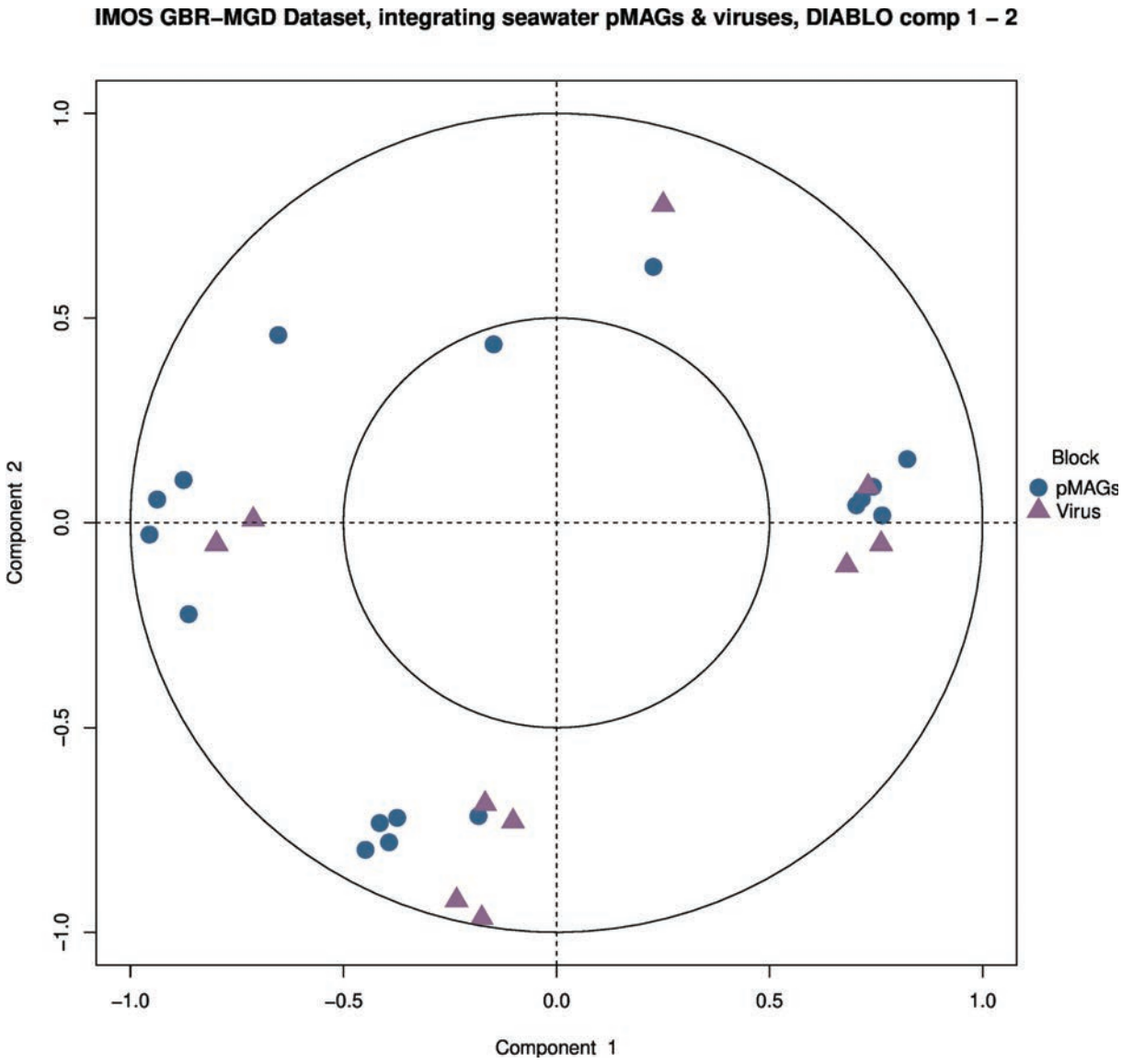

Figure S9. Correlation circle plot from DIABLO. Multiblock sPLS-DA was performed on the IMOS GBR-MGD data: with seawater pMAGs and viruses as two omics blocks (X), used to discriminate reef zoning status as categorical Y. The variable coordinates are defined according to their correlation with the first and second components for each data set. Variable types are indicated with different symbols and colours, and are overlaid on the same plot. The plot highlights the potential associations within and between different variable types when they are important in defining their own component.

circosPlot()

The circos plot represents the correlations between variables of different types to show within and between connections between blocks (pMAGs and Virus), and expression levels of each variable according to each class (NTMRs and fished reefs). The circos plot is built based on a similarity matrix, and a cutoff

argument can be further included to visualise correlation coefficients above this threshold in the multi-omics signature - this will be important considering the strong correlations between viruses and pMAGs.

The circos plot (**Fig. S10**) enables us to visualise cross-correlations between omics layers, and the nature of these correlations: positive (red) or negative (blue). Here we observe that correlations >0.7 are between a few seawater pMAGs and some viruses, as well as some (negative) correlations. The lines indicating the average enrichment levels per reef category (NTMRs and fished zones) indicate that the selected features are able to discriminate the sample groups.

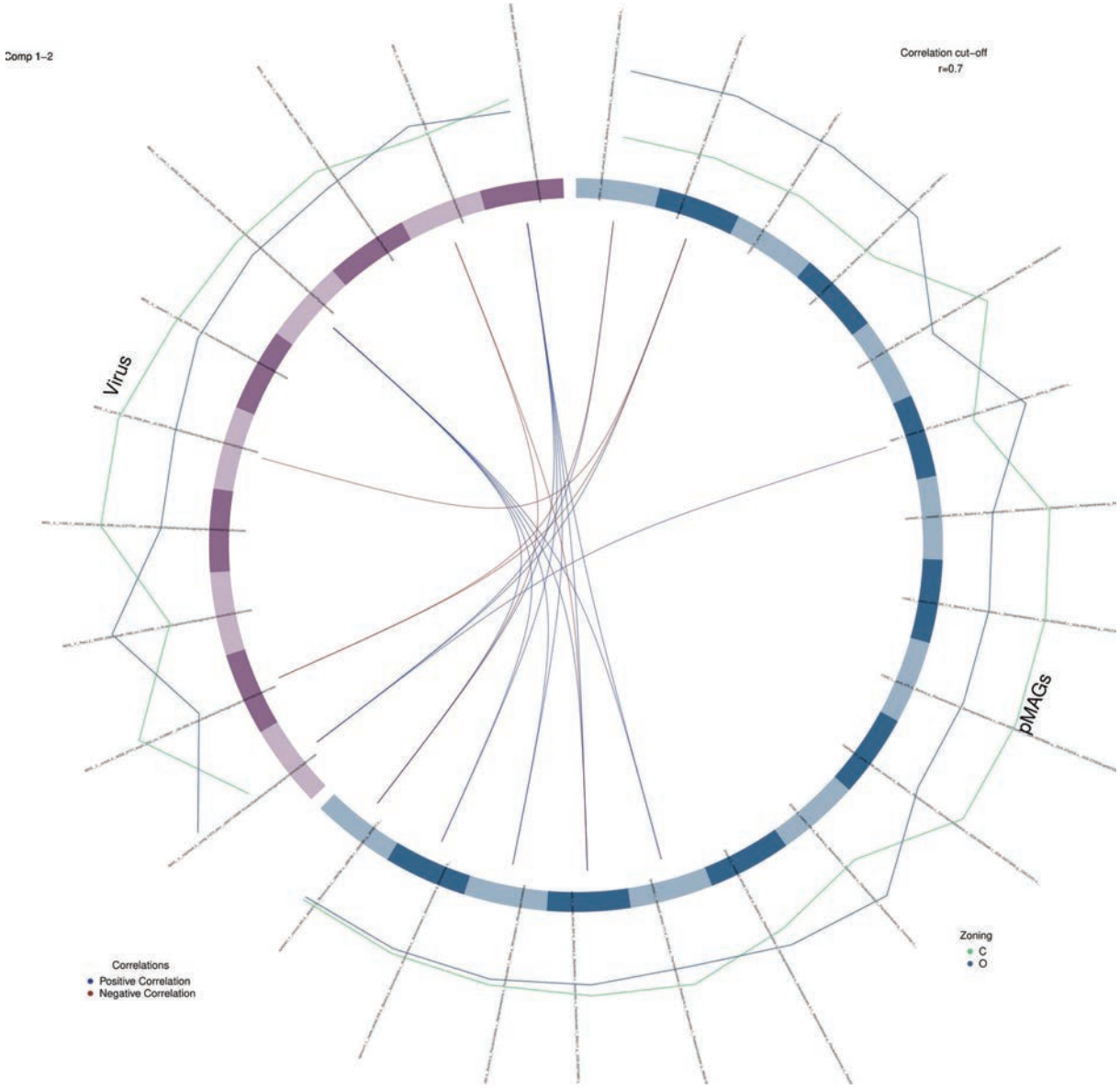

Figure S10. DIABLO Circos plot. Circos plot from multiblock sPLS-DA performed on the IMOS GBR-MGD data: with seawater pMAGs and viruses as two omics blocks (X), used to discriminate reef zoning status as categorical Y. The plot represents the correlations greater than 0.7 between variables of different types, represented on the side quadrants. The internal connecting lines show the positive (red) and negative (blue) correlations. The

outer lines show the enrichment levels of each variable (indicator pMAGs and viruses) in each sample group (No-take and Take reefs).

cimDiablo()

Similar to a classical hierarchical clustering, the cimDiablo() function is a clustered image map (heatmap) specifically implemented to represent the multi-omics molecular signature for each sample.

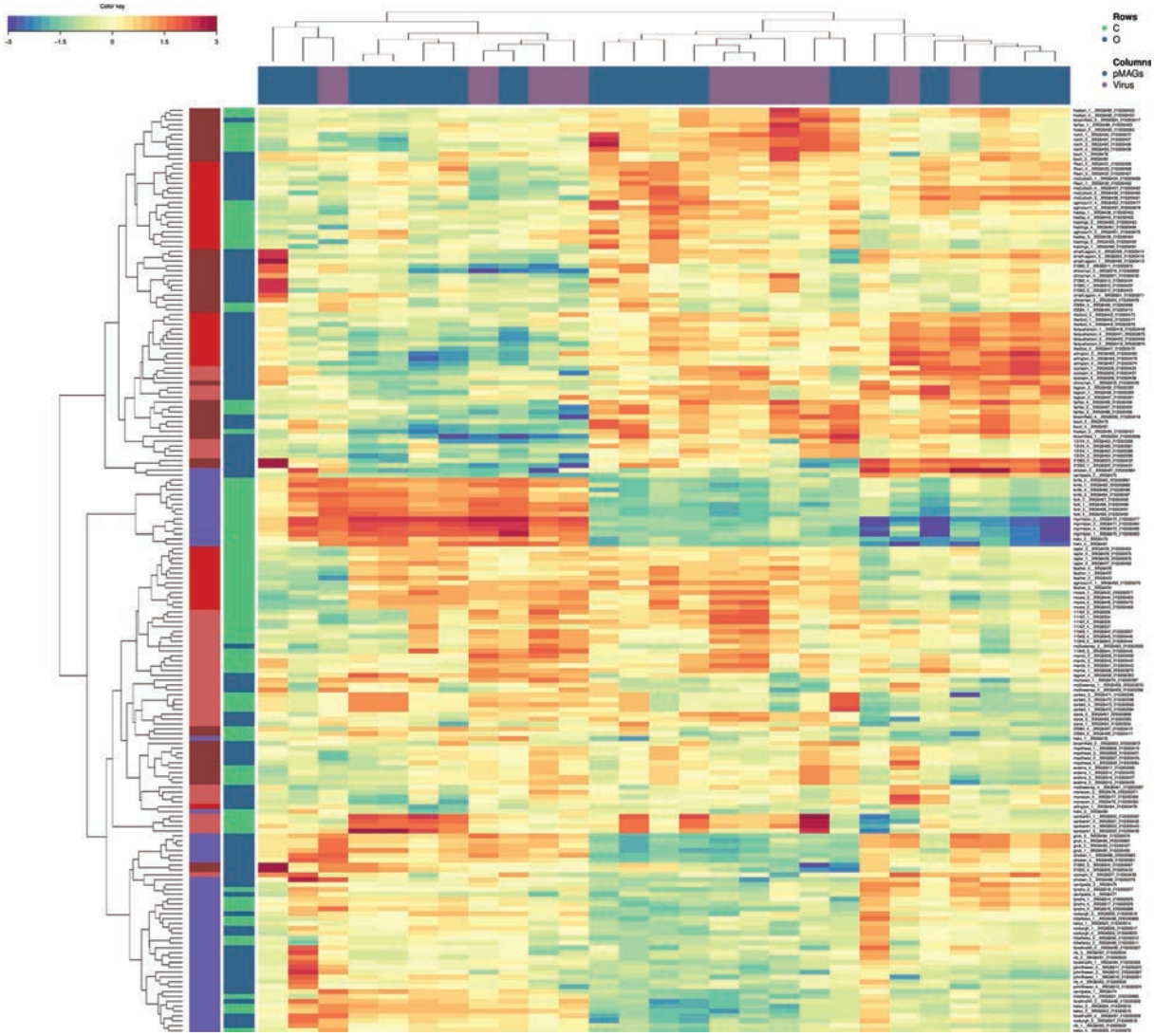

Figure S11. DIABLO Clustered Image Map (CIM). Variables are selected by multiblock sPLS-DA on component 1, performed on the IMOS GBR-MGD data: with seawater pMAGs and viruses as two omics blocks (X), used to discriminate reef zoning status as categorical Y. By default, Euclidean distance and Complete linkage methods are used. The CIM represents samples in rows (indicated by their reef zoning status and sampling trip on the left-hand side of the plot) and selected features in columns (indicated by their data type - pMAGs or Viruses, at the top of the plot).

Multivariate **INT**egration (**MINT**) **sPLS-DA** | Discriminating between reefs that are open or closed to fishing (sPLS-DA), while accounting for sector-specific effects (MINT)

pMAGs

Tuning the number of dimensions

The perf() function is used to estimate the performance of the MINT-sPLS-DA model using Leave One Group Out Cross Validation (LOGOCV, i.e. by training MINT sPLS-DA on six out of seven sectors and validating the model performance on the left-out subset, hence seven times until each of the seven GBR sectors is left out once), and to choose the optimal number of components for our final model.

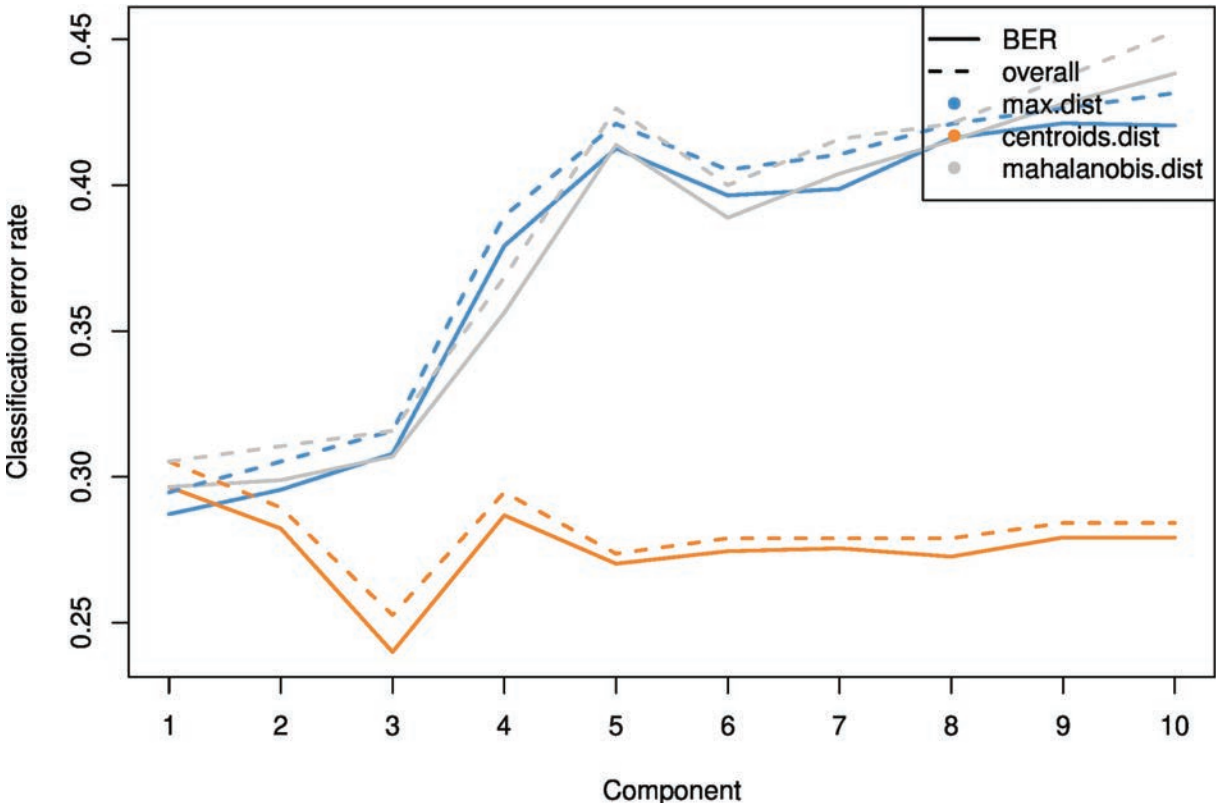

Fig S12. Choosing the number of components in mint.splsda using perf() with LOGOCV to discriminate between reefs that are open or closed to fishing, using a dataset of 876 IMOS GBR-MGD microbial genomes (pMAGs95%ANI). Classification error rates (overall and balanced - BER) are represented on the y-axis with respect to the number of components on the x-axis for each prediction distance. Overall and balanced error rates show largely the same trend as the design is balanced (i.e. the same number of NTMR and fished reefs in each GBR sector). The plot shows that the error rate reaches a minimum (~29%) with two or three dimensions with the centroids prediction distance. We therefore retained 2 PCs in downstream analysis.

Table S3. Numerical output associated with Fig. S1. Here, we show overall MINT sPLS-DA error rates when discriminating between reefs that are open or closed to fishing (Y), using 876 IMOS GBR-MGD pMAGs95%ANI as X, using 3 prediction distances (max.dist, centroids.dist, mahalanobis.dist), and across the seven GBR sectors. Sector-specific model accuracies were expressed as 1 – error (with centroids dist), and for each of the 10 tested MINT sPLS-DA components, we also show MINT sPLS-DA classification accuracy averaged across 7 GBR sectors.

MINT sPLS-DA comp	Study (GBR sector)	Max dist	Centroids dist	Mahalanobis dist	Accuracy (1 – error with centroids dist)	Average accuracy
comp1	01_Cape_Grenville	0.42	0.42	0.42	0.58	
comp1	02_Princess_Charlotte_bay	0.20	0.20	0.20	0.80	
comp1	03_Cairns	0.35	0.35	0.35	0.65	
comp1	04_Innisfail	0.07	0.07	0.07	0.93	0.71
comp1	05_Townsville	0.30	0.30	0.30	0.70	
comp1	06_Swains	0.35	0.35	0.35	0.65	
comp1	07_Capricorn_Bunker	0.32	0.32	0.32	0.68	
comp2	01_Cape_Grenville	0.58	0.42	0.54	0.58	
comp2	02_Princess_Charlotte_bay	0.67	0.20	0.73	0.80	
comp2	03_Cairns	0.10	0.25	0.10	0.75	
comp2	04_Innisfail	0.22	0.15	0.22	0.85	0.71
comp2	05_Townsville	0.32	0.27	0.32	0.73	
comp2	06_Swains	0.45	0.40	0.50	0.60	
comp2	07_Capricorn_Bunker	0.36	0.32	0.36	0.68	
comp3	01_Cape_Grenville	0.63	0.38	0.63	0.63	
comp3	02_Princess_Charlotte_bay	0.60	0.27	0.60	0.73	
comp3	03_Cairns	0.25	0.25	0.25	0.75	
comp3	04_Innisfail	0.30	0.15	0.30	0.85	0.71
comp3	05_Townsville	0.30	0.29	0.30	0.71	
comp3	06_Swains	0.35	0.40	0.35	0.60	
comp3	07_Capricorn_Bunker	0.18	0.29	0.21	0.71	
comp4	01_Cape_Grenville	0.63	0.42	0.63	0.58	
comp4	02_Princess_Charlotte_bay	0.93	0.27	0.93	0.73	
comp4	03_Cairns	0.25	0.30	0.25	0.70	
comp4	04_Innisfail	0.48	0.22	0.48	0.78	0.68
comp4	05_Townsville	0.39	0.29	0.39	0.71	
comp4	06_Swains	0.45	0.40	0.45	0.60	
comp4	07_Capricorn_Bunker	0.29	0.36	0.29	0.64	
comp5	01_Cape_Grenville	0.71	0.42	0.71	0.58	
comp5	02_Princess_Charlotte_bay	1.00	0.33	1.00	0.67	
comp5	03_Cairns	0.30	0.30	0.30	0.70	0.67
comp5	04_Innisfail	0.48	0.19	0.48	0.81	
comp5	05_Townsville	0.38	0.29	0.38	0.71	
comp5	06_Swains	0.45	0.40	0.45	0.60	

comp5	07_Capricorn_Bunker	0.21	0.36	0.21	0.64	
comp6	01_Cape_Grenville	0.63	0.42	0.63	0.58	0.67
comp6	02_Princess_Charlotte_bay	0.80	0.40	0.80	0.60	
comp6	03_Cairns	0.30	0.30	0.30	0.70	
comp6	04_Innisfail	0.41	0.19	0.41	0.81	
comp6	05_Townsville	0.34	0.29	0.34	0.71	
comp6	06_Swains	0.40	0.40	0.40	0.60	
comp6	07_Capricorn_Bunker	0.25	0.36	0.25	0.64	
comp7	01_Cape_Grenville	0.63	0.46	0.63	0.54	0.66
comp7	02_Princess_Charlotte_bay	0.87	0.40	0.87	0.60	
comp7	03_Cairns	0.20	0.30	0.20	0.70	
comp7	04_Innisfail	0.41	0.19	0.41	0.81	
comp7	05_Townsville	0.34	0.29	0.34	0.71	
comp7	06_Swains	0.45	0.40	0.45	0.60	
comp7	07_Capricorn_Bunker	0.32	0.36	0.32	0.64	
comp8	01_Cape_Grenville	0.63	0.46	0.63	0.54	0.65
comp8	02_Princess_Charlotte_bay	0.93	0.40	0.93	0.60	
comp8	03_Cairns	0.20	0.30	0.20	0.70	
comp8	04_Innisfail	0.44	0.19	0.44	0.81	
comp8	05_Townsville	0.32	0.29	0.32	0.71	
comp8	06_Swains	0.35	0.45	0.35	0.55	
comp8	07_Capricorn_Bunker	0.32	0.36	0.32	0.64	
comp9	01_Cape_Grenville	0.67	0.46	0.67	0.54	0.65
comp9	02_Princess_Charlotte_bay	0.93	0.40	0.93	0.60	
comp9	03_Cairns	0.20	0.30	0.20	0.70	
comp9	04_Innisfail	0.44	0.19	0.52	0.81	
comp9	05_Townsville	0.30	0.29	0.30	0.71	
comp9	06_Swains	0.35	0.45	0.40	0.55	
comp9	07_Capricorn_Bunker	0.43	0.36	0.39	0.64	
comp10	01_Cape_Grenville	0.75	0.46	0.75	0.54	0.67
comp10	02_Princess_Charlotte_bay	0.80	0.33	0.80	0.67	
comp10	03_Cairns	0.20	0.30	0.20	0.70	
comp10	04_Innisfail	0.56	0.19	0.56	0.81	
comp10	05_Townsville	0.29	0.29	0.29	0.71	
comp10	06_Swains	0.50	0.45	0.55	0.55	
comp10	07_Capricorn_Bunker	0.36	0.29	0.39	0.71	

Table S4. MINT sPLS-DA - error rate (centroids distance) across GBR sectors, and separately for C (reefs closed to fishing) and O (open to fishing).

	Study	comp1	comp2	comp3	comp4	comp5	comp6	comp7	comp8	comp9	comp10
C	01_Cape_Grenville	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
O	01_Cape_Grenville	0.25	0.25	0.17	0.25	0.25	0.25	0.33	0.33	0.33	0.33
C	02_Princess_Charlotte_bay	0.00	0.00	0.13	0.13	0.25	0.38	0.38	0.38	0.38	0.25
O	02_Princess_Charlotte_bay	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
C	03_Cairns	0.25	0.00	0.00	0.13	0.13	0.13	0.13	0.13	0.13	0.13
O	03_Cairns	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
C	04_Innisfail	0.07	0.13	0.13	0.20	0.13	0.13	0.13	0.13	0.13	0.13
O	04_Innisfail	0.08	0.17	0.17	0.25	0.25	0.25	0.25	0.25	0.25	0.25
C	05_Townsville	0.39	0.32	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
O	05_Townsville	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
C	06_Swains	0.25	0.38	0.38	0.38	0.50	0.50	0.50	0.50	0.50	0.50
O	06_Swains	0.42	0.42	0.42	0.42	0.33	0.33	0.33	0.42	0.42	0.42
C	07_Capricorn_Bunker	0.31	0.31	0.25	0.38	0.38	0.38	0.38	0.38	0.38	0.25
O	07_Capricorn_Bunker	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Average error [C]		0.27	0.25	0.26	0.31	0.33	0.35	0.35	0.35	0.35	0.31
Average error [O]		0.31	0.32	0.31	0.33	0.32	0.32	0.33	0.34	0.34	0.34
Average accuracy [C]		0.73	0.75	0.74	0.69	0.67	0.65	0.65	0.65	0.65	0.69
Average accuracy [O]		0.69	0.68	0.69	0.67	0.68	0.68	0.67	0.66	0.66	0.66

Tuning the number of features per dimension

We can choose the keepX parameter using the tune() function for a MINT object. The function performs LOGOCV for different values of test.keepX (we specified test.keepX = seq(10, 300, 10)) provided on each component (we tested 5 components), and no repeat argument is needed. Based on the mean classification error rate (overall error rate or BER) and a centroids distance, we output the optimal number of variables keepX to be included in the final model.

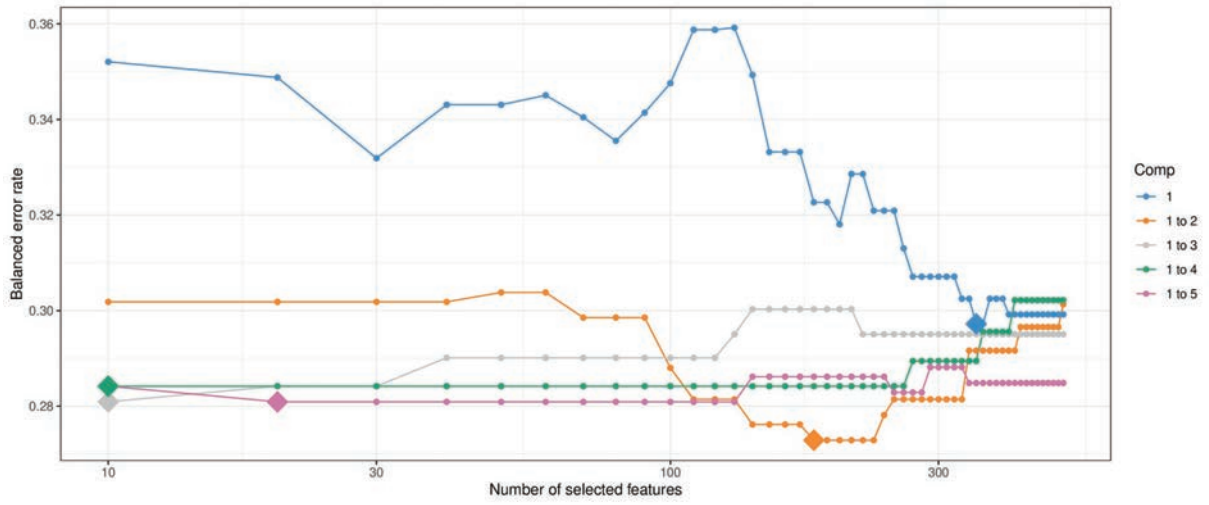

Figure S13. Tuning plot of the MINT sPLS-DA models with up to 5 components, testing a grid value of 10 to 500 indicators (with sequential increases of 10). Diamonds represent the optimal number of features on a given component. Balanced error rate found on the vertical axis and is the metric to be minimised.

Table S5. Numerical output associated with Fig S9, also showing sector-specific MINT sPLS-DA classification errors and average error/accuracy across sectors.

GBR_sector	comp1	comp2	comp3	comp4	comp5
CA	0.33	0.27	0.21	0.27	0.27
CB	0.32	0.32	0.32	0.35	0.35
CG	0.42	0.42	0.46	0.42	0.46
IN	0.18	0.23	0.23	0.23	0.23
PC	0.21	0.21	0.21	0.21	0.21
SW	0.44	0.4	0.44	0.44	0.44
TO	0.29	0.29	0.29	0.29	0.29
Average Error	0.31	0.31	0.31	0.32	0.32
Accuracy (1 – average error)	0.69	0.69	0.69	0.68	0.68

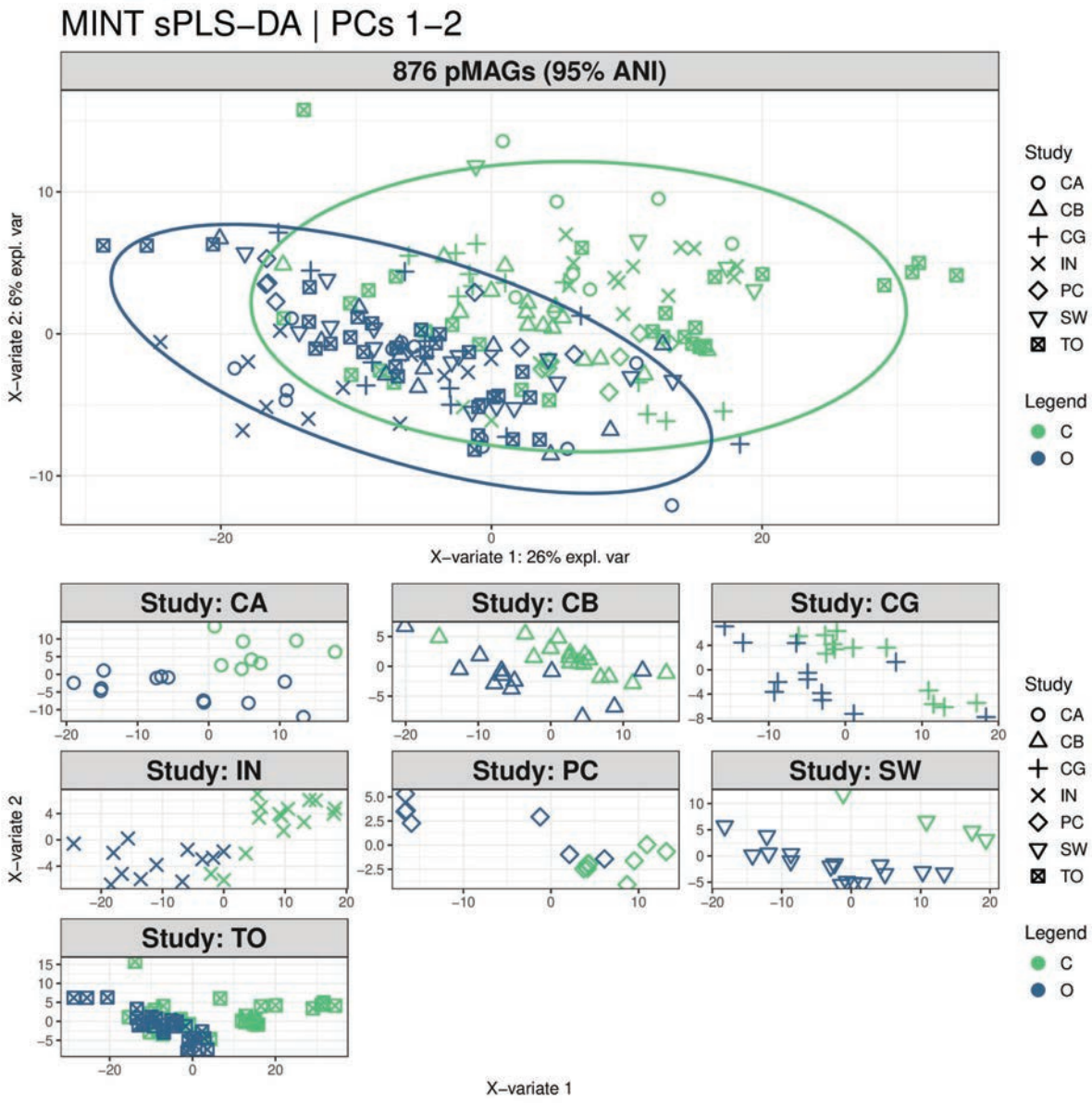

Figure S14. Sample plots from the MINT sPLS-DA performed on the 876 IMOS GBR-MGD seawater pMAGs95%ANI, aiming to find discriminatory microbes between reefs that are open or closed to fishing. Samples (48 reef sites x 4 replicates) are projected into the space spanned by the first two components. Reef sites are coloured by their protection level (open or closed to fishing) and symbols indicate the membership of reef sites to their corresponding LTMP trip/transect. (**top**) Global components from the model with 95% ellipse confidence intervals around each sample class. (**bottom**) Partial components per study show a good agreement across GBR sectors. Component 1 discriminates between reefs that are open or closed to fishing.

Performance of the final MINT sPLS-DA model

Use of the auroc() function will yield a visualisation of classification performance when undergoing the LOGOCV procedure from above. The interpretation of this output may not be particularly insightful in relation to the performance evaluation of mixOmics methods, but can complement the statistical analysis. For example, the MINT sPLS-DA classification of fished vs. NTMR sites had ~73 % accuracy in classifying samples in their corresponding zoning category.

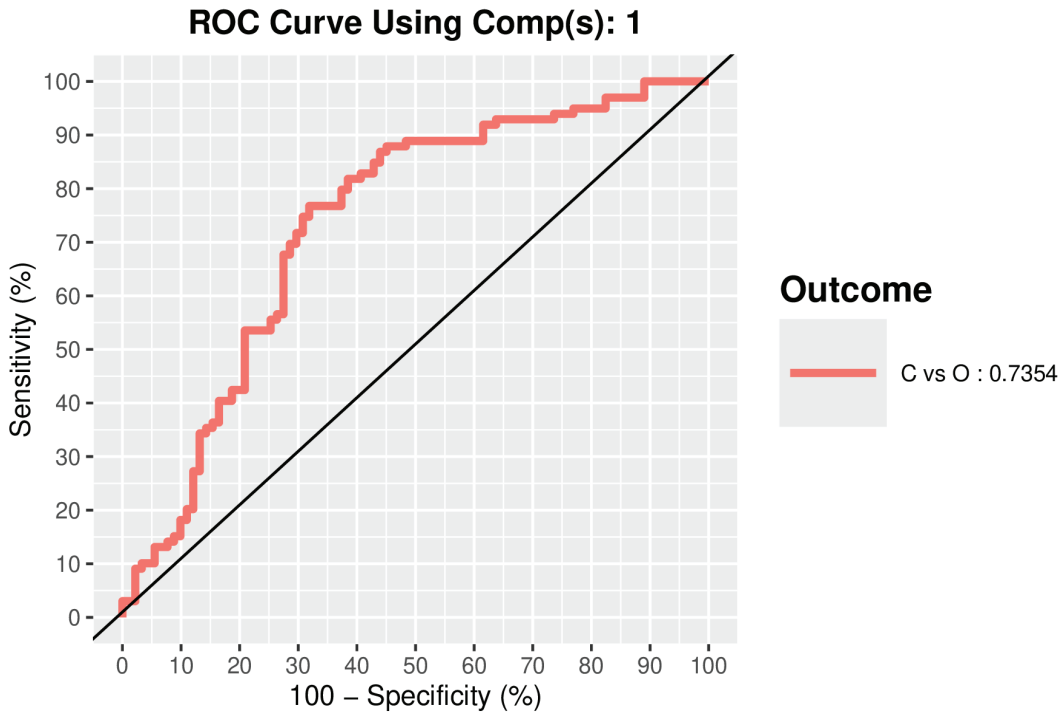

Fig. S15. ROC curve and AUC from the MINT sPLS-DA performed on the IMOS MGD MAGs (876 pMAGs95%ANI) for global component 1 for the fished vs. NTMRs reefs comparison. Numerical outputs include the AUC (0.7312) and a Wilcoxon test p–value (p = 3.69 × 10−8) for fished vs. NTMRs reefs class comparison that are performed per component.

Multivariate INTegration (MINT) sPLS-DA | Discriminating between reefs that are open or closed to fishing (sPLS-DA), while accounting for sector-specific effects (MINT)

Virus

To account for these confounding effects (seawater collections at a single time point across locations), we implemented a Multivariate INTegration Sparse Partial Least Squares Discriminant Analysis - MINT sPLS-DA[2,291,294] to stratify the IMOS GBR-MGD dataset by sector and identify viral contigs that consistently discriminate NTMRs from fished reefs across Great Barrier Reef sectors, while

removing sector-specific batch effects. MINT sPLS-DA was run on CLR-transformed viral abundance data with model tuning performed via Leave-One-Group-Out Cross-Validation (i.e. training the model on six sectors and validating on the left-out seventh; iterated seven times across sectors) to determine (1) the optimal number of components; and (2) select the most informative viral features per component.

Tuning the number of dimensions

The perf() function is used to estimate the performance of the MINT-sPLS-DA model using Leave One Group Out Cross Validation (LOGOCV, i.e. by training MINT sPLS-DA on six out of seven sectors and validating the model performance on the left-out subset, hence seven times until each of the seven GBR sectors is left out once), and to choose the optimal number of components for our final model (Fig. S16).

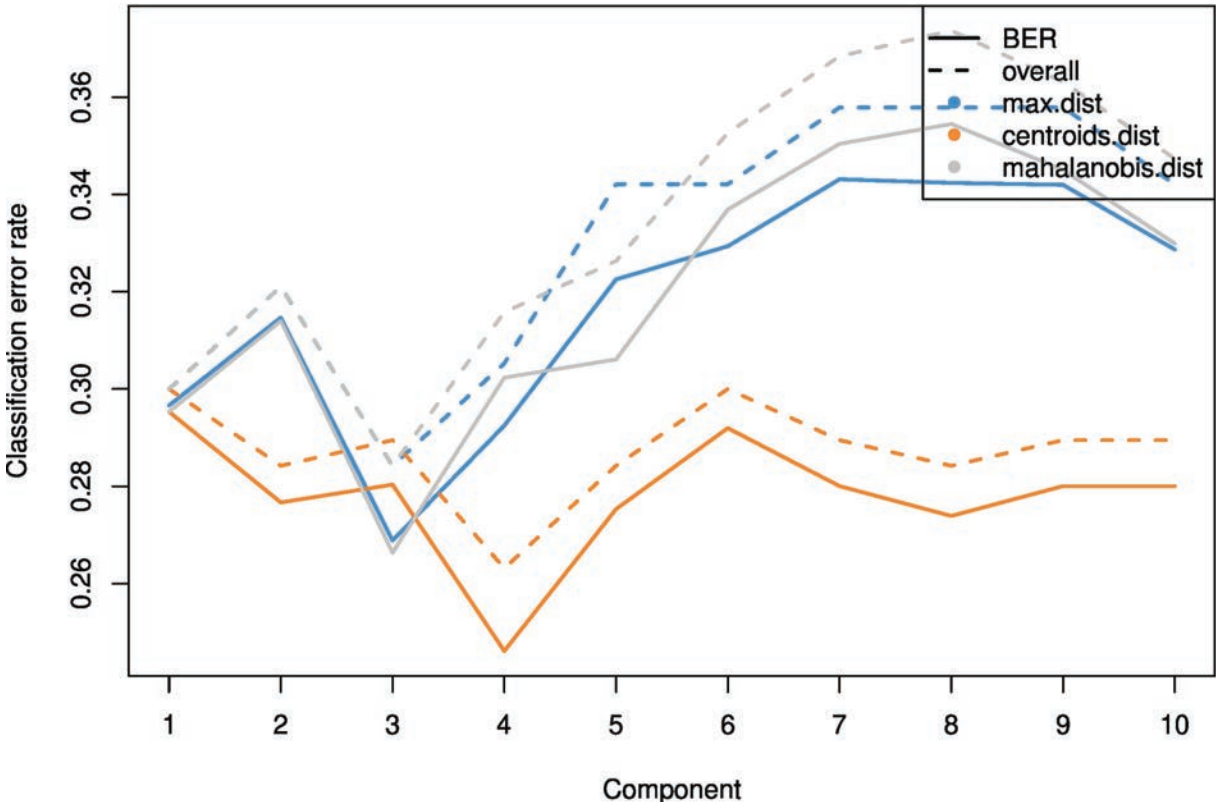

Figure S16. Choosing the number of components in mint.splsda using perf() with LOGOCV to discriminate between reefs that are open or closed to fishing, using a dataset of 6,024 IMOS GBR-MGD viral contigs with taxonomic annotations (>3kb; prevalence of >75% of samples within NTMRs and fished reefs; average rel. abund > 0.001%; with GENOMAD annotations). Classification error rates (overall and balanced - BER) are represented on the y-axis with respect to the number of components on the x-axis for each prediction distance. Overall and balanced error rates show largely the same trend as the design is balanced (i.e. the same number of NTMR and fished reefs in each GBR sector). The plot shows that the error rate reaches a minimum (~28%) with two dimensions with the centroids prediction distance. We therefore retained 2 PCs in downstream analysis.

Table S6. Numerical results for Figure S1. Sector-specific classification error rates in MINT sPLS-DA (X = 6,024 viral contigs; Y = reef zoning; study = GBR sector) based on 3 prediction distances (max.dist, centroids.dist, mahalanobis.dist), shown for the first 2 MINT sPLS-DA components. We also show MINT sPLS-DA classification accuracy averaged across 7 GBR sectors. Model accuracies were expressed as 1 – average error (for each prediction distance).

	GBR sector	Max dist	Centroids dist	Mahalanobis dist
comp1	CA	0.10	0.10	0.10
comp2	CA	0.15	0.10	0.15
comp1	CB	0.25	0.25	0.25
comp2	CB	0.21	0.25	0.21
comp1	CG	0.63	0.63	0.63
comp2	CG	0.54	0.63	0.54
comp1	IN	0.11	0.11	0.11
comp2	IN	0.22	0.11	0.22
comp1	PC	0.27	0.27	0.27
comp2	PC	0.33	0.27	0.33
comp1	SW	0.25	0.25	0.25
comp2	SW	0.30	0.25	0.30
comp1	TO	0.36	0.36	0.36
comp2	TO	0.43	0.38	0.43
Average error		0.30	0.28	0.30
Average accuracy		0.70	0.72	0.70

Tuning the number of features (viral contigs) per dimension

We can choose the keepX parameter using the tune() function for a MINT object. The function performs LOGOCV for different values of test.keepX provided on each component, and no repeat argument is needed. Based on the mean classification error rate (overall error rate or BER) and a centroids distance, we output the optimal number of variables keepX to be included in the final model. Here, we tested between 10 and 500 viral contigs in intervals of 10 (test.keepX = seq(10, 300, 10)) for 5 MINT sPLS-DA dimensions (ncomp = 5). For the final MINT sPLS-DA model, we retained the first two components with 240 and 110 features per dimension, as suggested by the *perf()* (see **Fig. S16**) and *tune()* functions (**Fig. S17**), respectively.

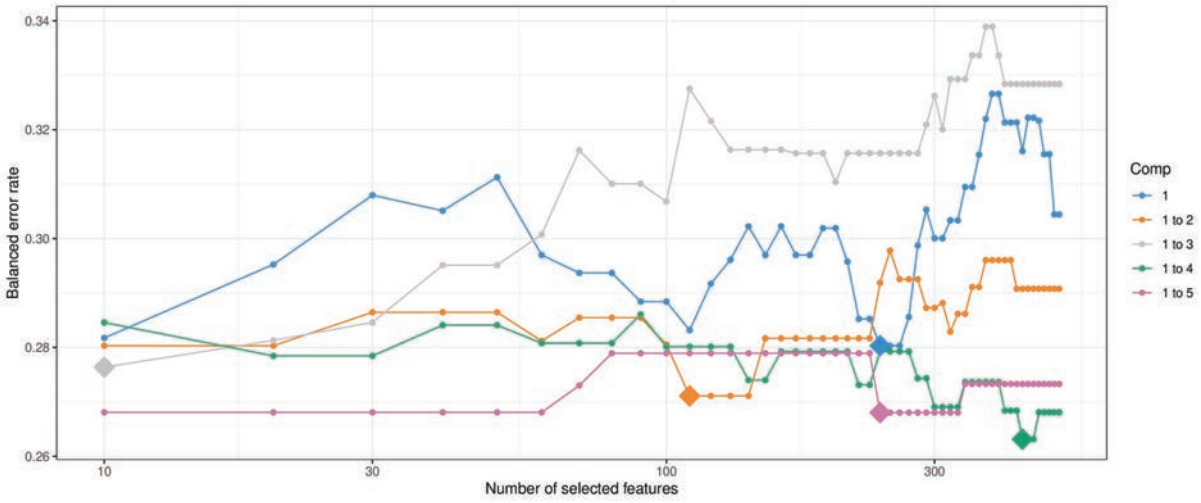

Figure S17. Tuning plot of the MINT sPLS-DA models with up to 5 components, testing a grid value of 10 to 500 viral indicators (with sequential increases of 10). Diamonds represent the optimal number of features on a given component. Balanced error rate found on the vertical axis and is the metric to be minimised.

Final MINT sPLS-DA model (Virus)

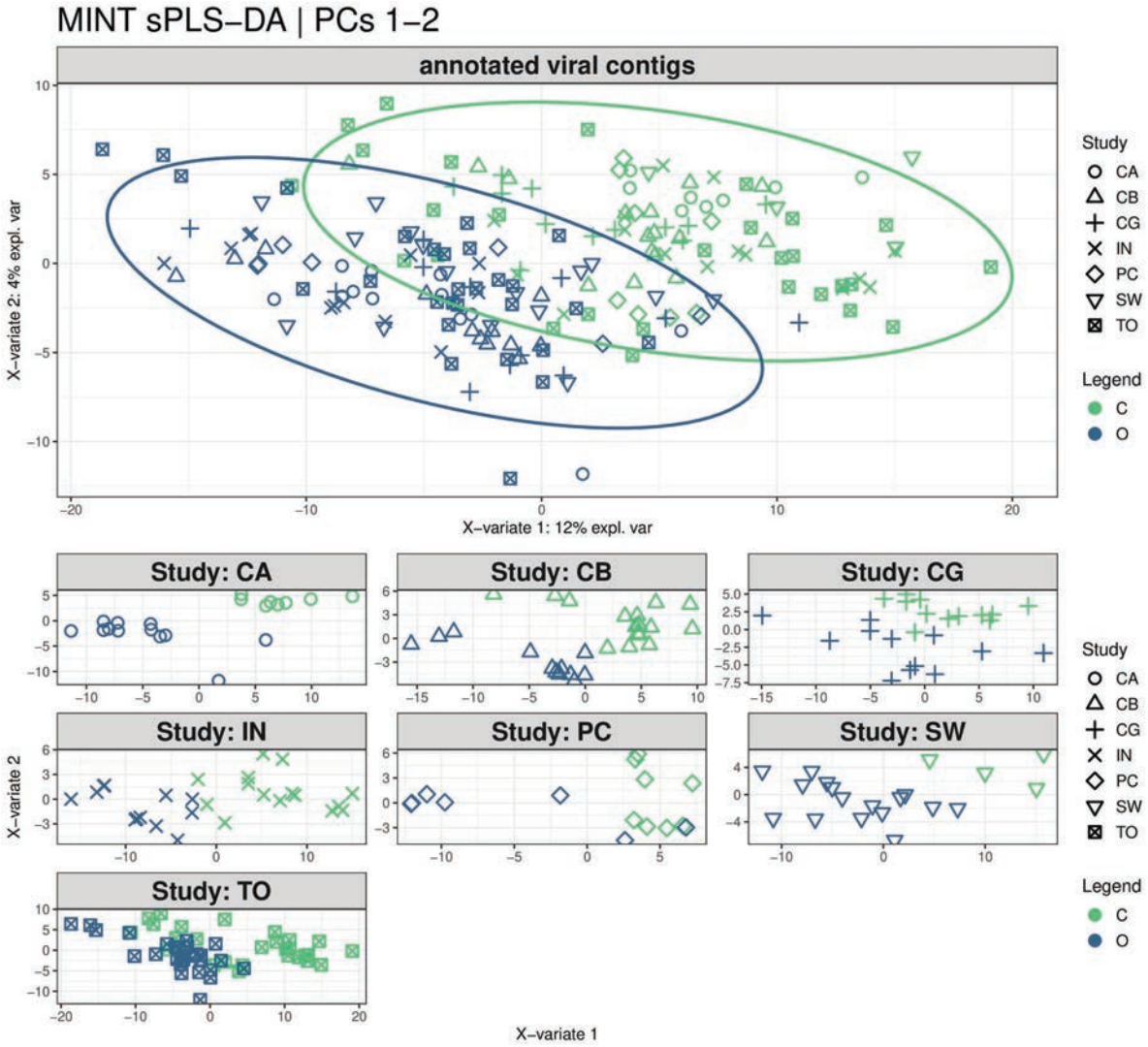

Figure S18. Sample plots from the MINT sPLS-DA performed on the 6,024 IMOS GBR-MGD seawater viral contigs, aiming to find discriminatory viruses between reefs that are open or closed to fishing. Samples (48 reef sites x 4 replicates) are projected into the space spanned by the first two components. Reef sites are coloured by their protection level (open or closed to fishing) and symbols indicate the membership of reef sites to their corresponding LTMP trip/transect. (**top**) Global components from the model with 95% ellipse confidence intervals around each sample class. (**bottom**) Partial components per study show a good agreement across GBR sectors. Component 1 discriminates between reefs that are open or closed to fishing.

We can choose the keepX parameter using the tune() function for a MINT object. The function performs LOGOCV for different values of test.keepX provided on each component, and no repeat argument is needed. Based on the mean classification error rate (overall error rate or BER) and a centroids distance, we output the optimal number of variables keepX to be included in the final model. Here, we

tested between 10 and 500 viral contigs in intervals of 10 (test.keepX = seq(10, 500, 10)) for 5 MINT sPLS-DA dimensions (ncomp = 5). For the final MINT sPLS-DA model, we retained the first two components with 240 and 110 features per dimension, as suggested by the *perf()* (see **Fig. S17**) and *tune()* functions (**Fig. S18**), respectively.

Performance of the final MINT sPLS-DA model

Use of the auroc() function will yield a visualisation of classification performance when undergoing the LOGOCV procedure from above. The interpretation of this output may not be particularly insightful in relation to the performance evaluation of mixOmics methods, but can complement the statistical analysis. For example, the MINT sPLS-DA classification of fished vs. NTMR sites had ~73 % accuracy in classifying samples in their corresponding zoning category.

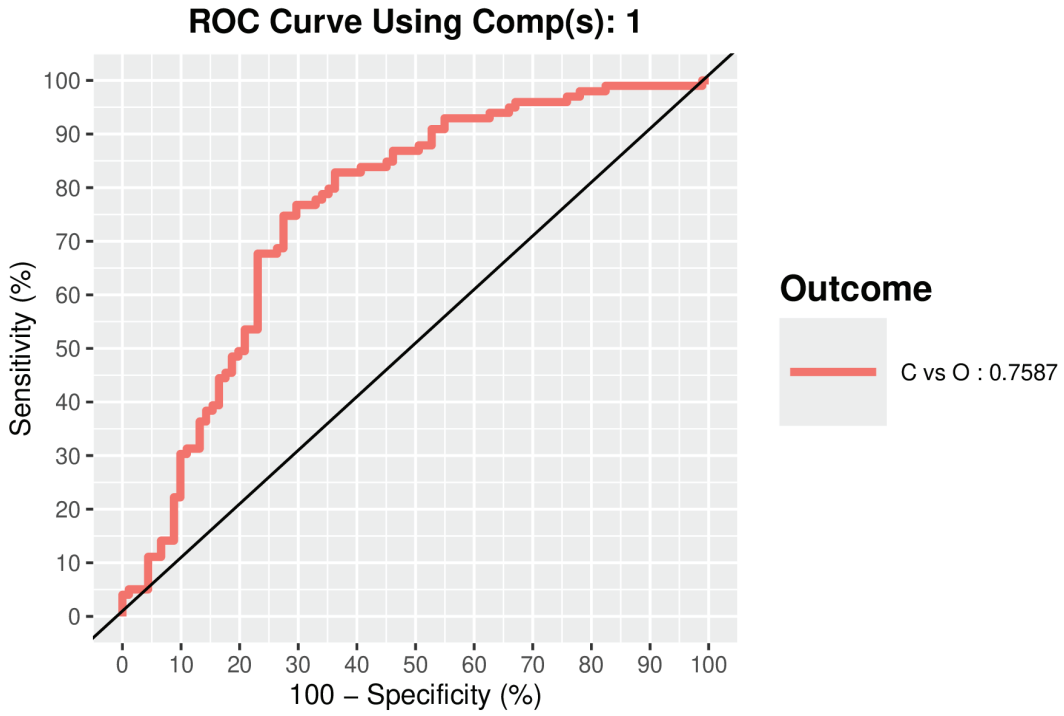

Figure S19. ROC curve and AUC from the MINT sPLS-DA performed on the IMOS GBR-MGD 6,024 viral contigs (>3kb; prevalence of >75% of samples within NTMRs and fished reefs; average rel. abund > 0.001%; with GENOMAD taxonomic annotations) for global component 1 for the fished vs. NTMRs reefs comparison. Numerical outputs include the AUC (0.7587) and a Wilcoxon test p–value ($p = 3.72 \times 10-8$) for fished vs. NTMRs reefs class comparison that are performed per component.

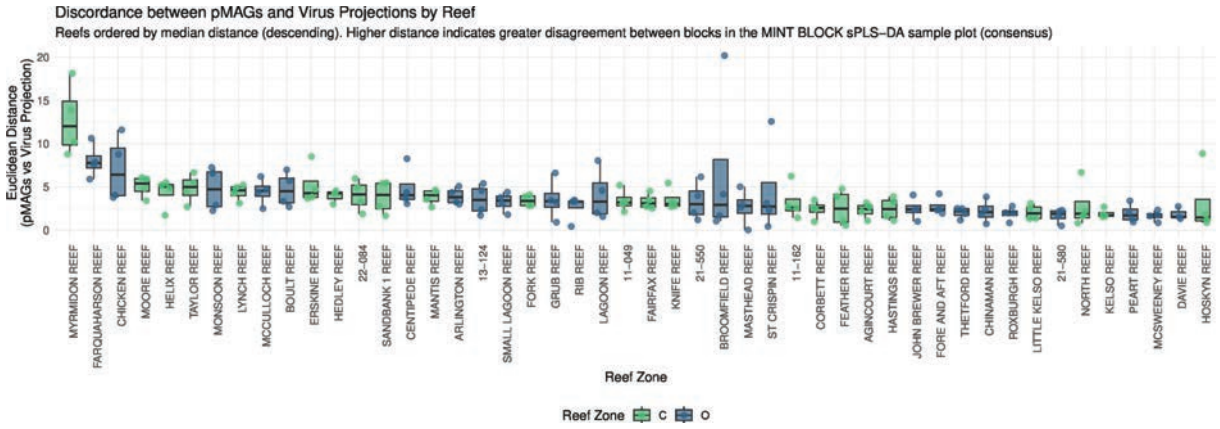

Figure S20. Quantifying discordance between pMAG and viral block projections across reef sites in the MINT BLOCK sPLS-DA consensus sample plot. Euclidean distances were calculated between each sample's projection in the pMAG-derived and virus-derived latent spaces from the MINT BLOCK sPLS-DA model. Reefs are ordered along the x-axis by their median distance (descending), highlighting sites with the greatest systematic disagreement between the two data blocks. Boxplots show the distribution of distances within each reef (colored by reef management zone), with points representing individual samples (four replicates). Higher distance values indicate greater discordance, potentially suggesting temporal or contextual decoupling between viral abundance and their microbial hosts at those specific sites.

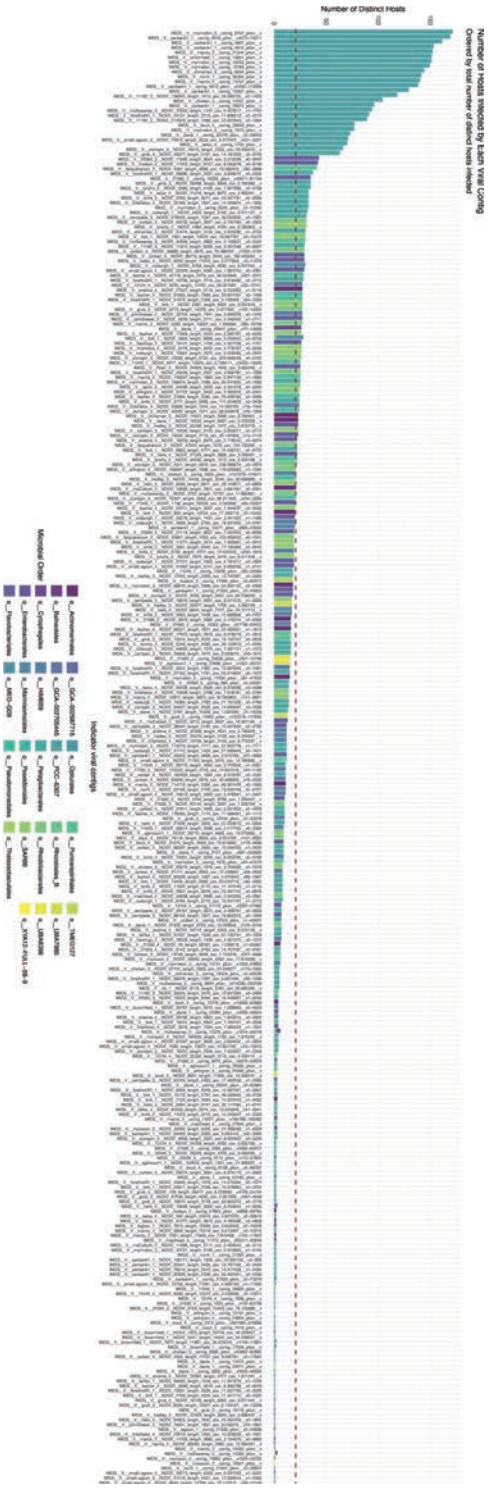

Figure S21. Host range (1-170) is shown for the indicator viral contigs. Average number of hosts per viral contig: 20.59; Standard deviation: 32.39. Microbial hosts were colored based on taxonomic order.

11 Appendix D – Comparison of NCBI vs GTDB Microbial Taxonomic Naming Conventions

This appendix clarifies differences in taxonomic nomenclature between the two primary classification systems used in this thesis. Specifically, microbial names used in Chapter 2 follow the NCBI (National Center for Biotechnology Information) taxonomy, whereas Chapters 3 and 4 use the Genome Taxonomy Database (GTDB). GTDB is a genome-based system that frequently revises prokaryotic taxonomy based on evolutionary relationships, leading to reassignments of families and orders compared to the NCBI system, which is more conservative and often retains historical names. Some names are stable across classification systems (e.g., *Flavobacteriaceae*, *Rhodobacteraceae*), though GTDB may reclassify them under revised higher ranks. Key differences include phylum names (Bacteroidota vs Bacteroidetes; Pseudomonadota vs Proteobacteria), formal order names for previously unclassified groups (e.g., Candidatus Poseidoniales, formerly known as Marine Group II archaea), and alphanumeric GTDB placeholder names for incompletely classified groups which would all fall under the "unclassified" category in NCBI. These differences reflect ongoing efforts to standardise microbial taxonomy and do not imply environmental absence of the organisms, and their understanding is essential for interpreting microbial community shifts across studies.

These differences are explained in more detail in the table below:

NCBI name (Chapter 2)	GTDB name (Chapters 3 and 4)	Notes
Bacteroidetes (phylum)	Bacteroidota (phylum)	GTDB uses standardised -ota suffix for phylum names.
Proteobacteria (phylum)	Pseudomonadota (phylum)	GTDB renamed Proteobacteria to Pseudomonadota based on phylogenomic consistency.
Deltaproteobacteria (class)Classes: Desulfobacteria, Desulfobulbia, Desulfovibrionia, etc. (split into multiple classes) GTDB splits Deltaproteobacteria into several distinct classes within Pseudomonadota	Classes: Desulfobacteria, Desulfobulbia, Desulfovibrionia, etc. (split into multiple classes)	GTDB splits Deltaproteobacteria into several distinct classes within phylum Pseudomonadota
Not used in Chapter 2	Candidatus Poseidoniales (order): comprising the families Candidatus Poseidonaceae fam. nov. (formerly subgroup MGIIa) and Candidatus Thalassarchaeaceae fam. nov. (formerly subgroup MGIIb)	GTDB provides formal order-level classification for MGII archaea in Rinke et al. (2018).
Not used in Chapter 2	Marinisomatales (order): former NCBI placeholder name SAR406 clade, or as part of the broader, unofficial "candidate phylum Marinimicrobia"	GTDB-specific order for certain marine bacteria.

Not used in Chapter 2	Puniccispirillales (order)	GTDB-specific order, not in NCBI classification.
Unclassified Bacteria (various)	*Classes*: UBA796, UBA1144, XYA12-FULL-58-9 *Orders*: UBA1151, UBA817, UBA796 *Families*: UBA1611, TMED-70, TCS55, TK06, D37C17, JABHHG01, TMED144, TMED161	GTDB assigns placeholder class/order/family names for deeply branching or unclassified bacterial lineages.
Unclassified Alphaproteobacteria (class)	*Orders*: TMED127, UBA7985, TMED109_A, MED-G09, HIMB59 *Families*: UBA1172, AAA536-G10, GCA-002684695 *Genera*: GCA-002690875, TMED13, CACBWF01, CACLCV01, GCA-2704625, GCA-2732015, GCA-002701455, MED-G40, MED-G52, AG-430-B22	GTDB assigns placeholder alphanumeric names at order, family, and genus levels for clades that are unclassified in NCBI.
Unclassified Gammaproteobacteria (class)	*Orders*: UBA11654, CACJXQ01, GCA-002705445 *Families*: TMED112, D2472, HTCC2089, GCA-002716945 *Genera*: SAR86A, D2472, UBA4421, TMED112, GCA-2707915, CACLCV01	GTDB assigns placeholder alphanumeric names at order, family, and genus levels for clades that are unclassified in NCBI.
Unclassified Flavobacteriia (class)	*Orders*: BACL11, UA16 *Families*: TMED113, TMED96, UBA11891, UBA8444, MED-G16 *Genera*: UBA11663, UBA8752, UBA10364, MED-G14, RFYP01, CAJXUB01, GCA-2697505, GCA-002715885, GCA-2862585, GCA-2718035, GCA-2693335, GCA-2863085, TMED96, MED-G20, MED-G13	GTDB assigns placeholder alphanumeric names at order, family, and genus levels for clades that are unclassified in NCBI.
Cyanophyceae (class)	Cyanobacteriia (class)	Incosistent names between NCBI and GTDB.
Synechococcus (genus)	*Synechococcus_C*, *Parasynechococcus*	Split into multiple genera in GTDB.
Pelagibacter or SAR11 (genus)	*Pelagibacter_A*, *Pelagibacter_B*, etc.	GTDB adds letter suffixes (A, B) to denote monophyletic genera within the *Pelagibacter* clades: Ia, Ic, II, IIIa.1, IIIa.2, IIIa.3, IIIb (LD12)