Article

The planktonic microbiome of the Great Barrier Reef

https://doi.org/10.1038/s41586-026-10778-z

Received: 18 May 2025

Accepted: 8 June 2026

Published online: 22 July 2026

Open access

Check for updates

Steven Robbins[1], Marko Terzin[2,3,4,5], Katherine Dougan[1], Julian Zaugg[1], Sara C. Bell[2], Patrick W. Laffy[2,3], J. Pamela Engelberts[1], Kim-Anh Lê Cao[5], Renee K. Gruber[2], Nicole S. Webster[1,6], David G. Bourne[2,3,4]✉, Philip Hugenholtz[1]✉ & Yun Kit Yeoh[2,3]✉

Large genome databases have markedly improved our understanding of marine microorganisms[1–5]. Although these resources have focused on prokaryotes, genomes from many dominant marine lineages, such as *Pelagibacter* and *Prochlorococcus*, are conspicuously underrepresented. Here we present the Great Barrier Reef Microbial Genomes Database (GBR-MGD), comprising 5,283 prokaryotic genomes obtained from Great Barrier Reef seawater samples using Nanopore and Illumina sequencing, including a collection of high-quality genomes of underrepresented groups. We show that standard short-read assemblies miss these populations owing to a combination of strain heterogeneity and low-GC-percentage sequencing bias. The GBR-MGD also comprises 20 chromosome-level picoeukaryote and 808,585 viral genomes, including a newly described clade of marine *Crassvirales*. We demonstrate the utility of the GBR-MGD to identify indicator taxa that can reliably predict the effects of reef management practices, such as the establishment of marine protected zones.

Microorganisms dominate life in the oceans[6], forming the foundation of marine food webs[7] and driving global biogeochemical cycles[8] through the metabolic interactions of bacterial, archaeal, viral and microbial eukaryotic communities. Understanding how these communities function, interact and change over time is critical for predicting ecological resilience, especially as oceans come under increasing pressure from climate change[9]. The overwhelming majority of microbial species found in marine ecosystems have yet to be cultured; however, the establishment of large-scale databases of prokaryote genomes via metagenomic sequencing has markedly improved our understanding of the identity, functional potential and distribution of marine microorganisms[1–5]. Among the prokaryotic inventories, however, many of the dominant species are conspicuously underrepresented, as evidenced by flow cytometry cell counts and metagenomic read mapping, including *Pelagibacter*, *Prochlorococcus*, SAR86 and others[5,10]. Further, the viral and eukaryote components of metagenomic surveys are less often included, with a few notable exceptions[11–17]. The oceans around Australia are also mostly absent in global marine genome surveys[4,18], as are its coral reefs[19,20], which are hotspots for biodiversity[9]. To address these gaps, we provide a detailed assessment of the pelagic microbiome of the Great Barrier Reef (GBR), the world's largest coral reef, using hybrid long-read and short-read DNA sequencing to construct a holistic database of bacteria, archaea, eukaryotes, viruses and plasmids, collectively termed the Great Barrier Reef Microbial Genomes Database (GBR-MGD).

We show that several prokaryote taxa are recalcitrant to recovery with Illumina short reads alone, owing to a combination of high strain heterogeneity and a bias against low-GC taxa, which can be rectified by inclusion of Nanopore long-read sequencing. This technology also enabled us to assemble *Crassvirales* genomes that form a distinct group within the order, as well as chromosome-level picoeukaryotic metagenome-assembled genomes (eMAGs) from multiple *Bathycoccus* structural variants and the as yet undescribed *Ostreococcus* clade B, often the most common picoeukaryote found on coral reefs. We were able to leverage the GBR-MGD prokaryote metagenome-assembled genomes (pMAGs) and machine learning techniques to identify reproducible shifts in community structure that reflect the GBR Marine Park Authority's management plan designating reefs as open or closed to fishing. To our knowledge, this is the first demonstration of a large-scale ecosystem-wide effect of fisheries management on microbial communities of the surrounding seawater. This study marks a substantial advance in our ability to recover complete and near-complete prokaryotic, eukaryotic and viral genomes from seawater samples, enabling marine ecosystem research and management.

Benchmarking hybrid metagenome assemblies

The absence of representative microbial genomic surveys of the GBR hampers our ability to understand how reef microbial communities are changing, especially in response to anthropogenic stressors such as climate change. To generate a comprehensive database of GBR marine microorganisms, seawater samples were collected from 48 sites spanning a north-to-south transect of the GBR (Fig. 1) for metagenomic sequencing (Supplementary Tables 1–3). To enhance metagenome-assembled genome (MAG) recovery, we used a hybrid assembly of long and short reads, which has been shown to substantially improve the quality of MAGs compared to short-read assemblies[21].

We hypothesized that common but difficult-to-recover marine taxa such as *Pelagibacter* could be obtained using Nanopore long-read

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Queensland, Australia. [2]Australian Institute of Marine Science, Townsville, Queensland, Australia. [3]AIMS@JCU, Research Division, James Cook University, Townsville, Queensland, Australia. [4]College of Science and Engineering, James Cook University, Townsville, Queensland, Australia. [5]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, Australia. [6]University of Tasmania, Battery Point, Tasmania, Australia. ✉e-mail: david.bourne@jcu.edu.au; p.hugenholtz@uq.edu.au; y.yeoh@aims.gov.au

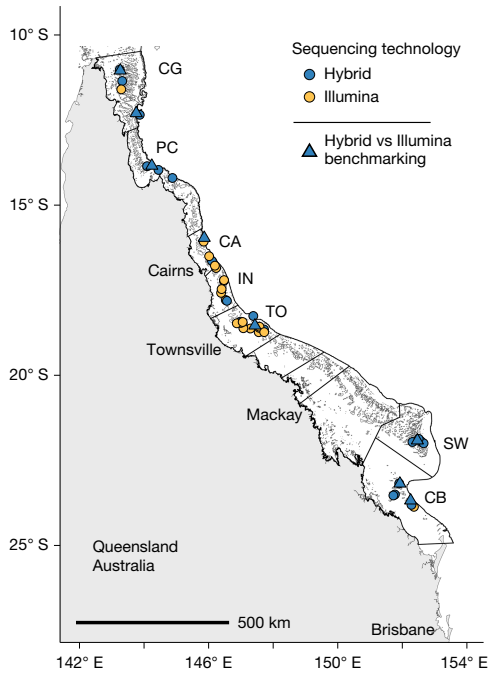


Fig. 1 | Location of the 48 sites sampled across the GBR. Sampling was conducted alongside reef surveys performed by the AIMS LTMP. At each site, 4 × 5 l replicate seawater samples were collected for DNA extraction. Sites with Nanopore and Illumina sequence data (hybrid) are indicated with blue circles or triangles while sites with Illumina data only are indicated in yellow. Blue triangles represent the eight sites used for benchmarking Nanopore and Illumina hybrid versus Illumina-only assemblies. Sector names are indicated next to their locations. CA, Cairns; CB, Capricorn Bunker; CG, Cape Grenville; IN, Innisfail; PC, Princess Charlotte Bay; SW, Swain; TO, Townsville.

sequencing to span strain-variable regions based on the expectation of high strain heterogeneity[22]. To test this hypothesis, we selected eight seawater samples for Illumina-only and hybrid (Illumina plus Nanopore) assemblies and compared the recovery of medium- to high-quality pMAGs with a focus on lineages previously found to be difficult to recover as MAGs. The total number of base pairs in both Illumina-only and hybrid assemblies was standardized to 30 Gbp to remove sequencing depth as a factor. Contiguity of the hybrid-assembled pMAGs was 29-fold higher than that of the Illumina-only pMAGs (N50 of 730 versus 25 kb), exceeding the genome length of some marine taxa (for example, Nanoarchaeota[23]; Fig. 2a, Supplementary Tables 4 and 5 and Extended Data Fig. 1). Commensurately, the average number of medium to high-quality pMAGs per sample increased about twofold in hybrid assemblies (64 to 123) with the increase even more apparent for high-quality pMAGs (13 to 40; Fig. 2a and Extended Data Fig. 1).

The increased number of hybrid pMAGs included lineages that were previously identified as difficult to recover with short-read-only sequencing, including Pelagibacterales, *Parasynechococcus*, *Prochlorococcus* and SAR86, along with other lineages that were not previously recognized as difficult to recover. For example, pMAGs from the Alphaproteobacteria class HIMB59, TMED109, TMED127 and MED-G09, the family AAA536-G10 within the order Puniceispirillales (formerly SAR116), and the Gammaproteobacteria order GCA-002705445, were not recovered from Illumina-only assemblies but were consistently recovered in hybrid assemblies (Fig. 2a). Notably, genomes belonging to the order Pelagibacterales and phylum Marinisomatota (formerly Marinimicrobia and SAR406) were recovered infrequently in Illumina-only assemblies, whereas 5 to 17 pMAGs were recovered per sample from the hybrid assemblies (Fig. 2a), resulting in substantial missed diversity using short read only assemblies—25 versus 2 different species across the 8 samples for Pelagibacterales and 13 versus 7 species in Marinisomatota (Supplementary Tables 4 and 5). Within the Cyanobacteriota, *Prochlorococcus* pMAGs were generally not recovered from Illumina-only assemblies but were recovered from all hybrid assemblies. Notably, whereas *Parasynechococcus* (formerly *Synechococcus*_E) pMAGs could be recovered from both short-read and hybrid assemblies, only the hybrid assemblies recovered representatives of *Parasynechococcus* species sp002724845 (Fig. 2a), the most dominant prokaryote lineage in GBR seawater, averaging 8.9 ± 5.9% relative abundance (maximum 35.9%) across the complete GBR-MGD dataset. By comparison, the next most abundant *Parasynechococcus* species averaged only 0.2% (Supplementary Table 6).

We next investigated whether strain heterogeneity was the driving factor behind the improved recovery of pMAGs using Nanopore hybrid assemblies. One way to assess strain heterogeneity is to quantify single nucleotide polymorphism frequencies, and indeed, a high proportion of hybrid pMAGs from difficult-to-recover taxa represent populations with high nucleotide diversity (>0.02; Fig. 2b), a measure of genetic diversity within microbial populations implemented in inStrain[24]. By contrast, Illumina-only pMAGs rarely have nucleotide diversity values greater than 0.02, indicating a threshold above which short-read pMAG recovery is severely affected (Fig. 2b). However, several difficult-to-recover taxa have sufficiently low strain heterogeneity (0.005 to 0.02 inStrain nucleotide diversity; Fig. 2b) to be recovered by Illumina-only assembly, suggesting a secondary reason for improved pMAG recovery in hybrid assemblies. By mapping both the Illumina and Nanopore reads used in hybrid assemblies back to their respective pMAGs, we found that difficult-to-recover taxa were consistently under-represented in Illumina datasets (up to 18× less than in Nanopore datasets; Fig. 2c). Plotting GC content and fold Nanopore overrepresentation for all pMAGs, defined as the relative abundance in Nanopore reads divided by the relative abundance in Illumina reads, we found that low-GC pMAGs were better represented in Nanopore than Illumina data (Fig. 2c). For example, Marinisomatota pMAGs are up to 18× less abundant (average 6×) in Illumina reads than Nanopore, *Prochlorococcus* are up to 16× less abundant (average 7×), Gammaproteobacteria GCA-002705445 are up to 15× (average 6×) less abundant, and *Pelagibacter* are up to 7× (average 3×) less abundant. In addition to high nucleotide diversity, this would account for the limitations of Illumina-only assemblies in recovering pMAGs from difficult taxa, as all have GC contents below 40%. This is consistent with a known GC bias in polymerase-based sequencing platforms, including Illumina[25]. When considering both GC bias and strain diversity as factors, short-read MAG recovery appears to be difficult above 0.005 inStrain nucleotide diversity and less than 40% GC, as few pMAGs were recovered within this range and those that were showed lower completeness values (Fig. 2b). The problem reaches a critical threshold above 0.02 nucleotide diversity and <40% GC, where no pMAGs could be recovered from the short-read assemblies, despite many difficult taxa in this range being readily recovered in hybrid assemblies (56 hybrid versus 0 Illumina pMAGs) (Fig. 2b). In summary, we show that assemblies that incorporate long reads substantially enhance the recovery of dominant and ubiquitous marine lineages that are often missed by short-read-only assemblies, not only because long reads can overcome elevated strain heterogeneity but because of GC bias inherent to Illumina sequencing.

Recovery of prokaryote MAGs

Having established that Nanopore sequencing substantially improves representation of cosmopolitan marine lineages, we extended our efforts to additional samples to produce the GBR-MGD. Hybrid assemblies from 27 of the 48 GBR seawater samples yielded 4,713 bacterial and archaeal pMAGs, including 1,505 high-quality (>90% completeness with <5% contamination) and 345 circular pMAGs assumed to be complete. We supplemented these data with Illumina-only assemblies from the remaining 21 sites resulting in an additional 570 pMAGs

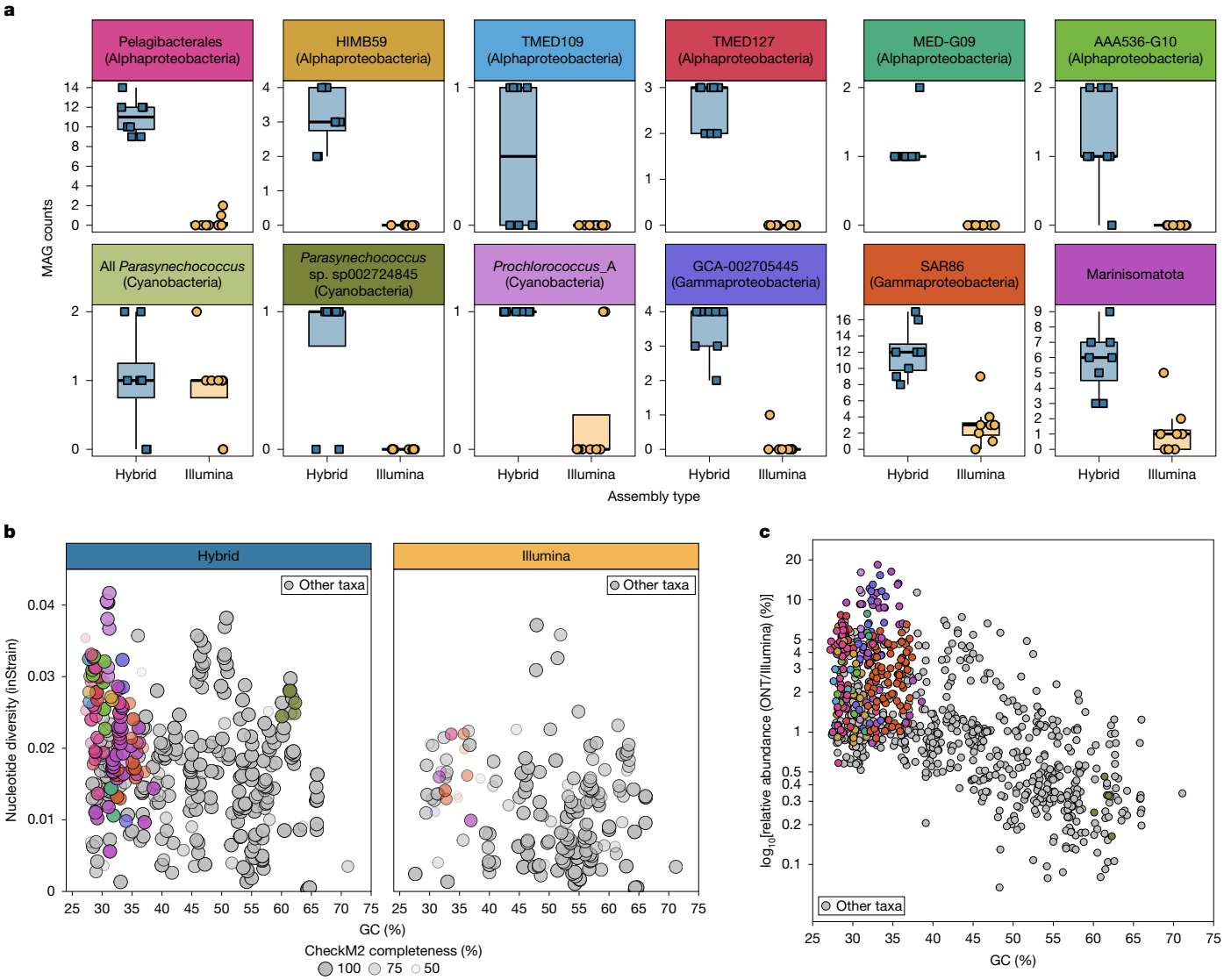


Fig. 2 | Plots illustrating recovery and sequencing bias between Illumina-only and Nanopore hybrid pMAGs. a, Box-and-whisker plots showing number of recovered pMAGs per sample for taxa that are underrepresented in Illumina-only assemblies ($n = 8$ samples from different sites used for comparative Illumina-only versus hybrid assembly benchmarking). Centre lines in the boxes represent median values, top and bottom box boundaries represent top and bottom quartiles, respectively, whiskers extend to 1.5× interquartile range, and minima and maxima are presented as individual data points. **b**, Plots of nucleotide diversity (a measure of genetic diversity within microbial populations implemented in inStrain) and average GC percentage for hybrid (left; $n = 382$) and Illumina-only (right; $n = 179$) benchmark pMAGs after species-level dereplication (95% ANI), highlighting taxa that are challenging to recover using short reads in **a**. Circle opacity and size are scaled according to CheckM2 completeness. **c**, Plot showing Nanopore overrepresentation as a function of average MAG GC percentage across all benchmark hybrid ($n = 987$) and Illumina-only ($n = 516$) pMAGs, where overrepresentation is defined as the ratio of pMAG relative abundance in Nanopore versus Illumina reads. All MAGs have CheckM2 quality scores >50 and contamination <10%. Circle colours in panels **b** and **c** correspond to taxa colours in **a**.

(85 in high quality), bringing the total number of bacterial and archaeal pMAGs to 5,283 in the final GBR-MGD database. Of the taxa identified to be recalcitrant to pMAG recovery, we obtained 319 Pelagibacterales pMAGs (22 high quality, 7 circular), 445 SAR86 (128 high quality, 35 circular), 222 Marinisomatota (70 high quality, 14 circular), 21 *Parasynechococcus* sp002724845 (1 high quality, 0 circular), 26 *Prochlorococcus* (8 high quality, 0 circular), 135 Gammaproteobacteria order GCA-002705445 (67 high quality, 34 circular) and 45 Alphaproteobacteria family AAA536-G10 (4 high quality, 0 circular). The recovery of circular pMAGs also enabled us to identify lineages for which CheckM version 1 (CheckM1) systematically underestimates completeness, as circular pMAGs should technically have a completeness value of 100% (the exception being multipartite genomes containing two or more replicons). Indeed, all circular genomes from the following taxa had completeness estimates below 85%: Gammaproteobacteria orders SAR86 and GCA-002705445, Bacteroidota genus MED-G13, Chloroflexota order UBA1151, Desulfobacterota class UBA1144, and the archaeal order Poseidonales (Supplementary Table 7). Completeness estimates were substantially higher for circular genomes using CheckM2, averaging 91% versus 79% for CheckM1, though groups GCA-002705445 and SAR86 are still underestimated using CheckM2 (average approximately 75% completeness) despite being circular. This underscores the need to use CheckM2 for completeness estimates of marine pMAGs to avoid discarding systematically underestimated genomes, which is likely to have contributed to underrepresentation of certain taxa.

Species-level dereplication (≥95% average nucleotide identity (ANI)) of the final set of GBR-MGD pMAGs resulted in a set of 876 unique species (Supplementary Table 6). To determine the novelty of the GBR-MGD

data, we compared all pMAGs to the approximately 32,000 pMAGs (≥50 quality based on CheckM1 or CheckM2) present in the Ocean Microbiomics Database[4] (OMD), revealing that 584 of the GBR-MGD species (66.7%) were not previously represented (<95% ANI to OMD sequences) (Supplementary Table 8). These may therefore represent either newly recognized marine species potentially specific to the GBR or taxa that are present elsewhere but are not represented in public databases owing to difficulties with assembly. Notably, of the 328 species clusters with members from both the GBR-MGD and OMD, GBR-MGD pMAGs were the best representative 65% of the time (213 clusters) based on CoverM Cluster scoring criteria, which take into account genome completeness, contamination and number of contigs in each pMAG. Thus, the GBR-MGD comprises a wealth of new species diversity captured in high-quality pMAGs. Mapping of the GBR-MGD Illumina reads from all samples to the dereplicated pMAGs (Supplementary Table 6) showed that an average of 47 ± 11% (maximum 66%) of the community was represented by prokaryotic taxa. This result exceeds previous read-mapping statistics of marine metagenomes against large prokaryote genome databases, including *Tara* Oceans (2,631 MAGs; 7% average mapping)[1], GEM[26] (8,578 MAGs; approximately 20% average mapping), single amplified genome (SAG)-only Global Ocean Reference Genomes Tropics Database[27] (20,288 SAGs; approximately 20–40% average mapping), and is nearly on par with the OMD[4], a gold standard aggregated set of >35,000 MAGs and SAGs from all previous marine studies (approximately 58% average mapping for the 0.2–3 µm size fraction). Mapping of the GBR-MGD Illumina reads to OMD MAGs and SAGs resulted in an average 40.4 ± 7.7% (maximum 53.7%) of the GBR community represented by OMD genomes, indicating that the 876 GBR-MGD pMAGs capture regional microbial diversity not yet present in the OMD. Additionally, average read mapping rates of seawater metagenomes from seven Australian National Reference Stations to the GBR-MGD pMAGs ranged from 1.3% to 32.8% (Extended Data Fig. 2), indicating that seawater microorganisms on the GBR are distinct compared with these marine monitoring sites across Australia.

Microbial community composition of the GBR differed across the seven sectors surveyed (adjusted $P < 0.05$, pairwise permutational multivariate analysis of variance (PERMANOVA); Fig. 3a), although a single *Parasynechococcus* species (sp002724845) was dominant in six of the sectors (8.5% to 14.7% average relative abundance) (Supplementary Table 6) and adjacent sectors tended to be compositionally similar (Fig. 3a). Communities in the seventh sector (TO) were distinct and significantly (adjusted $P < 0.1$, MaAsLin 3) enriched in *Prochlorococcus*_A sp003278705 (4.5% average relative abundance versus 0.5% in other sectors), *Pelagibacter* sp902612605 (0.6% versus 0.2%) and *Actinomarina* sp003213625 (0.5% versus 0.3%) (Supplementary Table 9). To determine potential drivers of the observed compositional variation, we compared two of the most divergent sectors, TO and IN. Water chemistry data fitted to community composition indicated that TO communities (Fig. 3a) were associated with fivefold higher phosphate and sixfold lower ammonium concentrations than IN (Supplementary Table 10). This was reflected in their metagenomes, with IN having higher abundances of Pst phosphate transport system and carbon-phosphorus lyase complex genes than TO (adjusted $P < 0.1$, DESeq2; Supplementary Table 11). These proteins facilitate the import of inorganic phosphate[28] and metabolism of phosphonate compounds as a source of phosphate[29], respectively, under phosphate-limiting conditions. Conversely, nitrate reductases were less abundant in IN compared with TO (Supplementary Table 11), possibly linked to inhibition of these enzymes by ammonium[30].

Microbial indicators of reef zoning

Marine protected areas are refuge zones that were created to protect sensitive reef-dwelling animals from human disturbance. When strictly enforced, such areas are known to be effective tools for reef conservation that can buffer the negative effects of fishing and potentially marine heatwaves on sensitive species[31]. Consequently, the GBR Marine Park Authority has categorized reefs into zones[32] that are open to fishing (Fished), and No-Take Marine Reserves (NTMRs), which are closed to fishing to protect ecosystems and maintain stable populations of commercially harvested fish (Supplementary Table 10). Marine microbial communities have been shown to shift rapidly and predictably in response to environmental change such as nutrient loads and temperature, which could be leveraged to inform management decisions[33]. To test whether the GBR-MGD can predict reef protection status, we analysed the abundances of prokaryotic, viral and eukaryotic MAGs and contigs (Methods, 'Final GBR-MGD database compilation') from co-located Fished and NTMR reefs (Extended Data Fig. 3) using multivariate integration sparse partial least squares discriminant analysis (MINT sPLS-DA) to perform feature selection controlling for sector, followed by classification with random forest. This approach predicted protection status with 74.6 ± 1.4% accuracy (mean ± s.d.), suggesting an ecosystem-wide effect that extends into the surrounding seawater to influence microorganisms. As prokaryotic communities are more easily profiled (for example, for development of environmental assessment tools[34]) and have been better characterized compared with viruses and eukaryotes, we repeated this analysis with pMAGs alone, resulting in 71.4 ± 0.9% accuracy (mean ± s.d.). Notably, the majority of pMAGs indicative of NTMR status belong to Pelagibacterales (7.9% versus 5.4% average relative abundance in NTMR versus Fished zones), SAR86 (1.6% versus 1.1%), HIMB59 (0.6% versus 0.5%), TMED109 (0.1% versus 0.06%) and Marinisomatota (0.6% versus 0.4%) (Fig. 3b), all identified in the present study as lineages displaying high strain heterogeneity and low GC content that are underrepresented in Illumina short-read sequence data (Fig. 2). Many of these taxa have undergone genome streamlining (reduced genome size and lower GC content; Fig. 3c) as an adaptation to oligotrophic habitats to reduce their requirement for nutrients[35–37] (Fig. 3d). By contrast, the majority of taxa indicative of Fished reefs have GC contents above 45% with no evidence of genome streamlining, including members of the genera UBA11663 and UBA8752, species UBA10364 sp003445735 in the class Bacteroidia, the family Halieaceae in the Gammaproteobacteria, and the species HIMB11 sp000472185 in the Alphaproteobacteria. In fact, at 60%, 63% and 45% average GC respectively, the UBA11663, UBA8752 and UBA10364 pMAGs have GC contents 9 to 27% higher than the average member of the Bacteroidia (36%) from the GBR-MGD (Supplementary Table 7). The association of NTMR reefs with streamlined, low-GC taxa that require less nitrogen for DNA synthesis[36–39] than their higher-GC counterparts on Fished reefs suggests that protection status may influence reef nutrient levels. Leveraging nutrient measurement data taken alongside these microbial samples, we find that dissolved nitrogen (ammonia, nitrate and nitrite) is indeed lower in NTMR reefs compared with Fished reefs (Extended Data Fig. 4), suggesting that differences in nutrient levels between NTMR and Fished reefs select for taxa with different ecological and evolutionary strategies. Decreased fish densities and thus grazing are postulated to result in increased macroalgae, which in turn release nutrients that may account for higher nutrient levels in Fished reefs[33,40–42].

Although NTMRs were generally associated with prokaryotic oligotrophs and Fished reefs with copiotrophs (Fig. 3e and Supplementary Table 12; false discovery rate-corrected $P < 0.05$, MaAsLin 3), the enrichment of specific taxa was region-dependent. For example, several oligotrophic *Pelagibacter* species were only enriched in IN NTMRs whereas *Prochlorococcus*_A was the most abundant genus in TO NTMRs. Conversely, inferred copiotrophic species such as *Luminiphilus* sp011524755 (order Pseudomonadales) were enriched in IN Fished reefs and MED-G52 sp001627375 (order Rhodobacterales) enriched in TO Fished reefs[43]. This regional specificity was also reflected by limited overlap of the most abundant species (3 out of 20) that distinguish reef protection zoning in each sector (adjusted $P < 0.1$, MaAsLin 3) (Supplementary Table 13). Regional specificity

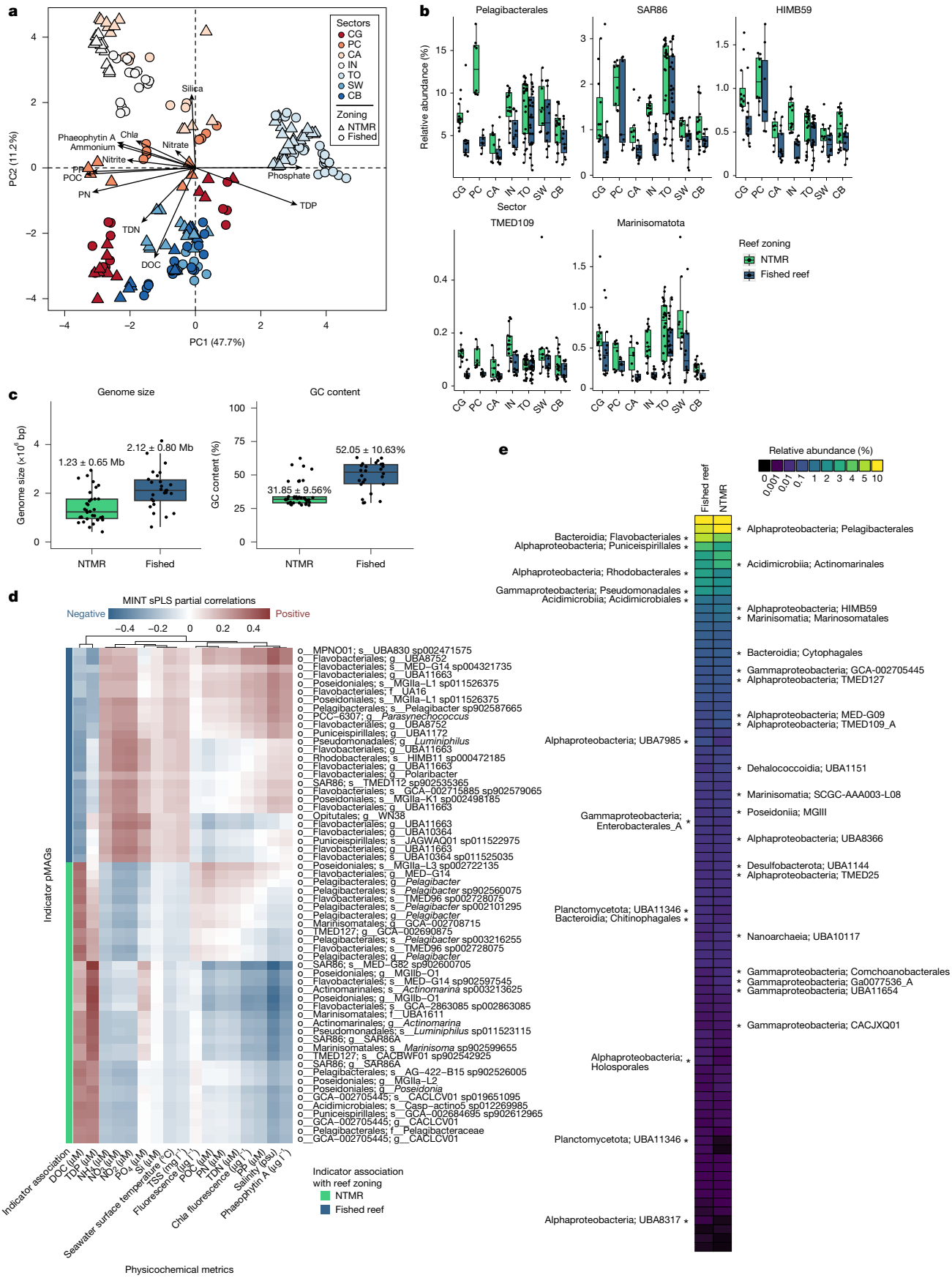


Fig. 3 | See next page for caption.

Fig. 3 | Microbial community composition on the GBR and indicators of reef protection zoning. a, Principal component analysis visualization of microbial community composition (n = 48 reef sites × 4 replicates). Chl*a*, chlorophyll *a*; DOC, dissolved organic carbon; PN, particulate nitrogen; POC, particulate organic carbon; PP, particulate phosphorus; TDN, total dissolved nitrogen; TDP, total dissolved phosphorus. **b**, Relative abundances of indicator taxa enriched in NTMRs (n = 23) compared with Fished reefs (n = 25) across the seven surveyed GBR sectors. **c**, Average genome size and GC content of the top 60 most influential indicator pMAGs (95% ANI dereplicated) from NTMR (n = 34 indicator pMAGs) and Fished (n = 26 indicators) reefs. Values quoted directly above the box plots are mean ± s.d. **d**, Heat map of MINT sPLS partial correlation coefficient values for each indicator pMAG showing positive (red) or negative (blue) correlation with seawater physicochemical variables. Si, silica; TSS, total suspended solids. **e**, Heat map showing microbial community composition of NTMRs (n = 23) versus Fished reefs (n = 25) based on relative abundances at order ranks. Asterisks indicate statistical association (q < 0.05, modelled using generalized linear models implemented in MaAsLin 3; P values generated using two-tailed tests and corrected for multiple comparisons with Benjamini–Hochberg false discovery rate) with either NTMRs (right side of heat map) or Fished reefs (left side of heat map). In box plots (**b**,**c**), centre lines represent median values, top and bottom box boundaries represent top and bottom quartiles, respectively, whiskers extend to 1.5× interquartile range, and minima and maxima are represented by data points.

was also reflected to some extent by enrichment of different gene families in different sectors (Supplementary Tables 14 and 15). For example, adenosine triphosphate-binding cassette transport systems for inositol-phosphate (K17238, K17239 and K17240) were enriched in IN NTMRs, and transport systems for polar amino acids, spermidine and putrescine (K02029, K02052 and K02053) were enriched in TO NTMRs (adjusted P < 0.1, DESeq2), broadly indicative of low nutrient availability but probably indicating regional differences in available nutrients. Similarly, different genes suggestive of copiotrophic conditions were enriched in the different sectors. For example, in IN Fished reefs genes linked to the degradation of aromatic compounds such as phenylacetyl-CoA epoxidase (*paaABCDE*, K02609–K02613; phenylacetate degradation pathway) and *pcaDFG* (K00448, K01055 and K07823; β-ketoadipate pathway) were enriched, and in TO Fished reefs phenol 2-monooxygenase (K03380) and 4-methoxybenzoate monooxygenase (K22553) were enriched. In addition, TO NTMRs were enriched in several *Kyanoviridae* protein orthologues (Supplementary Table 15), probably linked to the high relative abundance of *Prochlorococcus*[44] (~4.5% relative abundance) in this sector. Together, these data indicate that whereas NTMR zoning results in broad enrichment of oligotrophic taxa, region-specific differences are more nuanced and highlight that markers tailored to regions could increase predictive power for reef management. More samples from NTMR and Fished reefs in each sector will be needed to test this idea.

Recovery of DNA viruses

Viruses are the most abundant biological entities in marine environments, and have a large impact on marine ecology through interactions with their prokaryote and eukaryote hosts. For example, it is estimated that viruses kill 20% of marine microbial biomass per day, representing a substantial carbon turnover[45]. Current state-of-the-art viral detection tools have almost exclusively been benchmarked on short-read assemblies, and are known to be sensitive to contig length[46,47]. To assess their use on long-read assemblies, we first tested six widely used metagenomic viral prediction tools to identify viruses in the GBR-MGD—GeNomad, Virsorter2, ViralVerify, VIBRANT, PPR-Meta and DeepVirFinder—using CheckV to assess the number of complete viruses predicted by each tool. We further plotted the ratio of prokaryotic host to viral marker genes identified by CheckV within each putative viral contig to detect spurious assignments. DeepVirFinder showed substantially more false positives than the other tools tested, with the proportion of bacterial marker genes among putative viral contigs increasing with contig length, indicating bacterial origin (Extended Data Figs. 5 and 6). When these bacterial contigs were assessed by CheckV, longer contigs were increasingly likely to be designated as complete viral genomes, despite having a high ratio of bacterial to viral markers (Extended Data Figs. 5 and 6). We note that CheckV is not designed to identify viral versus non-viral sequences, but highlight this result because it is common practice to filter assemblies based on CheckV quality scores that could result in extensive misidentification of viral auxiliary metabolic genes if DeepVirFinder is applied to long-read assemblies. We therefore excluded DeepVirFinder from our pipeline as unsuitable for use with long-read metagenomes, and the remaining five tools were used in concert to produce a consolidated dataset of 808,585 single contig viral MAGs (vMAGs) from across the 48 GBR-MGD reefs, representing 362,802 species-level (95% ANI) viral operational taxonomic units (vOTUs) (Supplementary Table 16). Read mapping against the vOTU representatives accounted for 4–27% of each metagenome (averaging 13 ± 4%), highlighting that viruses comprise a substantial proportion of GBR seawater metagenomes[45,48]. The overall vOTU community composition significantly differed across sectors (P < 0.01, PERMANOVA) and closely tracked the prokaryotic communities (P < 0.01, Procrustes rotation; Extended Data Fig. 7).

The majority of vOTUs were classified as *Caudoviricetes* (67% of sequences), of which only 6% were resolved below this class. This included the families *Kyanoviridae* (2.8%) encompassing the T4-like cyanophages that commonly infect *Prochlorococcus*[44] and *Synechococcus*[49]; *Autographiviridae* (1.4%), common phages of *Prochlorococcus*, *Synechococcus* and *Pelagibacter*[50–53]; and the order *Crassvirales* (1%; see below). A further 6% of the vOTUs were classified as giant viruses belonging to the *Nucleocytoviricota*, which probably reflects our ability to recover low-GC genomes characteristic of this phylum. The remaining vOTUs (27%) were unclassified, which may indicate bona fide marine viral novelty or difficulties in discriminating viruses from other genetic elements. Therefore, we provide the CheckV and GeNomad marker gene frequency tables to enable users to filter vMAGs based on their own criteria (Supplementary Tables 17–19).

Recovery of marine *Crassvirales*

Crassphages were first identified in cross assemblies of human gut metagenomes[54] and are currently recognized as the most abundant phage in the human gut[55,56]. Although members of the *Crassvirales* have since been identified in a range of additional environments such as primate and termite gut samples, terrestrial systems, groundwater and marine habitats[57–60], *Crassvirales* from these habitats remain poorly characterized. Within the GBR-MGD, 312 vMAGs were classified by GeNomad as *Crassvirales*, of which 167 contained a terminase large subunit marker gene (*terL*) containing at least 400 amino acid residues. However, phylogenetic placement in a TerL tree containing representative genomes from the class *Caudoviricetes* indicated that most of these sequences (155, 93%) are members of sister clades to the *Crassvirales*, the families *Pachyviridae* and *Pervagoviridae* (Fig. 4a), which primarily comprise free-living marine flavobacterial phages[61]. To confirm the placement of the GBR-MGD *Crassvirales* within the order, we constructed whole genome phylogenies including all putative *Crassvirales* GBR-MGD vMAGs more than 20 kb in length and select reference genomes from the *Caudoviricetes*, including the *Pachyviridae* and *Pervagoviridae* (Fig. 4b). This tree also contained recently discovered *Crassvirales* sequences from Antarctic seawater[62], providing an opportunity to assess the phylogenetic distribution of the *Crassvirales* across global oceans and between tropical coral reefs and Antarctic waters (Fig. 4b). Four distinct clades of Antarctic *Crassvirales*, C1–C4,

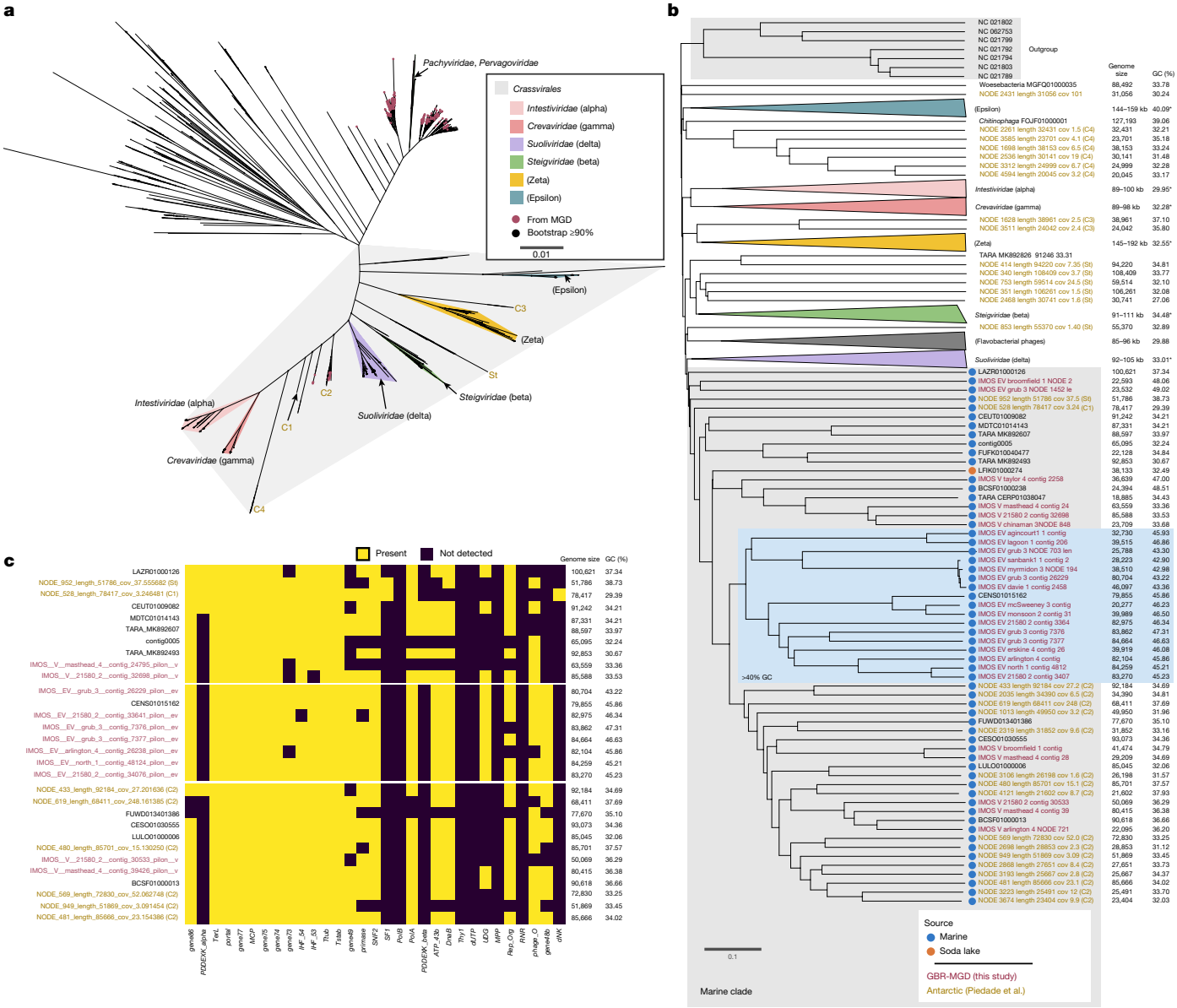


Fig. 4 | Expansion of the marine *Crassvirales*. **a**, TerL tree of *Crassvirales* and other *Caudoviricetes* showing placement of GBR-MGD vMAGs represented by red circles (MGD) in relation to existing *Crassvirales*. Families are labelled with International Committee on Taxonomy of Viruses (ICTV) names along with the previously proposed alpha, beta, gamma, delta, epsilon and zeta group names in parentheses[59]. C1–C4 and St groups are as proposed in Piedade et al.[62]. **b**, Whole-genome phylogeny (ViPTree) of GBR-MGD *Crassvirales* vMAGs larger than 20 kb ($n = 27$). **c**, Pattern of *Crassvirales* gene conservation in vMAGs and genomes within the marine clade. Only genomes larger than 50 kb are shown. Genes included here were previously identified by Yutin et al.[59]. ATP_43b, AAA domain ATPase; DnaB, phage replicative helicase DnaB family; dNK, deoxynucleotide monophosphate kinase; dUTP, dUTPase; UDG, uracil-DNA glycosylase; gene48b, phage endonuclease I; gene49, uncharacterized protein gene 49; gene73, putative structural protein gene 73; gene74, putative structural protein gene 74; gene75, putative structural protein gene 75; gene77, putative structural protein gene 77; gene86, putative structural protein gene 86; IHF_53, IHF subunit gene 53; IHF_54, IHF subunit gene 54; MCP, major capsid protein; MPP, metallophosphatase; PDDEXK, PD-(D/E)XK family nuclease; phage_O, bacteriophage replication protein O; PolA, DNA polymerase family A; PolB, DNA polymerase family B; portal, portal protein; primase, DnaG family primase; Rep_Org, replisome organizer protein; RNR, ribonucleotide reductase; SF1, SF1 helicase; SNF2, SNF2 helicase; TerL, terminase large subunit; Thy1, thymidylate synthase; Tstab, tail stabilization protein (P22 gp10-like); Ttub, tail tubular protein (P22 gp4-like). *Crassvirales* genomes recovered in this study are represented by red circles in **a** or red labels in **b**,**c**; genomes and groups from Piedade et al.[62] are represented in yellow (C1–C4 and St in **a**, genome labels in **b**,**c**). Colour labels of the *Crassvirales* clades are consistent across **a**,**c**.

have previously been proposed[62]. Notably, GBR-MGD *Crassvirales* are part of a larger monophyletic clade of exclusively marine sequences that includes the C1 and C2 clades (Fig. 4a). We confirm that all members of this marine lineage contain the full complement of core *Crassvirales* genes (Fig. 4c). Of note, most GBR-MGD *Crassvirales* form a separate clade within the marine lineage characterized by a higher GC content (average 45% versus 34%; Fig. 4b). However, several cluster with C2 genomes indicating that the C2 clade is not limited to the Antarctic but is also present in Australian tropical reefs, revealing a wider geographical distribution and temperature range (Fig. 4b). Similar to animal-associated *Crassvirales*, the marine lineage, both Antarctic and GBR, are predicted to infect Bacteroidota hosts[60], suggesting that this association is common to this viral order and that marine *Crassvirales* could be linked to marine carbon fluxes through interactions with their

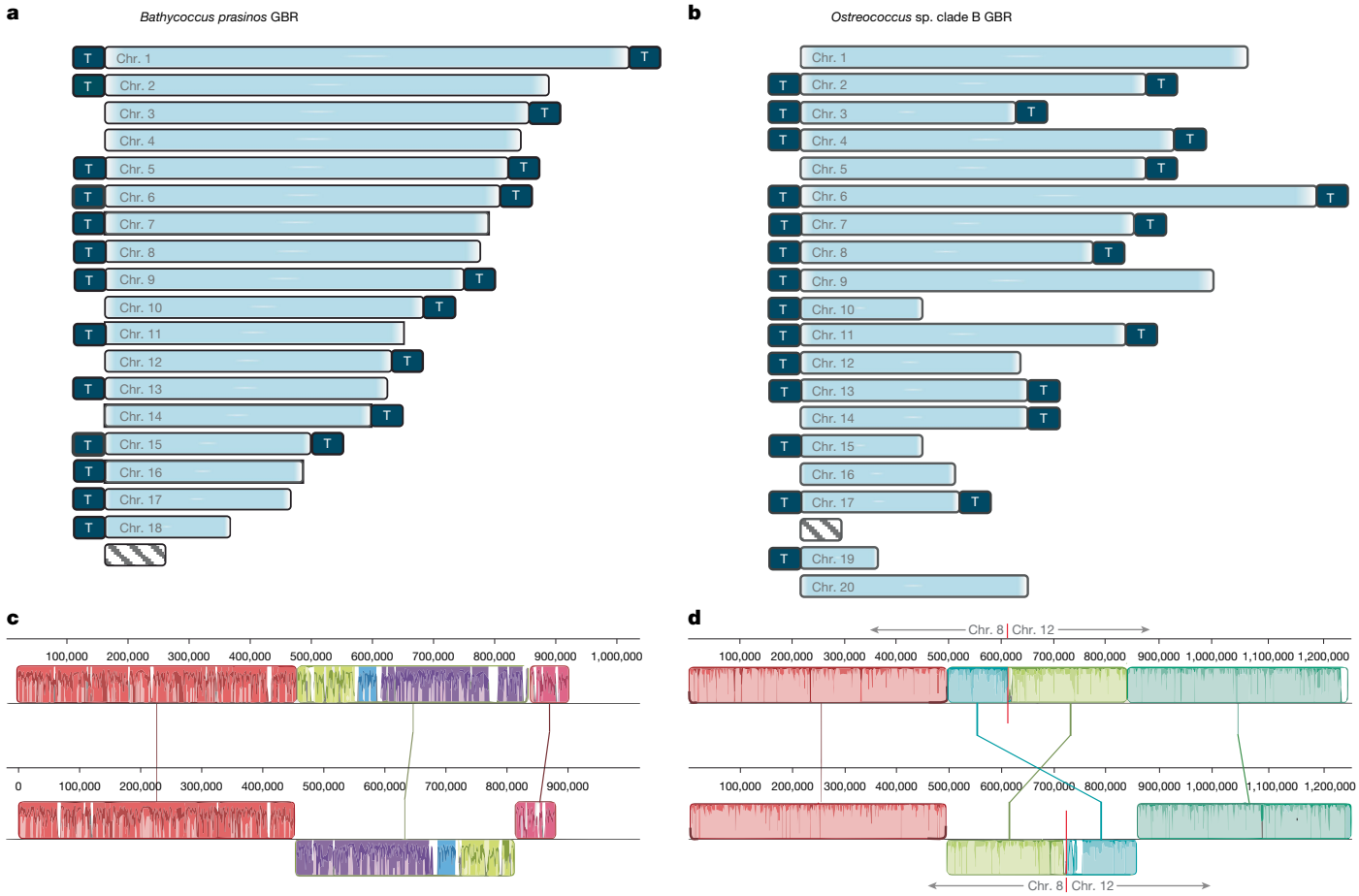


Fig. 5 | Chromosome-level eMAGs from marine picoeukaryotes. a, Composite illustration of the presence of telomeres (T) in GBR-MGD *Bathycoccus* eMAGs. Presence is noted in dark blue if telomeric repeats were found at the ends of chromosomes (Chr.) from at least one of the five eMAGs. **b**, Composite illustration of the presence of telomeres in GBR-MGD *Ostreococcus* sp clade B eMAGs. Presence is noted in dark blue if telomeric repeats were found at the ends of chromosomes from at least one of the 15 eMAGs. **c**, Sequence alignment of chromosome 4 for *Bathycoccus prasinos* RCC1105 (top) and GBR-MGD *Bathycoccus* variant 2 (bottom) from Grub Reef showing a large approximately 350 kb inversion in variant 2 compared with the reference. **d**, Sequence alignment of chromosomes 8 and 12 from *Ostreococcus* sp. clade B RCC809 (top) and GBR-MGD *Ostreococcus* sp clade B (bottom) from Masthead Reef showing an approximately 350 kb translocation between the ends of the two chromosomes. The striped chromosomal element in **a**,**b** represent the small outlier chromosome, which could not be recovered owing to its highly divergent sequence composition.

hosts[41]. Further work is necessary to delineate the global distribution and ecological characteristics of marine *Crassvirales*.

Recovery of picoeukaryote genomes

Marine phytoplankton are responsible for around half of all primary production on the planet[8], and although picoeukaryotes typically make up a smaller fraction of the phytoplankton community compared with prokaryotes, they exhibit much higher growth and turnover rates and often account for the majority of primary productivity[63,64]. Despite their ubiquity in marine ecosystems[12,65], much of our knowledge of these groups relies on analysis of a few cultured isolates, none of which are from coral reefs or Australian oceans[66–68]. Researchers have recently begun to recover genomes from uncultured representatives via metagenomics[12]. However, as picoeukaryotes have more complex genomes than their prokaryotic counterparts, including multiple chromosomes, large genome rearrangements and more extensive repeat regions than prokaryotes, the few available short-read-based eMAGs are mostly fragmented and incomplete[12].

Using Nanopore long reads, we were able to obtain telomere-containing (Supplementary Table 20) chromosome-level eMAGs from the GBR for the picoeukaryote genera *Bathycoccus* and *Ostreococcus* (green algae in the order Mamiellales, phylum Chlorophyta), taxa that are ubiquitously distributed in the world's oceans[63] (Fig. 5a,b). In total, five *Bathycoccus* eMAGs were recovered and found to represent two distinct genome variants, BV1 and BV2, with BV1 being closely related to *Bathycoccus prasinos* RCC1105 (96% ANI) and BV2 representing a more distantly related lineage with only 82% ANI to RCC1105, along with a large approximately 350 kb inversion on chromosome 4 that is not present in BV1 (Fig. 5c and Extended Data Figs. 8 and 9). Mapping of the metagenomic reads to representatives of each variant (BV1, reef 11049; BV2, Grub Reef) revealed that they had similar relative abundance across most samples (average 0.07 ± 0.11% for BV1 and 0.08 ± 0.08% for BV2), except for samples collected from TO in which BV2 was about sevenfold more abundant (average 0.47%, maximum 2.4%; Supplementary Table 21). Previous marker gene-based studies suggest the existence of multiple *Bathycoccus* ecotypes that may represent different species[65,69], though the genomic underpinnings of ecotype ecology are unclear and may be clarified through the recovery of additional chromosome-level genomes.

In contrast to *Bathycoccus*, all 15 GBR *Ostreococcus* eMAGs represent a single species that is closely related to *Ostreococcus* RCC809 (>99% ANI to *Ostreococcus* RCC809), a member of the clade B lineage, for which little genomic information is available. Despite high similarity to RCC809, all GBR *Ostreococcus* clade B eMAGs have an approximately 350 kb translocation between chromosomes 8 and 12 (Fig. 5d and Extended

Data Fig. 10), indicating that high sequence similarity can hide major structural variation detectable via chromosome-level assemblies. After *Parasynechococcus* sp002724845, *Ostreococcus* clade B represents the second most abundant microbial species in the GBR-MGD (average 3.5%, maximum 30%; Supplementary Table 21). The ability to readily recover eMAGs with long reads, especially to chromosome-level quality, is a substantial advance that we hope will encourage others to explore uncultured eukaryotic dark matter.

Discussion

Metagenomics has greatly enabled the study of microbial ecosystems by bypassing the cultivation bottleneck. However, marine ecosystems are unusual in that the genomes of many dominant microbial populations are conspicuously underrepresented in short-read MAG datasets. Prominent examples include the bacterial genera *Pelagibacter*, *Prochlorococcus* and *Synechococcus*, which are all underrepresented in international consortium-wide datasets. We show that this phenomenon is owing to limitations of short-read sequencing and assembly in handling the combination of high strain heterogeneity and low GC content of these taxa. By contrast, the Nanopore platform addresses both issues by being able to sequence low-GC templates and to traverse strain variable regions. However, Nanopore sequencing presents its own challenges, such as higher error rates in the raw reads and lower data output compared with Illumina platforms. Therefore, hybrid sequencing approaches that include long reads will produce the most comprehensive and high-quality genome inventory of marine microbial taxa using currently available technologies.

The GBR-MGD will be a highly valuable resource for marine researchers, not only because of its inclusion of dominant under-represented prokaryotic taxa, but also because it combines metagenomic data including high-quality prokaryote, viral and eukaryote genomes with seawater physicochemical measurements and metadata from the Australian Institute of Marine Science (AIMS) Long Term Monitoring Program (LTMP). We demonstrate the utility of this connection to explain the underlying physicochemical drivers of microbial community composition on the GBR and bioindicators of reef protection status. We expand the known diversity of marine *Crassvirales* and note that many of the GBR-MGD sequences form a high GC cluster, which is atypical for this viral order, and link *Crassvirales* to Bacteroidota that have important roles in nutrient cycling[70] and, more recently, are gaining attention as emerging pathogens in marine ecosystems[71]. We also demonstrate that chromosome-level genome recovery is possible for microbial eukaryotes, which comprise up to 30% of pelagic GBR microbial communities.

We suggest that long-read sequencing should become a standard part of metagenomic analyses of marine microbial communities, as many key populations are poorly accessible via short-read sequencing alone. As low-GC taxa and strain heterogeneity are not limited to marine environments[24,72–74], we predict that the sequencing limitations observed in this study will affect metagenomic analyses of other ecosystems that could be addressed through the incorporation of long reads. Finally, we suggest that integration of long-read metagenomics (biology) and physicochemistry measurements (nutrients) with reef monitoring will provide the necessary framework to enable adaptive reef management[34,75].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10778-z.

1. 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).
2. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**, e3558 (2017).
3. Biller, S. J. et al. Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 180176 (2018).
4. Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
5. Nishimura, Y. & Yoshizawa, S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).
6. Cavicchioli, R. et al. Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* **17**, 569–586 (2019).
7. Azam, F. et al. The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).
8. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
9. Roberts, C. M. et al. Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science* **295**, 1280–1284 (2002).
10. Haro-Moreno, J. M. et al. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. *Microbiome* **6**, 128 (2018).
11. Sunagawa, S. et al. *Tara* Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
12. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom.* **2**, 100123 (2022).
13. Gaïa, M. et al. Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
14. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
15. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).
16. Duncan, A. et al. Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome* **10**, 67 (2022).
17. Alexander, H. et al. Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio* **14**, e01676-23 (2023).
18. Laiolo, E. et al. Metagenomic probing toward an atlas of the taxonomic and metabolic foundations of the global ocean genome. *Front. Sci.* **1**, 1038696 (2024).
19. 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).
20. Lombard, F. et al. Open science resources from the *Tara* Pacific expedition across coral reef and surface ocean ecosystems. *Sci. Data* **10**, 324 (2023).
21. Sereika, M. et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
22. Delmont, T. O. et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* **8**, e46497 (2019).
23. Huber, H., Hohn, M. J., Stetter, K. O. & Rachel, R. The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life. *Res. Microbiol.* **154**, 165–171 (2003).
24. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
25. Browne, P. D. et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience* **9**, giaa008 (2020).
26. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
27. Pachiadaki, M. G. et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623–1635.e11 (2019).
28. Sowell, S. M. et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* **3**, 93–105 (2009).
29. Sosa, O. A., Repeta, D. J., DeLong, E. F., Ashkezari, M. D. & Karl, D. M. Phosphate-limited ocean regions select for bacterial populations enriched in the carbon–phosphorus lyase pathway for phosphonate degradation. *Environ. Microbiol.* **21**, 2402–2414 (2019).
30. Moreno-Vivián, C., Cabello, P., Martínez-Luque, M., Blasco, R. & Castillo, F. Prokaryotic nitrate reduction: molecular properties and functional distinction among bacterial nitrate reductases. *J. Bacteriol.* **181**, 6573–6584 (1999).
31. Benedetti-Cecchi, L. et al. Marine protected areas promote stability of reef fish communities under climate warming. *Nat. Commun.* **15**, 1822 (2024).
32. Great Barrier Reef Marine Park Authority. *Great Barrier Reef Marine Park Zoning Plan 2003* (Australian Government, 2004).
33. Terzin, M. et al. The road forward to incorporate seawater microbes in predictive reef monitoring. *Environ. Microbiome* **19**, 5 (2024).
34. Apprill, A. & Salerno, J. L. Reef water microorganisms as diagnostic indicators for coral reef ecosystem management and sustainability. *Cell Rep. Sustain.* **2**, 100403 (2025).
35. Viklund, J., Martijn, J., Ettema, T. J. G. & Andersson, S. G. E. Comparative and phylogenomic evidence that the Alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PLoS ONE* **8**, e78858 (2013).
36. 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. *Environ. Microbiol.* **25**, 1136–1154 (2023).
37. Martinez-Gutierrez, C. A. & Aylward, F. O. Strong purifying selection is associated with genome streamlining in epipelagic Marinimicrobia. *Genome Biol. Evol.* **11**, 2887–2894 (2019).
38. Grote, J. et al. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**, e00252-12 (2012).
39. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).

40. Roach, T. N. F. et al. Microbial bioenergetics of coral–algal interactions. *PeerJ* **5**, e3423 (2017).
41. Haas, A. F. et al. Global microbialization of coral reefs. *Nat Microbiol* **1**, 16042 (2016).
42. Dinsdale, E. A. & Rohwer, F. in *Coral Reefs: An Ecosystem in Transition* (eds Dubinsky, Z. & Stambler, N.) 231–240 (Springer, 2011).
43. Lauro, F. M. et al. The genomic basis of trophic strategy in marine bacteria. *Proc. Natl Acad. Sci. USA* **106**, 15527–15533 (2009).
44. Turner, D. et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **168**, 74 (2023).
45. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
46. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2023).
47. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
48. Tian, F. et al. Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. *Microbiome* **12**, 159 (2024).
49. Wang, T. et al. Characterization and genomic analysis of an oceanic cyanophage infecting marine *Synechococcus* reveal a novel genus. *Front. Microbiol.* **14**, 1231279 (2023).
50. Du, S. et al. Genome sequences of the first *Autographiviridae* phages infecting marine *Roseobacter*. *Microb. Genom.* **10**, 001240 (2024).
51. Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. & Chisholm, S. W. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* **3**, e144 (2005).
52. Pope, W. H. et al. Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a "horned" bacteriophage of marine *Synechococcus*. *J. Mol. Biol.* **368**, 966–981 (2007).
53. Zhao, Y. et al. Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013).
54. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
55. Guerin, E. et al. Biology and taxonomy of CrAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
56. Shkoporov, A. N. et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **9**, 4781 (2018).
57. Yutin, N. et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2017).
58. Edwards, R. A. et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* **4**, 1727–1736 (2019).
59. Yutin, N. et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* **12**, 1044 (2021).
60. Papudeshi, B. et al. Host interactions of novel *Crassvirales* species belonging to multiple families infecting bacterial host, *Bacteroides cellulosilyticus* WH2. *Microb. Genom.* **9**, 001100 (2023).
61. Bartlau, N. et al. Highly diverse flavobacterial phages isolated from North Sea spring blooms. *ISME J.* **16**, 555–568 (2022).
62. Piedade, G. J. et al. Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape. *Nat. Commun.* **15**, 9192 (2024).
63. Li, W. K. W. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol. Oceanogr.* **39**, 169–175 (1994).
64. Worden, A. Z., Nolan, J. K. & Palenik, B. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnol. Oceanogr.* **49**, 168–179 (2004).
65. Simmons, M. P. et al. Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North Pacific Ocean. *Appl. Environ. Microbiol.* **82**, 1693–1705 (2016).
66. Derelle, E. et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl Acad. Sci. USA* **103**, 11647–11652 (2006).
67. Moreau, H. et al. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13**, R74 (2012).
68. Palenik, B. et al. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA* **104**, 7705–7710 (2007).
69. Joli, N., Monier, A., Logares, R. & Lovejoy, C. Seasonal patterns in Arctic prasinophytes and inferred ecology of *Bathycoccus* unveiled in an Arctic winter metagenome. *ISME J.* **11**, 1372–1385 (2017).
70. Fernández-Gómez, B. et al. Ecology of marine Bacteroidetes: a comparative genomics approach. *ISME J.* **7**, 1026–1037 (2013).
71. Hudson, J. & Egan, S. Opportunistic diseases in marine eukaryotes: could Bacteroidota be the next threat to ocean life? *Environ. Microbiol.* **24**, 4505–4518 (2022).
72. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
73. Kazantseva, E., Donmez, A., Frolova, M., Pop, M. & Kolmogorov, M. Strainy: phasing and assembly of strain haplotypes from long-read metagenome sequencing. *Nat. Methods* **21**, 2034–2043 (2024).
74. Ma, B. et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
75. Gilbert, J. A. et al. Safeguarding microbial biodiversity: microbial conservation specialist group within the species survival commission of the International Union for Conservation of Nature. *ISME J.* **19**, wraf239 (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© Crown 2026

Methods

Sample collection

Samples were collected under the permit G12/35236–1 issued by the Great Barrier Reef Marine Park Authority. Seawater samples were collected from 48 coral reefs spanning the length of the GBR, aligning with in situ coral surveys by the AIMS LTMP in 2019 and 2020 (Fig. 1). Four 5-l biological replicates of seawater were collected above the reef (2–10 m depth) at each site (total $n = 192$) in either Nalgene or Niskin bottles and placed on ice until return to the vessel. Each 5 l volume was independently filtered successively through 5-µm, 33 mm diameter cellulose acetate and 0.22-µm Sterivex-GP filters (Millipore). The Sterivex filters were then snap frozen and stored at −75 °C until DNA extraction.

In addition to metagenomics data, a total of 17 physicochemical variables were measured as part of this survey on 3 of the 4 replicates (5 l each) from each site, including ammonia (NH_4^+), nitrite (NO_2^-), nitrate (NO_3^-), total dissolved nitrogen, phosphate (PO_4^{3-}), total dissolved phosphorus, dissolved organic carbon, silicate, total suspended solids, chlorophyll *a* pigment concentration, phaeophytin A, particulate organic carbon, particulate nitrogen, particulate phosphorus, temperature, salinity and chlorophyll *a* fluorescence, using methods described in Terzin et al.[76].

DNA extraction

DNA extractions were carried out inside each 0.22-µm Sterivex filter by adding 1.8 ml filter-sterilized lysis buffer containing 50 mM Tris-HCL (pH 8.0), 40 mM EDTA (pH 8.0), 256 mg ml^{-1} sucrose and 18 µl lysozyme (100 mg ml^{-1}). The filters were incubated with mild agitation for 1 h at 37 °C, followed by the addition of 20 µl proteinase K (20 mg ml^{-1}) and further incubation and agitation for 1 h at 55 °C. The lysate was then ejected through the Sterivex into a 5 ml microtube and an equal volume of phenol:chloroform:isoamyl alcohol (IAA; 25:24:1) was added, mixed by inversion and centrifuged at 16,000*g* for 10 min. The aqueous phase was then recovered and an equal volume of chloroform:IAA (24:1) was added and mixed by inversion, then centrifuged for 10 min at 16,000*g*. DNA was precipitated by adding 1 ml isopropanol and recovered by centrifugation at 20,000*g* for 25 min and the supernatant discarded. The DNA was washed with 500 µl 70% ethanol, centrifuged for 10 min, the ethanol discarded and the pellet air-dried. Finally, 20 µl PCR water was added, incubated at 4 °C overnight to resuspend the DNA pellet and frozen at −20 °C.

Illumina sequencing

All 191 replicates from the 48 sites were subjected to Illumina sequencing at the Centre for Microbiome Research (Brisbane, Australia). Libraries were prepared using the Nextera DNA Flex Library Preparation Kit (Illumina 20018705) according to manufacturer's protocol. Libraries were pooled at equimolar amounts of 2 nM per library. The library pool was quantified in triplicates using a Qubit dsDNA HS Assay Kit (Invitrogen) and sequenced using an S4 2×150 bp flow cell on a NovaSeq 6000. Additional 'deep' sequencing was carried out for one replicate from 27 of the 48 sites to a target depth of 40 Gbp for Nanopore polishing and hybrid assembly (Supplementary Tables 1 and 2).

Nanopore sequencing, basecalling and quality control

The replicate chosen for Illumina deep sequencing was also submitted for long-read sequencing on Oxford Nanopore Technology's (ONT) PromethION platform at the Australian Centre for Ecogenomics (Brisbane, Australia) sequencing facility. For ONT sequencing, DNA was first size-selected using the Circulomics SRE XS kit (PacBio, SKU 102-208-200) as this was shown in a small trial to increase total data output and read N50. Native barcoded libraries were prepared from the size-selected DNA following a genomic DNA by ligation protocol (Oxford Nanopore, LSK-109 with EXP-NBD104). Barcoded libraries were pooled with two samples per PromethION flow cell at equal concentration and the library pool was sequenced on a PromethION 24 (Oxford Nanopore) for a total of 72 h with a R9.4 flow cell using MinKNOW (v20.06.18) with default settings. We then performed rebasecalling of the fast5 files using the superaccuracy model with Guppy (v5.0.16) to generate higher accuracy raw reads before assembly. Finally, Porechop (https://github.com/rrwick/Porechop; v0.2.4) was used for barcode trimming using default settings.

Hybrid metagenome assembly

Deep Illumina and nanopore reads for each of the 27 sites were assembled using the Aviary assembly pipeline (https://github.com/rhysnewell/aviary; v0.3.3), specifying the ont_HQ flag for long reads. In brief, by supplying both short and long reads for hybrid assembly, Aviary first assembled the nanopore reads with metaFlye[77] (v2.9), then polished the metaFlye assembly using a series of Racon[78] (v1.4.3; three rounds), Pilon[79] (v1.23; one round), and Racon (one round). Next, a subset of high-coverage, polished metaFlye contigs were identified and Illumina reads from the same replicate were mapped to them. Illumina reads that did not map to these contigs were assembled using metaSPAdes[80] (v3.15.3) together with the nanopore reads. Finally, the hybrid-assembled contigs were collated with the high coverage polished metaFlye contigs to produce the final assembly. For the remaining 21 sites not chosen for Nanopore sequencing, a single replicate was assembled using the Aviary pipeline which used metaSPAdes for Illumina-only assembly.

Prokaryote genome binning and analysis

Assemblies were binned using the Aviary pipeline's recover workflow, which first mapped reads to each assembly using Minimap2[81] (v2.18). For differential coverage binning, three sites were chosen for mapping to each assembly based on geographical proximity and all nanopore and Illumina reads for all four replicates from each site were used (Supplementary Table 22). Aviary recover then ran binning programs MetaBAT[82], MetaBAT 2 (ref. 83) (v2.15), MaxBin 2.0 (ref. 84) (v2.2.7), Rosella (https://github.com/rhysnewell/rosella; v0.3.3), CONCOCT[85] (v1.1.0) and Vamb[86] (v4.0.0) and included a bin refinement step through Rosella to further refine bins based on kmer frequency and coverage. Finally, Das Tool[87] (v1.1.2) was used to collate a representative set of bins from across the different binning tools. pMAGs were retained if they passed a quality threshold ≥50 (quality = completeness − 3 × contamination) using either CheckM1 (ref. 88) (v1.2.1) or CheckM2 (ref. 89) (v1.0.2), such that pMAGs of lower completeness were allowed lower contamination scores for inclusion. Taxonomic classification of the pMAGs was performed using the Genome Taxonomy Database Toolkit[90] (v2.3.0) with database release R214. Dereplication of MAGs for mapping, and for assessment of presence within previously published databases and data sets including the OMD[4] and Australian National Reference Stations[91], was carried out using CoverM[92] cluster (v0.6) at 95% ANI. Relative abundance was calculated by mapping the Illumina reads to the dereplicated set of reference pMAGs using CoverM genome, requiring 95% identity and 75% read length alignment to consider a read mapped.

Benchmarking hybrid assembly MAG binning

To directly benchmark the effect of including long nanopore data on improving MAG recovery, we undertook separate re-assembly and binning of eight replicates spanning sites from north to south of the GBR. To remove the effect of sequencing depth, we compared two treatments: (1) Illumina-only assembly of 30 Gbp of sequence data using metaSPAdes; and (2) a separate assembly using a subset of 15 Gbp of Illumina reads combined with 15 Gbp of nanopore reads as described above for hybrid assembly. The resulting 16 assemblies were binned with Aviary recover to obtain pMAGs as described above. Genomic diversity of the recovered pMAGs was estimated with inStrain[24] (v1.7.1) by mapping Illumina reads to their respective metagenome assemblies using

Bowtie 2 (ref. 93) (v2.4.5) followed by CoverM filter to select mapped reads with at least 95% identity and 75% read alignment.

Microbial community composition and function

Read mapping counts to the 95% ANI dereplicated set of pMAGs and relative abundances generated using CoverM genome described above were used to visualize community composition and clustering through principal component analysis ordinations in R (ref. 94) (v4.4.1) using the vegan[95] package (v2.6.8). Relative abundances were summarized to genus and other higher taxonomic ranks using phyloseq[96] (v1.48.0). Associations between community composition, GBR sectors, and reef zoning were assessed using PERMANOVA implemented in vegan, while associations between pMAGs with NTMR and Fished reefs were assessed using generalized linear models implemented in MaAsLin 3 (ref. 97) (v0.99.4) controlling for variability across sectors and Benjamini–Hochberg false discovery rate correction for multiple comparisons. Figures showing maps of the GBR and zoning boundaries (for example, Fig. 1) were drawn in R using Geographic Information System data available in the gisaimsr (v0.0.1) R package (https://open-aims.github.io/gisaimsr/index.html).

To assess functional potential of microbial communities in IN and TO sectors, genes were first predicted from IN and TO metagenome assemblies using Pyrodigal[98] (v3.7.0) followed by clustering at 95% identity using MMseqs2 (ref. 99) (v13.45111) to create a non-redundant gene catalogue. Trimmed sequence reads from each IN and TO metagenome were then mapped to this catalogue using CoverM make and read mapping counts subsequently calculated from the output bam files using Samtools[100] (v1.21). Representative sequences in the non-redundant catalogue were assigned to Kyoto Encyclopedia of Genes and Genomes Orthology using KofamScan[101] (v1.3.0). Differential abundance statistics on the read mapping counts from CoverM and Samtools were generated using DESeq2 (ref. 102) (v1.44) in R.

Bioindicators of reef zoning

To identify bioindicators of NTMR and Fished reefs, we first performed feature selection using MINT sPLS-DA[103,104] implemented in mixOmics[105] (v6.26.0) to account for sectors on the GBR, using centred log ratio-transformed abundances of the 95% ANI dereplicated set of 876 pMAGs, 362,802 vOTUs and 1,136,487 eukaryote contigs (see 'Final GBR-MGD database compilation'). Random forest (ranger R package v0.17.0) was then used on the selected features to assess accuracy in predicting reef protection zones (average accuracy over 50 runs), using leave one group out cross validation to account for sectors on the GBR. This analysis was then repeated on pMAGs alone with the first MINT sPLS-DA component consisting of the top 60 selected indicator taxa. Differences in average genome size and GC content between indicators of NTMRs and Fished reefs were visualized as box plots using ggplot2 (v3.5.1), and the variation in genome size and GC content was compared between indicators of NTMRs and fished reefs with pairwise Wilcoxon rank-sum tests.

To identify environmental drivers that explained differences in indicator taxa between NTMR and Fished reefs, the 60 indicator pMAGs of reef zoning were correlated with each of the 17 physicochemical variables using MINT sPLS. sPLS[106,107] models the relationship between multiple predictors (physicochemical variables) and multiple responses (microbial taxa), while MINT[103] integrates samples from independent subsets to account for confounding effects, such as season and geography. Median values for each physicochemical variable were first calculated per reef site, as the number of replicates differed for molecular ($n = 4$) and water chemistry ($n = 3$) samples. We then correlated centred log ratio-transformed abundances of the 876 dereplicated pMAGs with the 17 physicochemical measurements. Similarity scores (partial correlations) were visualized in ggplot2 (v3.5.1). All analyses were implemented in R. Final composite plots were edited for clarity in Inkscape (v0.92.5).

Viral, plasmid and eukaryote contigs

The following viral prediction tools were used to identify putative viral contigs from the GBR-MGD metagenomic assemblies: GeNomad[46] (v1.5.018; --enable-score-calibration --min-score=0.7 --max-score=0.1) with database v1.2, VirSorter2 (ref. 47) (v2.2.431; --include-groups "dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae", --min-score=0.9) with database v1.2, ViralVerify[108] (v1.133; -thr 7), VIBRANT[109] (v1.2.132), PPR-Meta[110] (v1.134; -t 0.9), and DeepVirFinder[111] (not versioned; retrieved from GitHub on 8 March 2023; P value < 0.05). Putative viral contigs were then input into CheckV[112] (v1.0.1) with database v1.5 for quality trimming and assessment. Viral predictions were examined using CheckV's curated marker sets by visualizing the ratio of host:viral markers and the proportion of viral markers for each putative virus prediction to assess whether the contigs predicted by each tool were host- or virus-derived.

In addition to three viral prediction tools that also predicted plasmids (GeNomad, PPRMeta and ViralVerify), two additional plasmid-only prediction tools were used on the metagenomic assemblies, PlasForest[113] (v1.2), and PlasX[114] (unversioned; score ≥0.9), which uses annotated gene families generated in anvi'o[115] (v7.1) using the COG2014 and Pfam v32 databases. Whokaryote[116] (v1.1.2) and EukRep[117] (v0.6.6) were used for eukaryotic prediction.

Manual curation of picoeukaryote genomes

Genomes of two picoeukaryotes from the Mamiellales family, *Bathycoccus prasinos* and *Ostreococcus* sp. clade B, were manually curated after initial blast results of partial eMAGs revealed the presence of potential *Bathycoccus* and *Ostreococcus* long contigs in high abundance as determined by read mapping. Therefore, complete genomes for the sequenced Mamiellales isolates *B. prasinos* RCC1105, *Micromonas commoda* RCC299, *Micromonas pusilla* CCMP1545, *Ostreococcus lucimarinus* CCE9901, *Ostreococcus tauri* and *Ostreococcus* spp. RCC809 were retrieved from the RefSeq database and used as references for competitive mapping of large metagenomic contigs (≥50 kb) with Minimap2 (ref. 81) (v2.28) using its genome/assembly alignment mode, allowing up to 20% sequence divergence (--asm20). The majority of contigs were recruited by either *B. prasinos* RCC1105 or *Ostreococcus* spp. RCC809, and were selected for manual curation of two distinct species using the RCC1105 and RCC809 genomes as references. Contigs recruited by *Ostreococcus* spp. RCC809 also showed high recruitment to *O. lucimarinus* CCE9901, which was therefore used as a secondary reference as needed. Contigs were aligned to the corresponding reference using progressiveMauve from the Mauve aligner[118] (v2015-02-13). In the few cases where a chromosome was recovered in a sample fragmented in 2–4 contigs, Mauve Contig Mover from the Mauve program was used to reorder and link the chromosomal fragments with a linker of 100 Ns. Telomeres were identified at the ends of chromosomal contigs using Tidk[119] (v0.2.65) to search for telomeric repeat sequence AACCCT, followed by visual inspection to confirm elevated numbers of telomeric repeats at the ends of contigs using Tidk plot.

Marine *Crassvirales*

vMAGs classified by GeNomad as *Crassvirales* were annotated using Pharokka[120] (v1.6.1) with additional *Crassvirales* protein profiles provided by Yutin et al.[59]. To verify the GeNomad classifications, we downloaded the set of *Caudoviricetes* TerL sequences compiled by Piedade et al.[62] and aligned TerL sequences detected in the vMAGs (>400 amino acid residues) to this collection using MAFFT[121] (v7.508). The output sequence alignment was trimmed with trimAl[122] (v1.4.rev15; -gt 0.2) and used to infer a bootstrapped phylogenetic tree using IQ-TREE 2 (ref. 123) (v2.2.0.3). Separately, we constructed a proteomic tree using the vMAGs (>20 kb) based on genome-wide sequence similarities computed using tBLASTx as implemented in ViPTree[124] (v4.0). Both trees were visualized and edited for clarity in ITOL[125] (v7) and Inkscape (v0.92.5). Host

prediction was performed on vMAGs using iPHoP[126] (v1.3.3) with the supplied Aug_2023_pub_rw database including pMAGs dereplicated at 99% ANI (using CoverM) added to this host database.

Final GBR-MGD database compilation

The GBR-MGD database is composed of prokaryotic MAGs, manually curated Mamiellales genomes, and contigs identified as plasmid, viral, or eukaryotic. A standardized labelling scheme with the structure of IMOS_<TaxaGroup>_SAMPLENAME_CONTIG:<TaxaComponent> was applied to all sequences. TaxaGroup indicates all taxa that were linked to that contig (E = eukaryote, M = plasmid/mobile element, P = prokaryote, V = virus). Contigs marked with only one group were only identified to belong in that group, while contigs marked with multiple letters indicate they were linked to more than one group. For example, a contig marked with a V indicates that the contig was identified as a virus but is not found in another group. PV indicates that a contig is present in a pMAG and predicted to be a virus or provirus, MPV indicates that a contig is found in a pMAG and predicted to be a plasmid and virus. As viruses can be predicted within other taxa (for example, PV contigs), TaxaComponent indicates the taxa represented by a particular sequence as well as the bp position in the case of some viral sequences. For example, a proviral region identified within a prokaryotic bin would be represented by two distinct sequences. The sequence within the prokaryotic bin would be labelled IMOS_PV_SAMPLE_CONTIG:p while its counterpart representing the viral prediction would be labelled IMOS_PV_SAMPLE_CONTIG:v1001-10000. Additionally, contigs are labelled with the prediction tools used to group them into their respective TaxaGroup.

The prokaryotic bins and manually curated *Mamiellales* genomes were dereplicated together using CoverM. Contig predictions were dereplicated separately for each taxon group (for example, V, P, E or PV) using pairwise ANI (95% ANI + 85% alignment fraction) using the anicalc.py and aniclust.py scripts provided by CheckV. Relative abundance of each component of the database was calculated by mapping the Illumina metagenomic reads to each taxon group in the final dereplicated database separately using CoverM Genome, requiring 95% identity and 75% alignment to be considered mapped.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The GTDB R214 database was downloaded from https://data.gtdb.ecogenomic.org/releases/release214/214.0/. MAGs contained in the OMD v1 were downloaded from https://sunagawalab.ethz.ch/share/microbiomics/ocean/suppl_data/. Metagenomic sequence datasets from the Australian National Reference Stations are available through the Bioplatforms Australia Data Portal (https://data.bioplatforms.com/) by searching for keyword 'NRS' and filtering to 'illumina-shortread' sequence data type. Genomes of sequenced *Bathycoccus* and *Ostreococcus* are available at RefSeq (https://www.ncbi.nlm.nih.gov/refseq/). *Caudoviricetes* TerL sequences and *Crassvirales* genomes compiled in Piedade et al.[62] were downloaded from https://dataverse.nioz.nl/dataset.xhtml?persistentId=doi:10.25850/nioz/7b.b.wg. Conserved *Crassvirales* genes from Yutin et al.[59] were downloaded from https://ftp.ncbi.nlm.nih.gov/pub/yutinn/crassfamily_2020/. The COG2014 database is available at https://ftp.ncbi.nlm.nih.gov/pub/COG/COG2014/data/ and the Pfam32 database is available at https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/. Raw sequence data produced in this study are available at the European Nucleotide Archive under project accession PRJEB82623 and secondary accession ERP166296 with project name 'Great Barrier Reef seawater microbiomes genome database'. All Illumina and Nanopore sample identifiers are provided in Supplementary Table 23. The 5,283 pMAGs, 15 *Ostreococcus* clade B genomes, 5 *Bathycoccus* genomes, 27 *Crassvirales* and group contigs are available at Zenodo.org (https://doi.org/10.5281/zenodo.17109886 (ref. 127)).

Code availability

The R code used for the analyses performed in this study is archived at Zenodo (https://doi.org/10.5281/zenodo.17109886)[127].

76. 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).
77. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
78. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
79. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
80. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
81. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
82. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
83. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
84. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
85. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
86. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
87. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
88. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
89. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
90. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
91. Lynch, T. P. et al. IMOS National Reference Stations: a continental-wide physical, chemical and biological coastal observing system. *PLoS ONE* **9**, e113652 (2014).
92. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).
93. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
94. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2024).
95. Oksanen, J. et al. vegan: Community Ecology Package. R package version 2.6-8 https://CRAN.R-project.org/package=vegan (2024).
96. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
97. Nickols, W. A. et al. MaAsLin 3: refining and extending generalized multivariable linear models for meta-omic association discovery. *Nat. Methods* **23**, 554–564 (2026).
98. Larralde, M. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *J. Open Source Softw.* **7**, 4296 (2022).
99. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
100. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
101. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
102. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
103. 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 Bioinform.* **18**, 128 (2017).
104. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform.* **12**, 253 (2011).
105. 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).
106. 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 Bioinform.* **10**, 34 (2009).
107. 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**, 35 (2008).
108. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).

109. 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).
110. Fang, Z. et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* **8**, giz066 (2019).
111. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
112. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
113. Pradier, L., Tissot, T., Fiston-Lavier, A.-S. & Bedhomme, S. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinform.* **22**, 349 (2021).
114. 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).
115. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
116. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb. Genom.* **8**, mgen000823 (2022).
117. 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).
118. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
119. Brown, M. R., Manuel Gonzalez De La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, btaf049 (2025).
120. Bouras, G. et al. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2023).
121. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
122. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
123. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
124. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
125. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
126. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
127. Robbins, S. et al. Great Barrier Reef Microbial Genomes Database (GBR-MGD). *Zenodo* https://doi.org/10.5281/zenodo.20503966 (2026).

Acknowledgements The seawater samples in this study 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 recognize 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 Farquharson Reef. We honour the Girringun Aboriginal Corporation Traditional Use of Marine Resources Agreement (TUMRA) for their connection to Taylor Reef, and the Manbarra TOs of Rib, Kelso, Little Kelso and John Brewer reefs. We also recognize the Wulgurukaba TOs of Myrmidon, Grub and Helix reefs, the Bindal TOs of Knife, Fork, Centipede, Chicken and Lynchs reefs, and the continuing connection of the Manbarra TOs to Roxburgh, and Fore and Aft reefs. Finally, we acknowledge the Port Curtis Coral Coast TUMRA for their stewardship of North, Bloomfield, Erskine, 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. This research was undertaken at the Australian Institute of Marine Science (AIMS), and we wish to acknowledge the Wulgurukaba and Bindal peoples as the TOs of the land on which AIMS headquarters at Cape Ferguson is located. This research was also undertaken at the James Cook University Townsville Bebegu Yumba campus. We 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 thank the AIMS Water Quality team, especially U. Bove, K. Glasson and D. Moran for logistics, training and processing of water chemistry samples; the AIMS LTMP team and others involved in field collection and preparation of samples including M. Emslie, E. Botté, J. Davidson, V. Mocellin and J. Nielsen; and the crew of the RV *Solander* and RV *Cape Ferguson* for their excellent logistical support in the field.

Author contributions N.S.W. obtained the Queensland Government RICF funding for the project. N.S.W. and S.R. conceived the sampling design. S.C.B. collected seawater in the field and processed samples in the laboratory for metagenomic sequencing. R.K.G. generated water chemistry data. S.R., K.D., J.P.E., J.Z., P.H. and Y.K.Y. assembled the metagenomes, binned MAGs, benchmarked metagenome assemblies and performed contig taxa predictions. P.H. and Y.K.Y. analysed microbial community composition and function. M.T., P.W.L., K.-A.L.C., D.G.B. and Y.K.Y. identified indicators of reef zoning. Y.K.Y. analysed the *Crassvirales* genomes. K.D. curated and analysed the picoeukaryotic genomes. S.R. and P.H. wrote the original draft, and P.H. and Y.K.Y. revised the manuscript. All authors reviewed the submitted manuscript and approved the final version.

Funding This study is part of the Great Barrier Reef Microbial Genomics Database sub-Facility within the Marine Microbiome Initiative Facility of Australia's Integrated Marine Observing System (IMOS), and was funded by a Research Infrastructure Co-investment Fund (RICF) from the Queensland Government Department of Environment and Science. 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 AIMS and an AIMS@JCU PhD Scholarship to M.T. K.-A.L.C. was supported by a National Health and Medical Research Council Investigator Grant (GNT2025648). The funders had no role in sampling design, data collection, processing and interpretation, preparation of the manuscript, or decision to publish.

Competing interests The authors declare no competing interests.

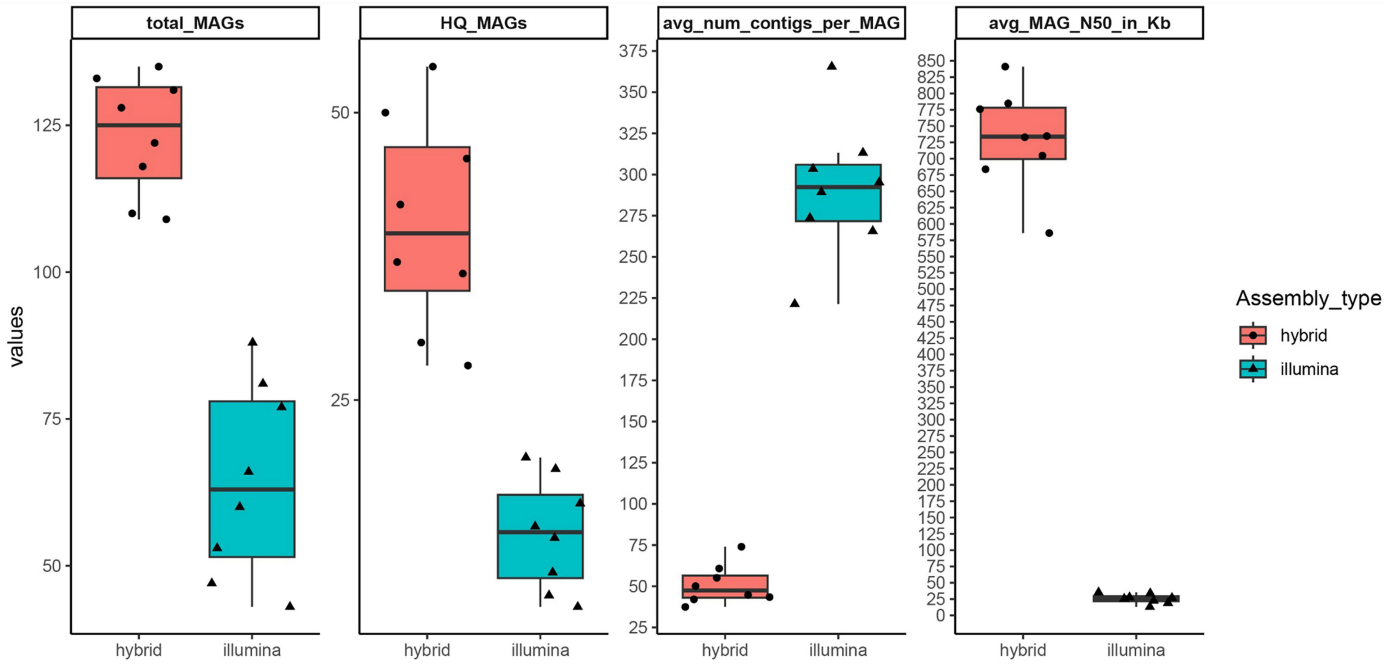
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10778-z.

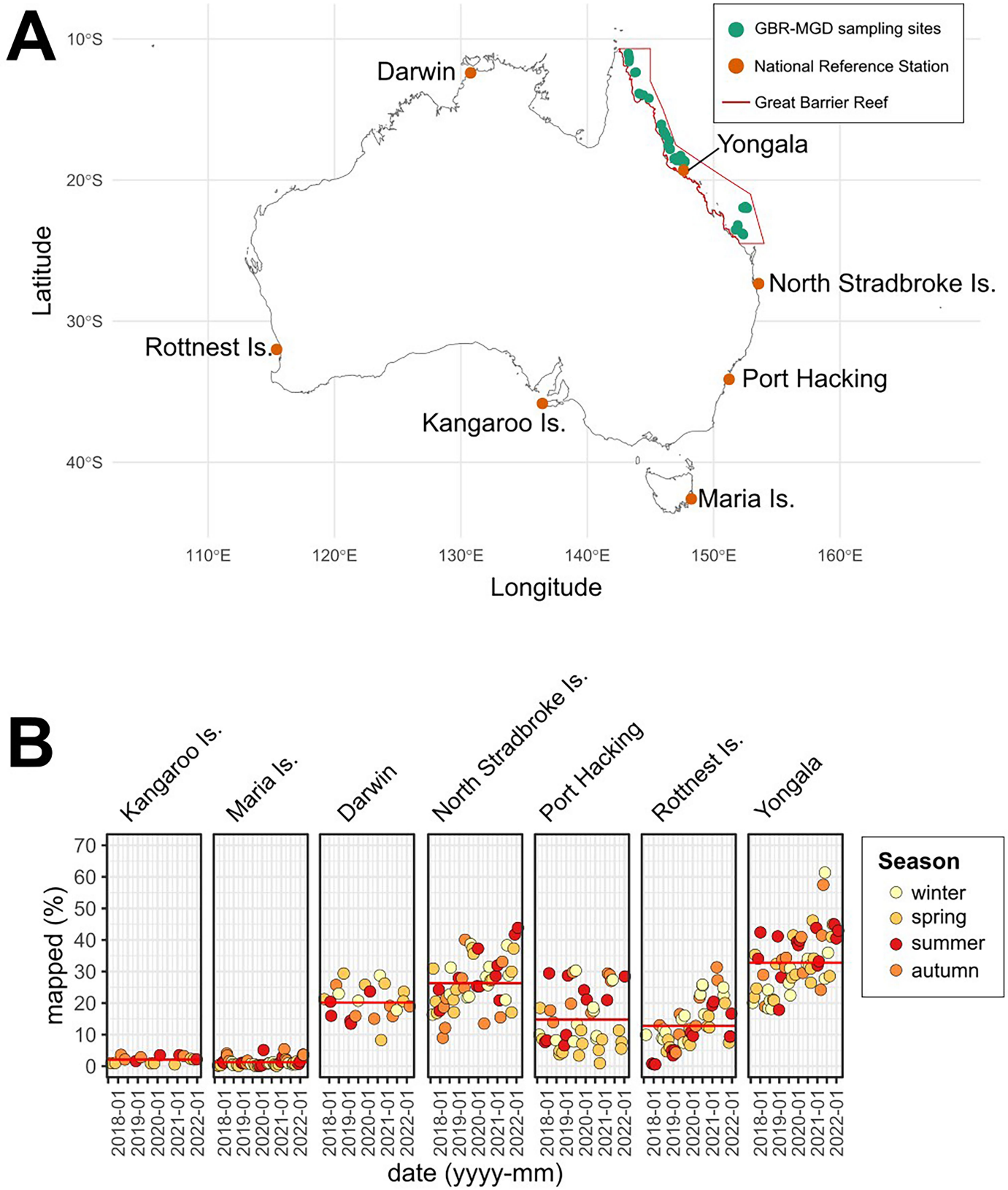
Correspondence and requests for materials should be addressed to David G. Bourne, Philip Hugenholtz or Yun Kit Yeoh.

Peer review information *Nature* thanks Harriet Alexander and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

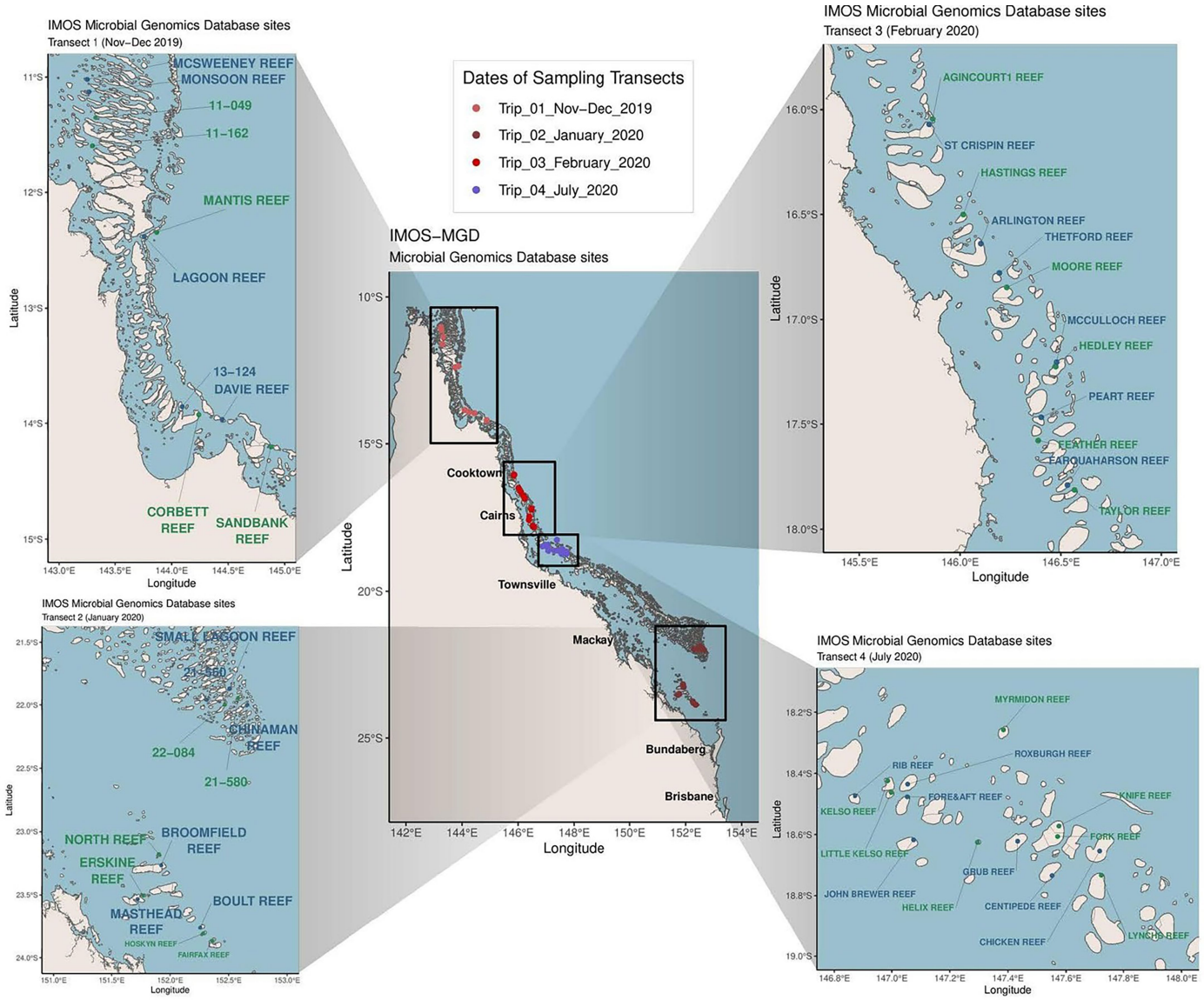
Reprints and permissions information is available at http://www.nature.com/reprints.



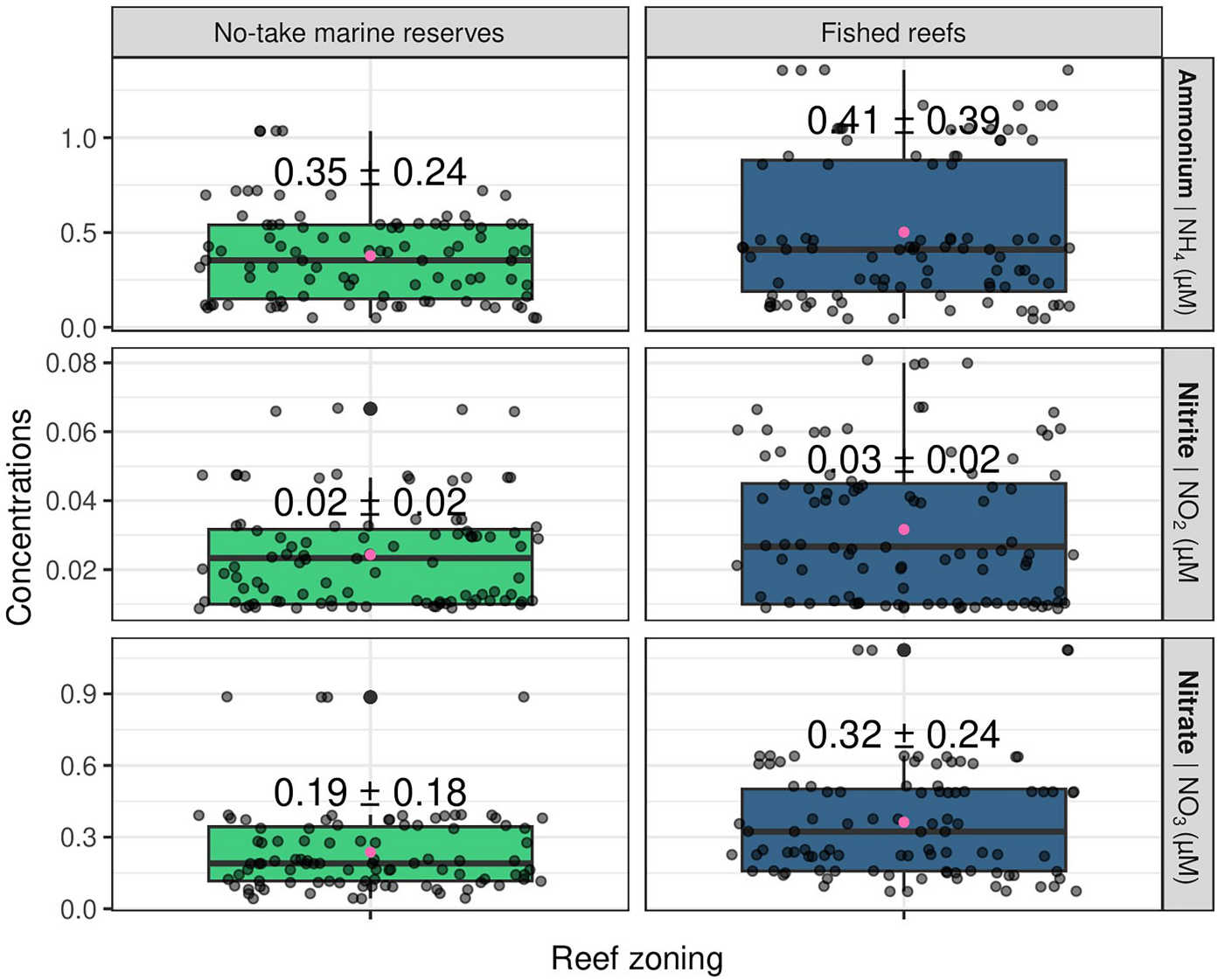
Extended Data Fig. 1 | Metagenome-assembled genome (MAG) quality statistics for Illumina-only and hybrid metagenome assembly benchmarking. Box and whisker plots showing the number of MAGs recovered with quality >50 (%completeness – 3x %contamination) from Illumina-only and Nanopore Hybrid metagenome assemblies (n = 8 samples from eight different reefs used for comparative Illumina-only vs. hybrid assembly benchmarking and MAG binning), number of high-quality MAGs recovered, average number of contigs per MAG, and the average MAG N50. Centre lines of the box plots represent median values, upper and lower box boundaries represent upper and lower quartiles, respectively, whiskers extend to 1.5x interquartile range, and value minima and maxima are represented by data points.



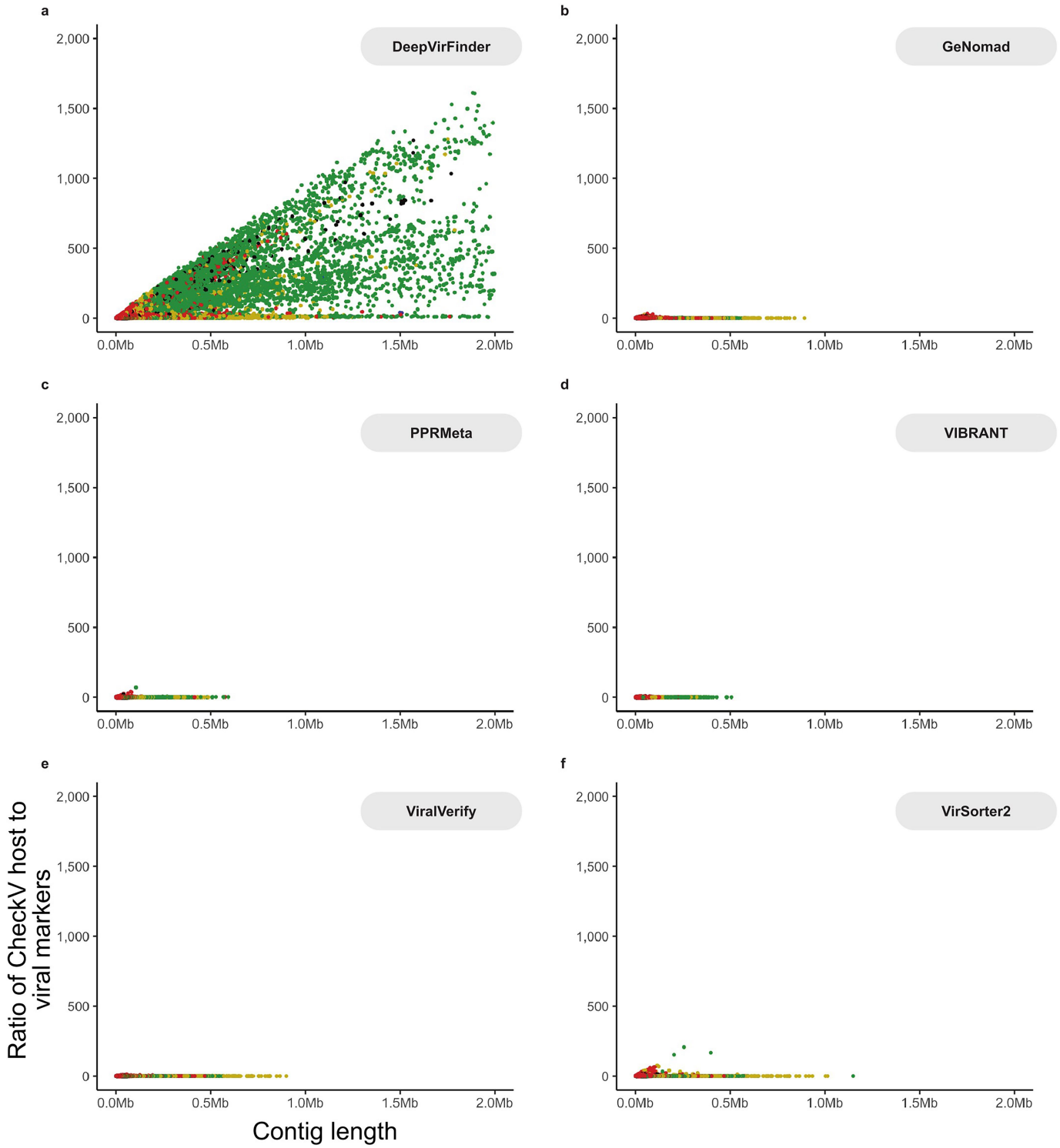
Extended Data Fig. 2 | Read mapping rates of metagenomes from Australia's seven National Reference Stations (NRS) to prokaryotic metagenome-assembled genomes (pMAGs) from the Great Barrier Reef Microbial Genomes Database (GBR-MGD). A, map of Australia showing location of GBR-MGD sampling sites and the seven NRS[91]. **B**, percent reads from NRS metagenomes mapped to GBR-MGD pMAGs (dereplicated at 95% average nucleotide identity). Each circle represents one metagenome data set coloured to indicate season. Red lines represent average % mapped reported on the top of the respective panels. The 272 NRS metagenomes were provided by Andrew Bissett (CSIRO Australia) and are available through the Bioplatforms Australia data portal (https://data.bioplatforms.com/).



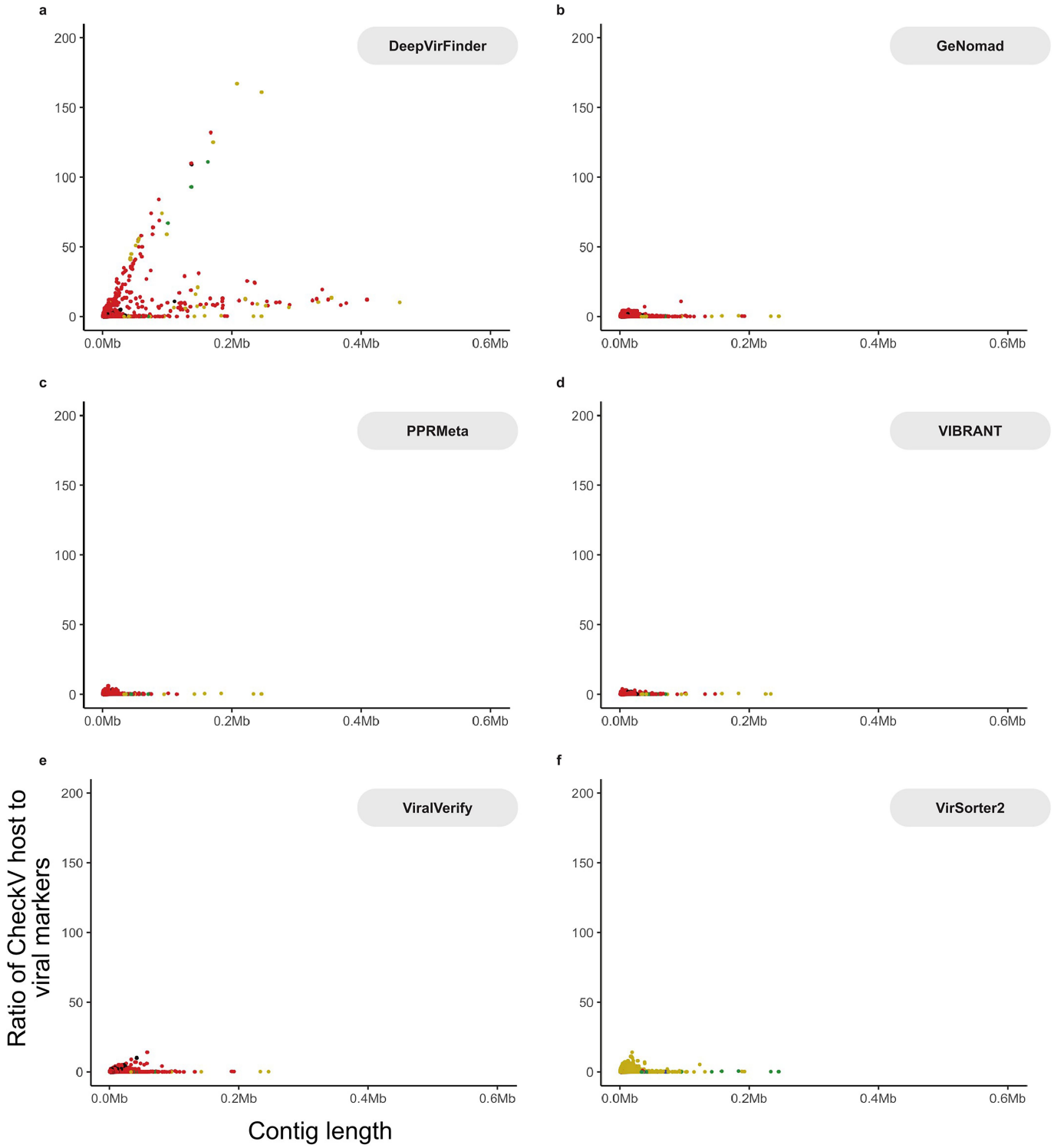
Extended Data Fig. 3 | Site map showing Fished and No-Take Marine Reserve (NTMR) reefs designated under the 2004 Great Barrier Reef (GBR) Marine Park Authority (MPA) management plan. Reefs names in green font indicate NTMR reefs, blue font indicates Fished reefs. See GBR zoning maps for more information, available via the GBR MPA website https://www2.gbrmpa.gov.au/access/zoning/zoning-maps.



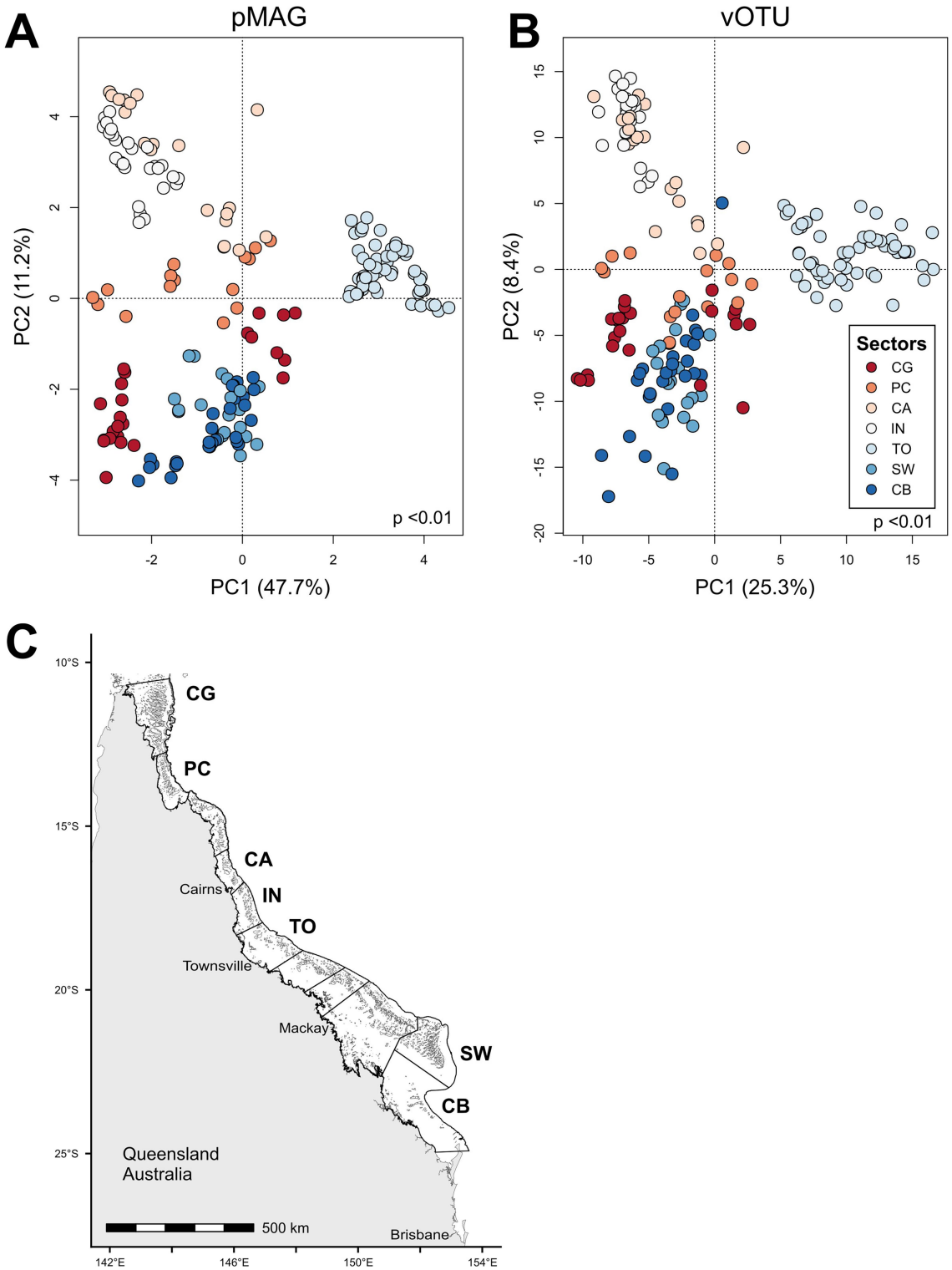
Extended Data Fig. 4 | Dissolved nitrogen concentrations at No-Take Marine Reserve (NTMR) and Fished reefs. Three measurements with taken at each reef using three of the four water samples collected (n = 69 for NTMRs, n = 75 for Fished reefs). NTMR reefs (n = 23) are represented in green, Fished reefs (n = 25) in blue. Red dots represent average values, while the thick middle line, upper and lower box boundaries represent median, upper and lower quartiles, respectively. Whiskers extend to 1.5x interquartile range. Values quoted above the individual box plots are mean ± standard deviation.



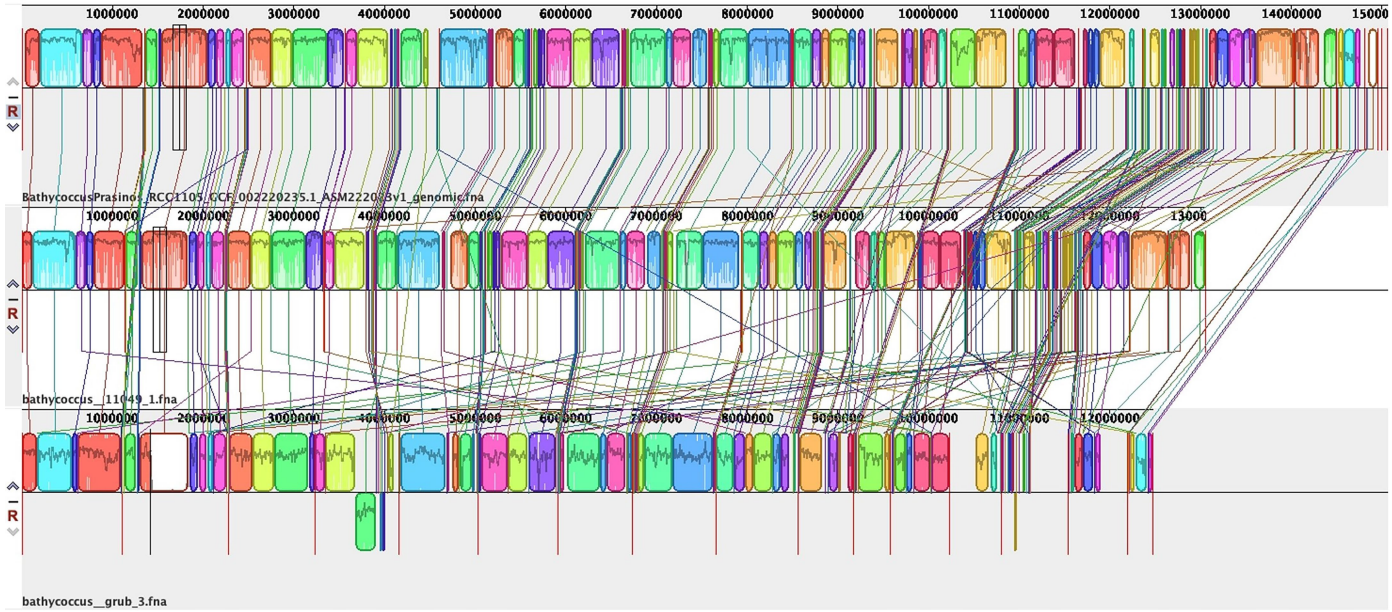
Extended Data Fig. 5 | Ratio of CheckV host to viral markers vs contig length in hybrid assemblies. Dotplots show the ratio of CheckV host to viral marker genes in contigs predicted to be viral by a) DeepVirFinder, b) GeNomad, c) PPRMeta, d) VIBRANT, e) ViralVerify, and f) VirSorter2 in hybrid Nanopore-Illumina assemblies from the GBR-MGD. Colours indicate the assigned CheckV completeness label of Complete (blue), High-Quality (green), Medium-quality (yellow), Low-Quality (red), and Not Determined (black).



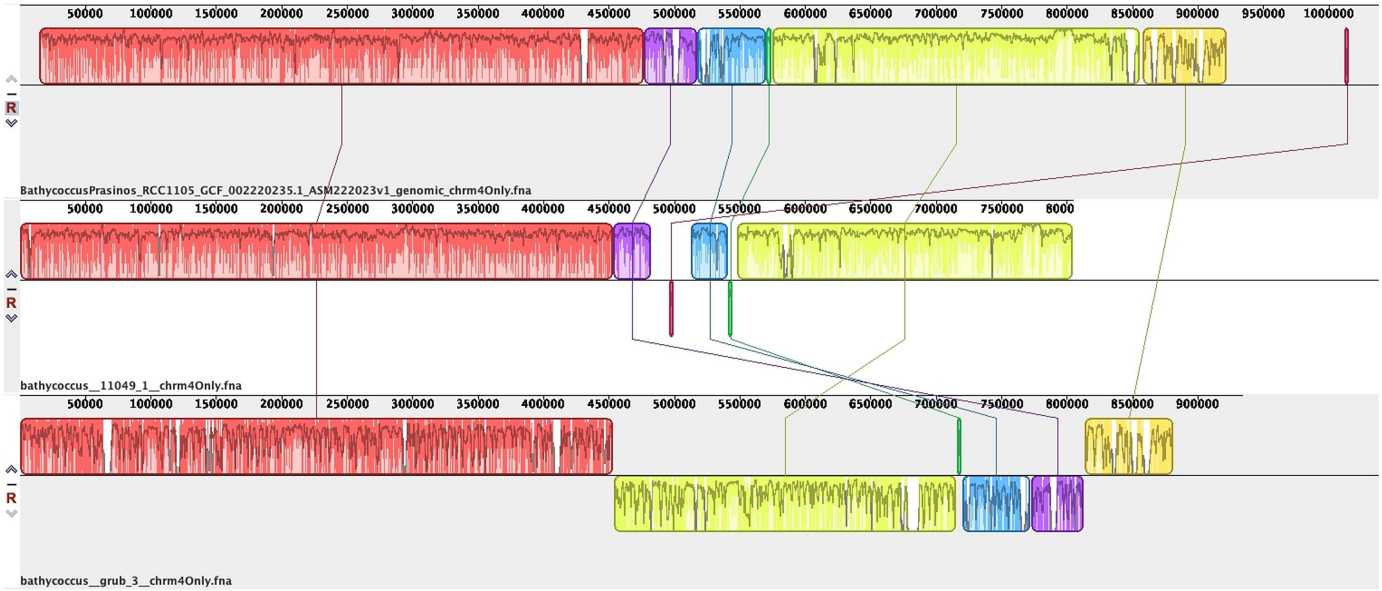
Extended Data Fig. 6 | Ratio of CheckV host to viral markers vs contig length in Illumina-only assemblies. Dotplots show the ratio of CheckV host to viral marker genes in contigs predicted to be viral by a) DeepVirFinder, b) GeNomad, c) PPRMeta, d) VIBRANT, e) ViralVerify, and f) VirSorter2 in Illumina-only assemblies from the GBR-MGD. Colours indicate the assigned CheckV completeness label of Complete (blue), High-Quality (green), Medium-quality (yellow), Low-Quality (red), and Not Determined (black).



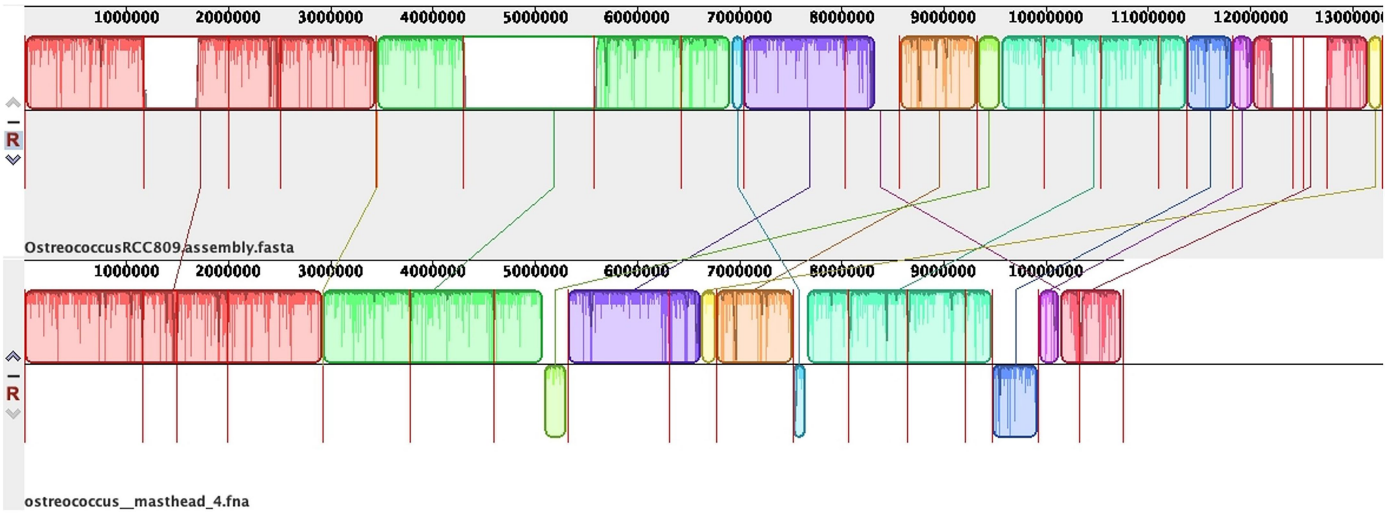
Extended Data Fig. 7 | Community composition of prokaryotic metagenome-assembled genomes (pMAGs) and virus operational taxonomic units (vOTUs) on the Great Barrier Reef (GBR). A, Principal component analysis of community composition of pMAGs. **B**, Principal component analysis of community composition of vOTUs. **C**, location of GBR sectors sampled in this study. Circles in panels A and B are coloured by GBR sector; refer to panel C for sector locations. Community composition significantly differed by sector ($p < 0.01$, PERMANOVA). CG: Cape Grenville, PC: Princess Charlotte Bay, CA: Cairns, IN: Innisfail, TO: Townsville, SW: Swain, CB: Capricorn Bunker.



Extended Data Fig. 8 | Whole genome alignment of GBR-MGD *Bathycoccus* vs *Bathycoccus prasinos* RCC1105. Alignment of the isolate reference genome from *Bathycoccus prasinos* RCC1105 (top row), *Bathycoccus prasinos* GBR_11049 from the GBR-MGD (middle), and *Bathycoccus prasinos* GBR_Grub from the GBR-MGD (bottom) using the Mauve aligner (progressiveMauve). Each chromosome is represented by one contig and vertical red lines represent contig boundaries. The small outlier chromosome (SOC) is not shown as it is highly divergent from the reference and could not be recovered.



Extended Data Fig. 9 | Alignment of *Bathycoccus prasinos* chromosome 4. Alignment of *Bathycoccus prasinos* RCC1105 (top row), *Bathycoccus prasinos* GBR_11049 from the GBR-MGD (middle), and *Bathycoccus prasinos* GBR_Grub from the GBR-MGD (bottom) using the Mauve aligner (progressiveMauve). Vertical red lines represent contig boundaries.



Extended Data Fig. 10 | Whole genome alignment of GBR-MGD *Ostreococcus* clade B vs *Ostreococcus* clade B RCC809. Alignment of the isolate reference genome from *Ostreococcus* clade B RCC809 (top row) and *Ostreococcus* clade B from Masthead reef from the GBR-MGD (bottom) using the Mauve aligner (progressiveMauve). Each chromosome is represented by one contig and vertical red lines represent contig boundaries.

nature portfolio

Corresponding author(s): David Bourne, Philip Hugenholtz, Yun Kit Yeoh

Last updated by author(s): Jun 1, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	- MinKNOW v20.06.18 for Nanopore long read sequencing on PromethION 24, Guppy v5.0.16 for Nanopore basecalling - Porechop v0.2.4 for splitting nanopore reads with internal adapters - Aviary pipeline v0.3.3 for metagenome assembly, which invoked metaFlye v2.9 for nanopore long read assembly, racon v1.4.3 and pilon v1.23 for polishing, and metaSPAdes v3.15.3 for hybrid metagenome assembly. - Aviary pipeline v0.3.3 for genome binning, which invoked minimap2 v2.18 for read mapping, and metabat v2.15, maxbin2 v2.2.7, rosella v0.3.3, CONCOCT v1.1.0, VAMB v4.0.0, CheckM v1.2.1, CheckM2 v1.0.2, GTDB-TK v2.3.0, Das Tool v1.1.2, for binning, genome dereplication and taxonomic assignment. - CoverM v0.6 for species dereplication based on average nucleotide identity and read mapping for genome coverage and abundances. - inStrain v1.7.1 and bowtie2 v2.4.5 for hybrid vs Illumina-only genome binning benchmarking - DeepVirFinder (not versioned; retrieved from github 2023-03-08), Genomad v1.5.018, PPR-Meta v1.134, Vibrant v1.2.132, viralverify v1.133, Virsorter2 v2.2.431, CheckV v1.0.1, Plasforest v1.2 and PlasX (unversioned, retrieved from github 20230308) for identification of viral and mobile elements. - Pharokka v1.6.1, MAFFT v7.508, trimAl v1.4rev15, IQTREE2 v2.2.0.3, iPHoP v1.3.3, ViPTree v4.0 and ITOL v7 for phylogenetic inference and visualisation of Crassvirales phylogeny. - Whokaryote v1.1.2, EukRep v0.6.6, anvi'o v7.1 for eukaryotic predictions. - Minimap2 v2.28, Mauve v2015-02-13 and Tidk v0.2.65 for eukaryotic genome alignment and identifying telomeric repeats. - R v4.4.1 with Phyloseq v1.48.0, vegan v2.6.8, ranger v0.17.0, mixomics v6.26.0, maaslin3 v0.99.4, DESeq2 v1.44.0, gisaimsr v0.0.1, ggplot2 v3.5.1 packages for microbial community composition and reef zoning indicator taxa analyses, plotting figures. - Pyrodigal v3.7.0, MMseqs2 v13.45111, samtools v1.21, kofamscan v1.3.0 for community functional analysis using metagenomic sequence

data
- Inkscape v0.92.5 for editing figures in scalable vector graphics (.svg) file format

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The GTDB R214 database was downloaded from https://data.gtdb.ecogenomic.org/releases/release214/214.0/. MAGs contained in the Ocean Microbiomics Database (OMD) v1 were downloaded from https://sunagawalab.ethz.ch/share/microbiomics/ocean/suppl_data/. Metagenomic sequence datasets from the Australian National Reference Stations are available through the Bioplatforms Australia Data Portal (https://data.bioplatforms.com/) by searching for keyword "NRS" and filtering to "illumina-shortread" sequence data type. Genomes of sequenced Bathycoccus and Ostreococcus are available at RefSeq https://www.ncbi.nlm.nih.gov/refseq/. Caudoviricetes TerL sequences and Crassvirales genomes compiled in Piedade et al70 were downloaded from https://dataverse.nioz.nl/dataset.xhtml?persistentId=doi:10.25850/nioz/7b.b.wg. Conserved Crassvirales genes from Yutin et al67 were downloaded from https://ftp.ncbi.nlm.nih.gov/pub/yutinn/crassfamily_2020/. The COG2014 database is available at https://ftp.ncbi.nlm.nih.gov/pub/COG/COG2014/data/, Pfam32 database is available at https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/. Raw sequence data produced in this study are available at the European Nucleotide Archive under project accession PRJEB82623 and secondary accession ERP166296 with project name "Great Barrier Reef seawater microbiomes genome database". All Illumina and Nanopore sample identifiers are provided in Supplementary Table 23. The 5,283 pMAGs, 15 Ostreococcus Clade B genomes, 5 Bathycoccus genomes, 27 Crassvirales and group contigs are available at Zenodo.org (https://doi.org/10.5281/zenodo.17109886).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	no human participants involved in this study.
Reporting on race, ethnicity, or other socially relevant groupings	no human participants involved in this study.
Population characteristics	no human participants involved in this study.
Recruitment	no human participants involved in this study.
Ethics oversight	no human participants involved in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Seawater samples were collected from 48 coral reefs across the length of the Great Barrier Reef (GBR) for metagenome sequencing and genome assembly to obtain genome representation of the resident microbial communities. We show that dominant and ubiquitous marine bacteria are often underrepresented by short-read DNA sequencing but can be recovered using Nanopore long read DNA sequencing technology. Analysis of the recovered genomes, including microbes, viruses and eukaryotes demonstrate as yet uncharacterised diversity at the GBR and an ecosystem-wide effect of fisheries management on the microbial communities of surrounding seawater, especially microbes recovered through Nanopore sequencing.
Research sample	At each reef, we collected four 5 L biological replicates of seawater above the reef at about 2 - 10 m depth for a total of 192 samples.
Sampling strategy	Seawater samples were collected using nalgene or Niskin bottles. Each 5 L replicate was filtered through a 5 um filter followed by 0.22 um Sterivex filters to capture microbial cells. Sterivex filters were snap frozen immediately and stored at -75 C.

Data collection	DNA was extracted from the Sterivex filters using a phenol:chloroform protocol at the Australian Institute of Marine Science. Extracted DNA was sent to a sequencing service provider for DNA sequencing.
Timing and spatial scale	Seawater sampling was performed from Nov 2019 - July 2020 at 48 sites across the GBR along Australia's north eastern coast alongside the Australian Institute of Marine Science's GBR Long Term Monitoring Program. These 48 sites span approximately 1600 km of the GBR.
Data exclusions	One of the 192 seawater samples was excluded as wrong chemical was added to Sterivex filter during DNA extraction.
Reproducibility	Four technical seawater replicates were collected at each of the 48 coral reefs. DNA was extracted and sequenced from 191 samples except the one excluded replicate.
Randomization	We used mixomics to account for batch effects associated with geography in analysing microbial taxa indicative of reef management and protection status.
Blinding	Blinding not relevant for the recovery and analysis of microbial genomes.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The first three sampling trips occurred in summer and the last trip in winter. No extreme weather events were encountered during the trips.
Location	Seawater samples were collected at 48 coral reefs across the GBR spanning approximately 1600 km. Four 5 L replicates were collected at each reef at 2 - 10 m depth. Figure 1 of the manuscript shows the location of the specific reefs on the GBR.
Access & import/export	Seawater samples were collected under permit G12/35236–1 issued by the Great Barrier Reef Marine Park Authority of the Australian Government. Access was provided through the Australian Institute of Marine Science (AIMS) GBR Long Term Monitoring Program field trips using AIMS research vessels.
Disturbance	Our seawater collection method is unlikely to disturb the reef. Niskin bottles attached to a wire were lowered into the water from the research vessel and brought back to the surface for filtering.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	No plants involved in this study.
Novel plant genotypes	No plants involved in this study.
Authentication	No plants involved in this study.